Exploring the performance of transgender athletes *p. 322*

A threat to innovation in biomedicine *p. 335*

How social amoebae manage a microbiome *p. 402*

Science

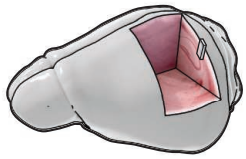
$15
27 JULY 2018
sciencemag.org

AAAS

SPECIAL ISSUE

FRONTIERS IN COMPUTATION

CONTENTS



328 & 380
Mapping cells in tissue

27 JULY 2018 • VOLUME 361 • ISSUE 6400

SPECIAL SECTION

Frontiers in Computation

ON THE COVER

Strobe-lit view of the Summit supercomputer, capable of a peak performance of 200 petaflops, at the U.S. Department of Energy's Oak Ridge National Laboratory. Widespread advances in computing power and algorithm development have greatly enhanced modeling capabilities in astronomy, physics, chemistry and materials science, and biology. See the special section starting on page 342. *Photo: Carlos Jones, courtesy of ORNL; Creative Services: Renee Manning, courtesy of ORNL*

CREDITS: (CLOCKWISE FROM TOP) C. BICKEL/*SCIENCE*; MATTHIEU PALEY/NATIONAL GEOGRPAHIC CREATIVE; ANDY POTTS

NEWS

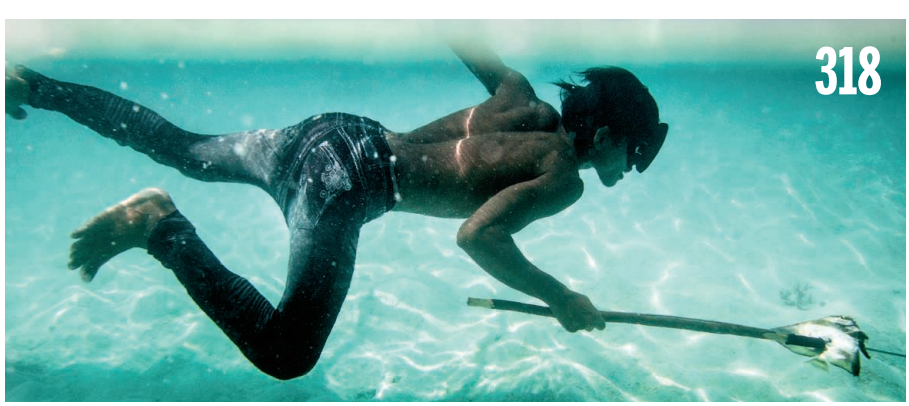
IN BRIEF

IN DEPTH

FEATURE

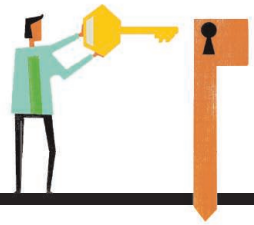
INSIGHTS

PERSPECTIVES



Published by AAAS

CONTENTS



335 Biomedical innovation under threat

27 JULY 2018 • VOLUME 361 • ISSUE 6400

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $1808; Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery): $89. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act is granted by AAAS to libraries and others who use Copyright Clearance Center (CCC) Pay-Per-Use services provided that $35.00 per article is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

CREDITS: (FROM TOP) ADAPTED FROM SHAW NIELSEN/THE ISPOT; KRIS WIKTOR/SHUTTERSTOCK.COM

Published by AAAS

Science
AAAS
1200 New York Avenue, NW, Washington, DC 20005
Clarendon House, Clarendon Road, Cambridge, UK CB2 8FH

SUBSCRIPTION SERVICES For change of address, missing issues, new orders and renewals, and payment questions: 866-434-AAAS (2227) or 202-326-6417, FAX 202-842-1065. Mailing addresses: AAAS, P.O. Box 96178, Washington, DC 20090-6178 or AAAS Member Services, 1200 New York Avenue, NW, Washington, DC 20005

INSTITUTIONAL SITE LICENSES 202-326-6730 **REPRINTS:** Author Inquiries 800-635-7181 **COMMERCIAL INQUIRIES** 803-359-4578 **PERMISSIONS** 202-326-6765, permissions@aaas.org **AAAS Member Central Support** 866-434-2227 www.aaas.org/membercentral.

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

EDITORIAL

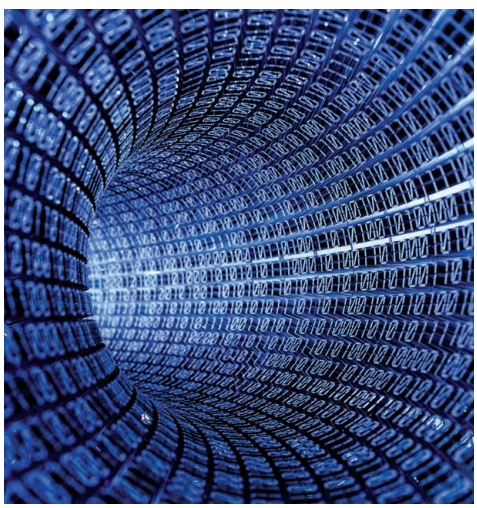
A quantum future awaits

A century ago, the quantum revolution quietly began to change our lives. A deeper understanding of the behavior of matter and light at atomic and subatomic scales sparked a new field of science that would vastly change the world's technology landscape. Today, we rely upon the science of quantum mechanics for applications ranging from the Global Positioning System to magnetic resonance imaging to the transistor. The advent of quantum computers presages yet another new chapter in this story that will enable us to not only predict and improve chemical reactions and new materials and their properties, for example, but also to provide insights into the emergence of spacetime and our universe. Remarkably, these advances may begin to be realized in a few years.

From initial steps in the 1980s to today, science and defense agencies around the world have supported the basic research in quantum information science that enables advanced sensing, communication, and computational systems. Recent improvements in device performance and quantum bit ("qubit") approaches show the possibility of moderate-scale quantum computers in the near future. This progress has focused the scientific community on, and engendered substantial new industrial investment for, developing machines that produce answers we cannot simulate even with the world's fastest supercomputer (currently the Summit supercomputer at the U.S. Department of Energy's Oak Ridge National Laboratory in Tennessee).

Achieving such quantum computational supremacy is a natural first goal. It turns out, however, that devising a classical computer to approximate quantum systems is sometimes good enough for the purposes of solving certain problems. Furthermore, most quantum devices have errors and produce correct results with a decreasing probability as problems become more complicated. Only with substantial math from quantum complexity theory can we actually separate "stupendously hard" problems to solve from just "really hard" ones. This separation of classical and quantum computation is typically described as approaching quantum supremacy. A device that demonstrates a separation may rightly deserve to be called the world's first quantum computer and will represent a leap forward for theoretical computer science and even for our understanding of the universe.

Once a real quantum computer is realized, what's next? In the coming decade, we can expect that some problem-solving will be optimized much more rapidly using quantum devices. We can also expect that efficient sampling from a probability distribution—the theoretical version of a machine learning algorithm—will become a place where quantum computers can shine. In the longer term, error correction and factoring may change the landscape further.

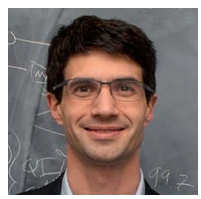


"Achieving such quantum computational supremacy is a natural first goal."

However, the lowest hanging fruit will be improving the ability to work with quantum mechanics. In the past, knowledge of quantum mechanics has been refined by comparing classical computational techniques to what has been observed by experiments—from solving differential equations to brute force simulation to new approximation methods in chemistry and materials science. If quantum computational supremacy is achieved, we may be able to test new techniques without requiring such comparison. This will reduce the cycle of research and transform how science is conducted.

Aiming toward these outcomes, the White House Office of Science and Technology Policy has established a new interagency group that is tasked with creating a national strategy to nurture a full quantum ecosystem through coordinated research between government, academia, and industry. This will include engagement across community boundaries and between disciplines to ensure a strong, quantum-smart future workforce. Matching this with similar efforts worldwide should allow us to catch glimmers of new scientific horizons and to develop the potential industries and new technologies that may emerge from investing in quantum information science.

–Jacob M. Taylor



Jacob M. Taylor
is a Fellow at the U.S. National Institute of Standards and Technology, Gaithersburg, MD, and is assistant director for Quantum Information Science at the White House Office of Science and Technology Policy, Washington, DC. jacob.taylor@nist.gov

10.1126/science.aau8256

NEWS

13% Portion of the world's oceans that remains as wilderness, minimally affected by humans. Most is in polar regions and around remote Pacific Island nations—and less than 5% is protected (*Current Biology*).

IN BRIEF

Edited by **Jeffrey Brainard**

CONSERVATION

Trump to overhaul species protections

The Endangered Species Act covers the woodland caribou (*Rangifer tarandus caribou*).

President Donald Trump's administration announced last week it plans major changes to how federal agencies interpret and apply the Endangered Species Act. One change proposed on 25 July by the U.S. Fish and Wildlife Service and the National Oceanic and Atmospheric Administration would mean that, in the future, species designated as threatened could receive less protection than those listed as endangered. The agencies also want to reduce how far into the future they must look when evaluating extinction risk and to have the flexibility to shrink the area of so-called critical habitat a species might need to survive. In addition, the agencies plan to begin to describe the financial costs of protecting certain species, although they will continue to base determinations to protect them solely on biological considerations. Industry groups are generally welcoming the revisions, which the executive branch expects to adopt administratively later this year after a public comment period; conservation groups warn the changes will undermine efforts to save many species.

Hungary takes academy's funding

SCIENCE POLICY | Hungary's budget for 2019, which its Parliament approved on 20 July, transfers two-thirds of the Budapest-based Hungarian Academy of Sciences's (MTA's) €124 million budget to the Ministry for Innovation and Technology (ITM). The government says this will improve efficiency, but many scientists in Hungary and abroad see the move as a power grab by an increasingly authoritarian regime that will extend political influence over science spending and research agendas. "We accept the outcome, but we expect to continue negotiating with [ITM] Minister László Palkovics to see what the exact specifics of this new situation will be," says Tamás Simon, an MTA spokesperson. MTA President László Lovász says Palkovics has not supported basic, "bottom-up" research proposals, which Lovász calls necessary to produce cutting-edge science.

Viagra study halted over risks

BIOMEDICINE | A Dutch study testing whether administering sildenafil, better known as Viagra, to pregnant women can help poorly growing fetuses was halted on 19 July after evidence emerged that the drug may cause a potentially fatal complication in the newborns. Fetal growth restriction, which can lead to stillbirth and neonatal death, is caused by a lack of blood flow from the placenta to the fetus. Sildenafil dilates these blood vessels, which smaller clinical studies had suggested might benefit the fetus. But an interim analysis of the new study at Amsterdam University Medical Center found that pregnant women who took the drug had a higher risk of losing their babies after birth because the newborns experienced lung hypertension—11 in the treatment group died of this cause. Similar studies in three other countries found that sildenafil had no benefit, but did no harm either; a study underway in Canada has been put on hold based on the new data.

Fight brews over census choice

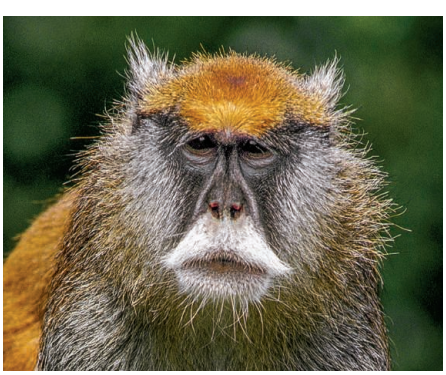
STATISTICS | If there's a fight over Steven Dillingham, whom President Donald Trump nominated on 18 July to lead the U.S.

PHOTO: NICK TREHEARNE/GETTY IMAGES

Census Bureau, it's likely to focus on the administration's commitment to an accurate count in 2020, not his experience leading two smaller federal statistical agencies or his multiple degrees. Secretary of Commerce Wilbur Ross has been sued for adding a citizenship question to the 2020 census, and there's continuing concern over whether the agency can pull off the first-ever head count using internet responses. "Steve is a tip-top scholar, a progressive administrator, and a very ethical guy," says Geoff Alpert, a professor of criminology at the University of South Carolina in Columbia, who worked under Dillingham at the Bureau of Justice Statistics in the early 1990s. But Representative Carolyn Maloney (D–NY) wants Dillingham "to reject the Administration's attempt to add a citizenship question" before any Senate vote.

The real-life Lorax?

POPULAR CULTURE | In Theodor "Dr. Seuss" Geisel's beloved picture book *The Lorax*, the titular character tries to defend his forested home from an insatiable businessman—a lasting allegory for environmental protection. This week, scientists report they have identified a species of monkey that may have inspired the Lorax's implike features—and in a twist the Lorax himself might lament, they note that environmental changes have reduced the monkey's range by half since 1991. Scientists report this week in *Nature Ecology & Evolution* that Geisel may have based the Lorax on the patas monkey (*Erythrocebus patas*). Although no evidence directly links the fictional and real creatures, historical documents show that Geisel, who published the book in 1971 and died in 1991, was writing it while vacationing in Nanyuki, Kenya. There he could have encountered the animal and its main food source, the whistling thorn acacia (*Acacia drepanolobium*), which may have inspired the book's "truffula" trees. To make a stronger case, the researchers conducted computer-aided facial analysis of several species of Kenyan monkeys and found strong similarities between the Lorax and the patas monkey. Kenyan patas populations are losing the acacias to drought and land clearance for agriculture.



Patas monkey populations in Kenya are at risk.

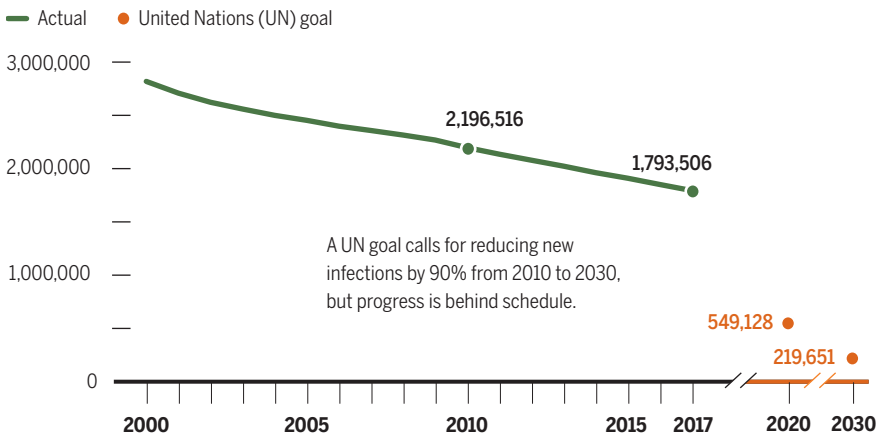
CREDITS: (PHOTO) ARCO IMAGES GMBH/ALAMY STOCK PHOTO; (GRAPHIC) A. CUADRA/*SCIENCE*; (DATA) JOINT UNITED NATIONS PROGRAMME ON HIV/AIDS

PUBLIC HEALTH

Progress lags in HIV/AIDS fight

The world is behind on meeting an ambitious United Nations target of ending AIDS as a public health threat by 2030. To accomplish this goal, set in 2015 by the Joint United Nations Programme on HIV/AIDS (UNAIDS) in Geneva, Switzerland, new infections must drop to 90% of their 2010 levels. In a report titled *Miles to Go* and released on the eve of this week's International AIDS Conference in Amsterdam, UNAIDS Executive Director Michel Sidibé warns that "complacency" is setting in. To meet the 2030 target, new infections must decline by 75% between 2010 and 2020. But new infections fell only 18% between 2010 and 2017. Sidibé stresses that "action now can still put us back on course to reach the 2020 targets," adding, "This report is a wake-up call."

New HIV infections worldwide



Sloppy science more widespread

STATISTICS | Several studies have shown that many psychology researchers admit to questionable practices in analyzing and reporting data—a phenomenon that helps explain why so many of the field's studies have proved impossible to reproduce. But ecologists and evolutionary scientists aren't behaving much better, according to a study published on 16 July in *PLOS ONE* by Hannah Fraser of the University of Melbourne in Australia and colleagues. Among 807 researchers surveyed in those disciplines, 64% reported they had at least once failed to report outcomes that were statistically insignificant; 42% had extended a study to collect more data and reach statistical significance, a form of "p-hacking"; and 51% had reported an unexpected finding as if it had been hypothesized from the beginning. The rates resemble those found in psychology; remedies such as education, changing the incentive structure in science, and preregistration of studies could help bring them down, the authors write.

Recent geological age named

NATURAL HISTORY | Scientists last week announced a new name for the past 4200 years of geologic history: the Meghalayan age. The declaration by the International Commission on Stratigraphy dividesthe Holocene, our current, warm, human-dominated epoch, into three ages based on global changes: the Greenlandian, which started as glaciers retreated 11,700 years ago; the Northgrippian, marked by a cooling trend 8300 years ago; and the Meghalayan, which began 4200 years ago with a mysterious drought and cooling trend that coincided with the decline of several civilizations. The Meghalayan's marker in the rock record—called the "golden spike"—comes from oxygen isotopes found in a stalagmite from the Meghalaya plateau of northeast India. The newly defined age names and dates provide standardized versions of divisions that scholars had long drawn between an early, middle, and late Holocene. The commission has not finished a study of whether the world, by the 1950s, left the Holocene and entered a new geological epoch, the Anthropocene, dominated by human influence.

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

PHOTO: AP PHOTO/JACQUELYN MARTIN

SCIENTIFIC COMMUNITY

Report details harassment by famed biologist

Allegations that felled Francisco Ayala divide scientists, university community

By **Meredith Wadman**

The investigative report that earlier this month triggered the ouster of prominent evolutionary geneticist Francisco Ayala from the University of California (UC), Irvine, concluded he had a pattern of harassing behavior toward women. The document concludes that he told a pregnant colleague "you're so huge" and repeatedly put his hands under a female administrator's jacket, rubbing them up and down her sides. According to the report, Ayala told a female professor that she had been so animated while giving a talk that he thought she would "have an orgasm." In another instance, the report states, he invited a junior professor in a crowded meeting to sit on his lap, saying he would enjoy the presentation more that way.

The 97-page report, completed in May and obtained by *Science* (currently available at http://bit.ly/AyalaReport), describes evidence of unwelcome conduct by Ayala that continued after he was warned to stop. The internal university investigation, prompted by complaints submitted by four UC Irvine women, detailed off-color remarks and repeated unsolicited compliments on women's physical appearances; these behaviors often were witnessed by one or more of the 61 people interviewed for the investigation. The investigators said the complainants felt professionally undermined by his conduct and they concluded that Ayala, 84, violated UC Irvine's sexual harassment and sex discrimination policies in the cases of three of the four women who lodged complaints. The university accepted Ayala's resignation, effective 1 July, and plans to strip his name from its science library and biology building.

In responses included in the report, Ayala strenuously denies most of the allegations. He told investigators that the entire complaint of Department of Ecology and Evolutionary Biology (EEB) Professor and Chair Kathleen Treseder, who reported the "huge" and orgasm comments, "was a lie."

"I saw my compliments as courtesies. And they turned those courtesies into sexual harassment," Ayala told *Science* in an interview last week. "I have never intentionally caused sexual harassment to anybody," he also wrote in an email to UC Irvine Chancellor Howard Gillman days after the probe was launched. "To the extent that my actions may have caused harm to others … I apologize from the deepest of my heart and of my mind."

Three of the women who lodged complaints directed questions to their lawyer, who did not respond to repeated requests for comment. *Science* was unable to reach the fourth complainant.

Ayala's firing has drawn praise from some scientists and sharp criticism from others on the UC Irvine campus and in his native Spain. Critics questioned whether he received due process and complained that the university had not detailed the charges against him. But the report's contents seem only to have deepened the division.

"I'm just shocked that this man's life was ruined over this collection of reactions to his behavior," says Elizabeth Loftus, a UC Irvine experimental psychologist who has vocally supported Ayala. After reading the report, she said the "thin" allegations "are far, far from the obviously bad behaviors that we want to be punishing. I feel like: 'Who's next?'"

But Jane Zelikova, an ecologist at the University of Wyoming in Laramie and a founder of the international advocacy group 500 Women Scientists, says: "No form of harassment is OK. … He could have corrected his behavior. He did not. Being fired for doing something that is illegal is justice."

Ann Olivarius, a senior attorney at McAllister Olivarius in Saratoga Springs, New York, who specializes in sexual harassment, reviewed the UC Irvine report at *Science*'s request. "Unlike many harassers who have sex with students or pressure them directly

In the wake of a harassment report, Francisco Ayala resigned from the University of California, Irvine.

for sex, Ayala did not cross those boundaries," she notes. "But he clearly made multiple women feel degraded. … Senior university officials warned him to stop acting in these ways, but he continued."

Robert Cook-Deegan, a historian of science with Arizona State University who is based in Washington, D.C., and also read the report, concluded that "Norms are changing really fast and I think this 84-year-old got caught in a norm shift."

Ayala, who was elected to the National Academy of Sciences in 1980, did pioneering work in evolution and genetics and was a prominent public defender of evolution. In 1995, he was president of AAAS, which publishes *Science*. Since 2010, he has donated more than $11 million to UC Irvine.

Ayala admitted to making the "sit in my lap" comment to complainant Jessica Pratt, an assistant teaching professor, in 2015, as she prepared to present at a crowded faculty meeting. Ayala told investigators it was a one-time lapse showing "a horrendous lack of judgment." (A graduate student who was interviewed as a witness described a separate instance in which she said Ayala invited her, too, to sit in his lap.)

Pratt complained to the then–EEB department chair, prompting Ayala to visit Pratt in her office, according to the report. Ayala told investigators that he "apologized profusely" to Pratt, telling her he intended the comment to be playful, like he was addressing a niece or granddaughter. But when Pratt told Ayala that the comment was overheard—something other witnesses corroborated—Ayala called her a liar, according to the report. Pratt lodged an informal complaint with UC Irvine's Office of Equal Opportunity and Diversity. She told investigators that Ayala's conduct "made her question whether he respected her work. … She even began to question her own merit as a scientist."

> ***"Dr. Ayala's very public punishment will send a loud signal that times are changing …"***
>
> **Ann Olivarius**, McAllister Olivarius

As a result of Pratt's complaint, Associate Chancellor Kirsten Quanbeck warned Ayala in 2015 to watch his language with women and told him that his conduct was viewed as unwelcome and was out of line with university policy. The EEB department chair gave him a similar warning, the report says.

One complainant, Benedicte Shipley, an assistant dean in UC Irvine's School of Biological Sciences, told investigators she felt she had to put up with Ayala's attentions because of his power as a major donor. "I just learned that women don't like to be told they're beautiful, but I know you don't mind," Shipley recalled Ayala saying in 2016, rubbing her sides while kissing her cheeks in greeting—a behavior that occurred repeatedly, she said. A male professor noted this encounter and asked Shipley afterward whether she was all right, the report says.

Shipley told investigators she was relieved when Ayala's attention shifted to Treseder, to whom she said Ayala was "glued" at a department social event not long after. Distressed, Treseder asked a male colleague, who corroborated her claim to investigators, to attend such events with her.

In early November 2017, just before the complaints against Ayala were filed, Treseder, who had recently been named the department chair, proposed a code of conduct concerning sexual harassment at a faculty meeting. Ayala pushed back, according to the report, saying the "ladies" in the dean's office wanted him to hug and kiss them.

In the fourth complaint, graduate student Michelle Herrera alleged that Ayala put his hands on her bare shoulders—a behavior Ayala admitted to—and leaned his front against her back as she sat at a picnic table. Ayala vigorously denied leaning against her. Investigators concluded that the incident probably occurred but might not have been gender-based. They found Herrera's claims that Ayala repeatedly commented on her appearance not credible. Herrera could not be reached for comment.

UC Irvine's investigators also assert that Ayala "has engaged in a campaign with the highest University officials to influence the outcome of this investigation." The report says he wrote to Gillman and to Janet Napolitano, president of the UC system, reminding them of his financial and academic contributions to UC Irvine. According to the report, Ayala told the senior investigator as the probe launched that it "needed to end quickly and in his favor and [that] he had lawyers waiting if [it] did not."

Speaking with *Science* last week, Ayala said he had wanted to avoid a protracted legal struggle so he could focus on his science. He added: "I didn't say anything about lawyers."

Olivarius predicts the Ayala case will have an impact beyond UC Irvine. "Dr. Ayala's very public punishment will send a loud signal that times are changing—that harassment … does not mean just extreme misconduct," she says. ■

NUCLEAR PHYSICS

Electron-ion collider wins key endorsement

National Academies report calls for accelerator to look inside protons and neutrons

By **Adrian Cho**

The next dream machine for U.S. nuclear physicists got an important boost this week in a report from the National Academies of Sciences, Engineering, and Medicine. The report committee glowingly approved of the science that could be done with the proposed Electron-Ion Collider (EIC), a billion-dollar accelerator that would probe the innards of protons and neutrons. The endorsement should help the Department of Energy (DOE) justify building the EIC at one of two national laboratories competing to host it, although the project probably won't get the go-ahead for several years.

"We're basically saying, 'You've really got to do this,'" says Ani Aprahamian, a nuclear physicist at the University of Notre Dame in South Bend, Indiana, and co-chair of the report committee.

The inner structure of the proton and the neutron remains mysterious. Crudely, a proton consists of three subatomic particles called quarks, bound by the strong nuclear force. In actuality, a proton is far more complex. Because of the uncertainties inherent in quantum mechanics, its interior roils with quark-antiquark pairs popping in and out of virtual existence. It also teems with gluons, the quantum particles that convey the strong force. The mess is so complex that even basic properties of the proton remain unexplained. For example, its three quarks account for less than 5% of its mass, the rest arising somehow from energy of the virtual quarks and gluons.

By blasting a beam of electrons into a beam of protons or ions, the EIC would help solve this mystery and a parallel one: how the proton gets its spin. Just as in the case of mass, the proton's spin is not simply the sum of the spins of the three quarks; it also has unknown contributions from glu-

ons and from the quarks orbiting around one another. Finally, the EIC could probe the gluons' behavior for so-called emergent properties. For example, some theories predict that a proton's gluons crowd into a single quantum wave a bit like laser light. The EIC would be better for such studies than a machine that smashes protons into protons, such as the Large Hadron Collider (LHC) near Geneva, Switzerland. That's because the electron is an infinitesimally small particle and produces cleaner, easier to interpret collisions, Aprahamian says.

The EIC would help maintain U.S. expertise in colliding beams, says Gordon Baym, a theorist at the University of Illinois in Urbana and co-chair of the report committee. "It's the only collider in the U.S. that's being considered for the next 50 years or so," he says. Now, the United States only has the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory in Upton, New York, which was completed in 1999. Europe has the newer LHC and Japan runs the newly upgraded SuperKEKB electron-positron collider in Tsukuba.

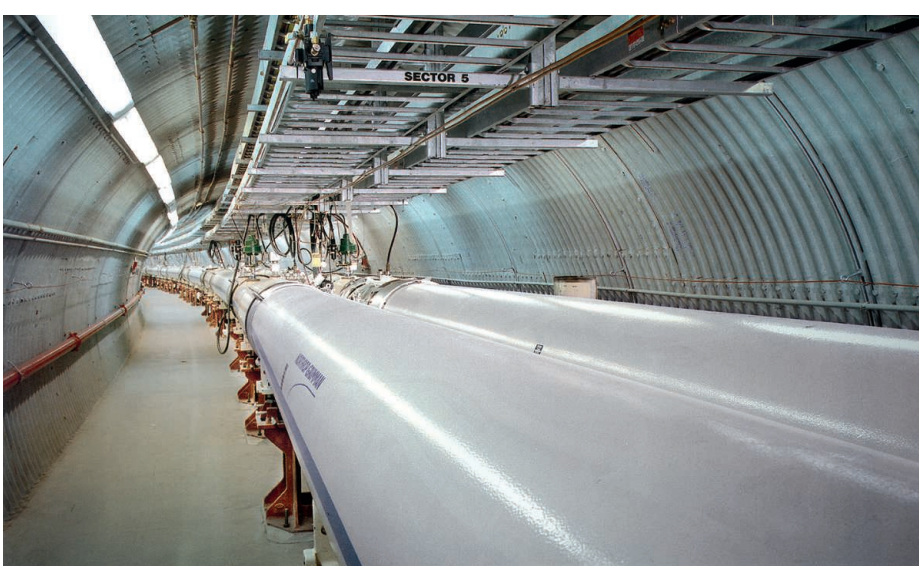
Physicists at two DOE labs propose to assemble the EIC in different ways. At Brookhaven, RHIC smashes together heavy nuclei such as gold to melt them and create a soup called a quark-gluon plasma, the stuff that filled the universe just after the big bang. Brookhaven physicists eventually want to add an electron accelerator to convert RHIC into the EIC.

However, the Thomas Jefferson National Accelerator Facility in Newport News, Virginia, has recently upgraded its Continuous Electron Beam Accelerator Facility (CEBAF), which fires electrons into stationary targets to study protons, neutrons, and nuclei. Jefferson Lab researchers hope to add an ion accelerator to it to make the EIC. Jefferson Lab is a smaller facility that's more narrowly focused on nuclear physics than Brookhaven, and its long-term survival could depend on landing the EIC.

The report avoids comparing the two labs' proposals, but offers DOE officials a scientific case for the EIC. However, on top of running RHIC and CEBAF, DOE's $684 million office of nuclear physics is also building the $730 million Facility for Rare Isotope Beams at Michigan State University in East Lansing, which upon completion in 2020 will generate exotic nuclei. Given the costs of operating those facilities, DOE probably can't afford the EIC any time soon, notes Donald Geesaman, a nuclear physicist at Argonne National Laboratory in Lemont, Illinois, and former chair of DOE's Nuclear Science Advisory Committee.

More time may be welcomed, anyway, as the Brookhaven and Jefferson Lab proposals cannot yet meet key technical requirements. From 1992 to 2007, physicists in Germany ran the Hadron-Electron Ring Accelerator (HERA), which also collided electrons and protons and revealed the gluon. The EIC will run at lower energy than HERA did, but to achieve its goals, it will have to generate collisions at a rate 100 to 1000 times higher with highly polarized electron and proton beams.

For now, Brookhaven and Jefferson Lab scientists are collaborating on the R&D rather than competing. "It's been very exciting to see the community self-assemble," says Jefferson Lab Director Stuart Henderson. Brookhaven Director Doon Gibbs says, "The immediate objective is to keep walking down the road together with both labs in lockstep." If all goes well, however, the two labs' ambitions will eventually collide. ■



The main accelerator at Brookhaven National Laboratory could be repurposed into an electron-ion collider.

SCIENTIFIC ETHICS

Study of 'sea nomads' under fire in Indonesia

Critics say study lacked local ethical approval and role for Indonesian experts

By **Dyna Rochmyaningsih**

In April, a paper showing why Indonesia's Bajau people are such great divers drew worldwide attention as a striking example of recent human evolution (*Science*, 20 April, p. 244). But the study, published in *Cell*, has created a different kind of stir in Indonesia, where some say it is an example of "helicopter research" carried out by scientists from rich countries with little consideration for local regulations and needs.

"Too many mistakes were made here," says geneticist Herawati Sudoyo, who heads the Eijkman Institute for Molecular Biology in Jakarta. Indonesian officials say the research team failed to obtain ethical approval from a local review board and took DNA samples out of the country without the proper paperwork. And some Indonesian scientists complain that the only local researcher involved in the study had no expertise in evolution or genetics. But Eske Willerslev, director of the University of Copenhagen's (KU's) Centre for GeoGenetics, says the team he headed had a permit from the Indonesian government and worked hard to follow the rules. "I would never participate in research that I felt was unethical," Willerslev says. The government hasn't informed him about problems, he says, but, "If we have made an error that violates national or international guidelines, we would like to apologize for that."

The issue escalated in late May, when Pradiptajati Kusuma, a geneticist at the Eijkman Institute who has also studied the Bajau, suggested in a tweet that the team could have faced prosecution under strict new rules on foreign research, proposed by the Indonesian government and now under debate. "Jail? Possible," Kusuma wrote. He later deleted the tweet, but Melissa Ilardo, the *Cell* study's first author, says she was so rattled that she canceled a July trip to Indonesia during which she planned to

A controversial study showed that the Bajau are good divers thanks in part to an enlarged spleen.

PHOTO: MATTHIEU PALEY/NATIONAL GEOGRAPHIC CREATIVE

inform the Bajau about her study. "I did everything I could to conduct this research ethically and properly, and this is breaking my heart," says Ilardo, a Ph.D. student at KU at the time of the fieldwork and now at the University of Utah in Salt Lake City.

Sometimes called sea nomads, the Bajau have lived off the ocean for centuries; men spend much of the day underwater to spear fish and harvest sea cucumbers. In 2015, Ilardo took saliva samples from 59 Bajau individuals in Central Sulawesi and measured their spleen size. The team found that, compared with controls, the Bajau have bigger spleens, which may help prevent hypoxia during long dives by releasing extra blood cells. The researchers also identified a gene variant that may be responsible.

Willerslev's group received a permit for the study from Indonesia's Ministry of Research, Technology and Higher Education, commonly known as RISTEK, in Jakarta and ethical clearance from the Danish National Committee on Health Research Ethics. "We were told that RISTEK permit included local ethical approval as well, thus there was no ethical violation," Willerslev says.

Sadjuga, secretary of RISTEK's Foreign Research Permit Coordinating Team, disputes that account. "We always request ethical clearance from at least one Indonesian research ethics commission," Sadjuga says. (Like many Indonesians, he uses only one name.) Triono Soendoro, who heads the Ethical Commission for National Health Research and Development at the Indonesian Ministry of Health in Jakarta, confirms that the team should have had approval from an ethical panel in Indonesia; guidelines from the Council for International Organizations of Medical Sciences also call for local approval.

The team may also have run afoul of regulations when it shipped DNA samples to Copenhagen for analysis. Ilardo says she filed a material transfer agreement (MTA)—a contract governing the shipment of research samples—with her application to RISTEK. But for the transfer of human DNA, she should have sought approval from the National Institute of Health Research and Development in Jakarta, says Siswanto, who chairs that institute. "If this was a requirement, I would have expected that RISTEK would have told me if my MTA was invalid when I submitted it," Ilardo says.

Some Indonesian scientists, meanwhile, are miffed that the only Indonesian name on the paper is that of Suhartini Salingkat, an education researcher at Tompotika Luwuk Banggai University, a small private institution in Central Sulawesi; according to the paper, she "provided logistical support." Foreign teams "should involve Indonesian scientists in all stages of research," says Mohamad Belaffif, an Indonesian bioinformatician at the HudsonAlpha Institute for Biotechnology in Huntsville, Alabama.

Ilardo says she did try to collaborate with scientists at the Eijkman Institute after RISTEK requested she do so. An email exchange between Ilardo and Sudoyo, provided to *Science* by Willerslev, shows Sudoyo didn't respond to several requests for a meeting in October 2015, before fieldwork began, and later effectively declined a partnership. "As far as I understand, you have your own partner already in the Bajau project, therefore we are not needed," she wrote. (Sudoyo declined to answer *Science*'s questions on this matter.) Given Ilardo's overtures to the Eijkman Institute, "I would love to understand what went wrong and why they suddenly are so angry," says Rasmus Nielsen of the University of California, Berkeley, another senior author on the paper.

Ilardo says she shared her genetic expertise with Tompotika students in an informal seminar, and made the partnership worthwhile for Salingkat by helping her with a research paper. In Ilardo's application to RISTEK, she also promised to organize a meeting with the Bajau people to tell them about the results of the study. But even if she hadn't abandoned that plan following Kusuma's tweet, some argue it would have been too late. "In general, the return [of research results] should coincide with or slightly precede publication so that the participants are not the last to know," says Conrad Fernandez, a bioethicist at Dalhousie University in Halifax, Canada.

Berry Juliandi, a biologist at Bogor Agricultural University in Indonesia, says the country's "tangled" system of permits can be hard to navigate for foreign researchers. "The root of this problem is the weak management of foreign research permits in Indonesia," he says. "How could RISTEK approve Ilardo's permit proposal [when] she doesn't have a valid MTA and ethical clearance from an Indonesian institution?" Working with a bigger, more experienced local institution than Tompotika might have helped the researchers avoid pitfalls, he says.

The case comes at a sensitive time, when Indonesian and foreign scientists are debating rules, proposed in 2017, that would strengthen MTA regulations, compel foreign researchers to include Indonesian colleagues as "equal partners" on projects, and include them as authors on every peer-reviewed paper about the work. Outside researchers would also have to submit raw data to the country's research ministry; some violations would carry prison sentences. Some scientists, both in Indonesia and abroad, say the law is unworkable and could stifle scientific progress. At the same time, RISTEK says it wants to promote research collaborations, and on 5 July, it launched an online system that makes the paperwork easier and less time-consuming for foreign researchers.

Neither RISTEK nor the Ministry of Health has taken action against the researchers over the Bajau study. A spokesperson for *Cell* says the journal is satisfied by the researchers' explanation. "The authors sent us documentation indicating that they received consent from the Indonesian government to conduct this research," he says. "We have no evidence that further investigation of this matter is warranted." ■

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

PLANETARY SCIENCE

Lake spied deep below polar ice cap on Mars

Orbiting radar instrument finds martian analog to pools under Antarctic glaciers

Liquid water lies unseen under ice (white) at Mars's south pole in an image from Mars Express.

By **Daniel Clery**

Far beneath the deeply frozen ice cap at Mars's south pole lies a lake of liquid water—the first to be found on the Red Planet. Detected from orbit using ice-penetrating radar, the lake is probably frigid and full of salts—an unlikely habitat for life. But the discovery, reported online in *Science* this week, is sure to intensify the hunt for other buried layers of water that might be more hospitable. "It's a very exciting result: the first indication of a briny aquifer on Mars," says geophysicist David Stillman of Southwest Research Institute in Boulder, Colorado, who was not a part of the study.

The lake resembles one of the interconnected pools that sit under several kilometers of ice in Greenland and Antarctica, says Martin Siegert, a geophysicist at Imperial College London, who heads a consortium trying to drill into Lake Ellsworth under West Antarctica. But the processes that gave rise to a deep lake on Mars are likely to be different. "It will open up a very interesting area of science on Mars," he says.

Water is thought to have flowed across the surface of Mars billions of years ago, when its atmosphere was thicker and warmer, cutting gullies and channels that are still visible. But today, low atmospheric pressures mean that any surface water would boil away. Water survives frozen in polar ice caps and in subsurface ice deposits. Some deposits have been mapped by the Mars Advanced Radar for Subsurface and Ionospheric Sounding (MARSIS), an instrument on the European Space Agency's Mars Express orbiter, which launched in 2003. MARSIS beams down pulses of radio waves and listens for reflections. Some of the waves bounce off the surface, but others penetrate up to 3 kilometers and can be reflected by sharp transitions in the buried layers, such as going from ice to rock.

Several years into the mission, MARSIS scientists began to see small, bright echoes under the south polar ice cap—so bright that the reflection could indicate not just rock underlying the ice, but liquid water. The researchers doubted the signal was real, however, because it appeared in some orbital passes but not others.

Later the team realized that the spacecraft's computer was averaging across pixels to reduce the size of large data streams—and in the process, smoothing away the bright anomalies. "We were not seeing the thing that was right under our noses," says Roberto Orosei, a principal investigator (PI) for MARSIS at the Italian National Institute for Astrophysics in Bologna.

To bypass this problem, the team commandeered a memory chip on Mars Express to store raw data during short passes over intriguing areas. Between 2012 and 2015, the spacecraft confirmed the existence of the bright reflections during 29 passes over the south polar region. The brightest patch, offset 9° from the pole, lies 1.5 kilometers under the ice and spans 20 kilometers, Orosei and his colleagues report.

The radar brightness alone isn't enough to prove that liquid water is responsible. Another clue comes from the permittivity of the reflecting material: its ability to store energy in an electric field. Water has a higher permittivity than rock and ice. Calculating permittivity requires knowing the signal power reflected by the bright patch, something the researchers could only estimate. But they find the permittivity of the patch to be higher than anywhere else on Mars—and comparable to the subglacial lakes on Earth. Although the team cannot measure the thickness of the water layer, Orosei says it is much more than a thin film.

Not everyone on the MARSIS team is convinced. "I would say the interpretation is plausible, but it's not quite a slam dunk yet," says Jeffrey Plaut, the other MARSIS PI at NASA's Jet Propulsion Laboratory in Pasadena, California, who is not an author on the study.

After all, it isn't easy to explain the presence of water at Mars's south pole. In Earth's polar regions, the pressure of the overlying ice lowers its melting point, and geothermal heat warms it from below to create the subglacial lakes. But there's little heat flowing from the geologically dead interior of Mars, and under the planet's weak gravity, the weight of 1.5 kilometers of ice does not lower the melting point by much. Orosei suspects that salts, especially the perchlorates that have been found in the planet's soils, could be lowering the ice's melting point. "They are the prime suspects," he says.

High levels of salt and temperatures dozens of degrees below zero do not bode well for any microbes trying to live there, Stillman says. "If martian life is like Earth life, this is too cold and too salty." But he says researchers will want to look for other lakes under the ice and find out whether they are connected—and whether they point to an even deeper water table.

Lakes might even turn up at lower, warmer latitudes—a location more suitable for a martian microbe, says Valèrie Ciarletti of the University of Paris-Saclay, who is developing a radar instrument for Europe's ExoMars rover, due to launch in 2020. "The big, big finding would be water at depth outside the polar cap." ■

ELECTRICAL ENGINEERING

Chipmakers look past Moore's law, and silicon

Novel computer chip materials explored in $1.5 billion U.S. military research program

By **Robert F. Service**

Silicon computer chips have been on a roll for half a century, getting ever more powerful. But the pace of innovation is slowing. This week, the U.S. military's Defense Advanced Research Projects Agency (DARPA) announced dozens of new grants totaling $75 million in a program for academic and industry scientists. The program, called the Electronics Resurgence Initiative, aims to spur new chip designs and materials, such as carbon nanotubes. Over the next few years, the program will grow to $300 million per year, up to a total of $1.5 billion over 5 years.

"It's a critical time to do this," says Erica Fuchs, a computer science policy expert at Carnegie Mellon University in Pittsburgh, Pennsylvania. "It's a good first step."

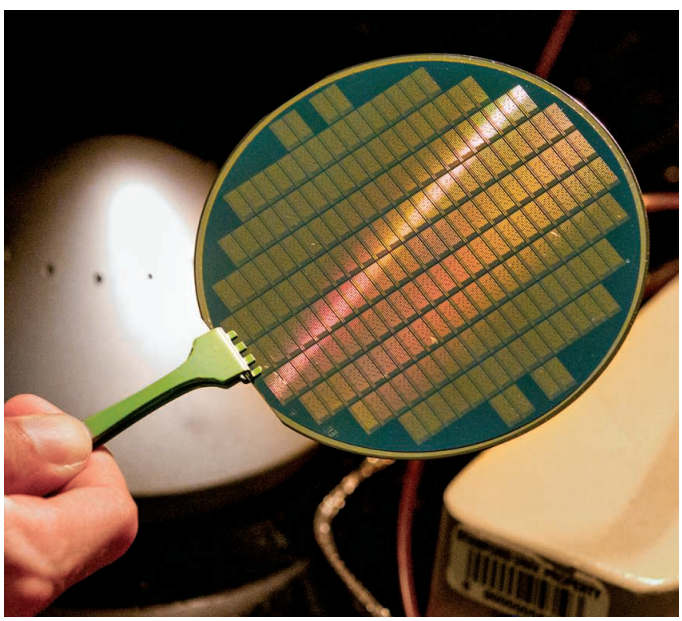
In 1965, Intel co-founder Gordon Moore made the observation that would become his eponymous "law": The number of transistors on chips was doubling every 2 years, a time frame later cut to every 18 months. But the gains from miniaturizing the chips are dwindling. Today, chip speeds are stuck in place, and each new generation of chips brings only a 30% improvement in energy efficiency, says Max Shulaker, an electrical engineer at the Massachusetts Institute of Technology in Cambridge. Fabricators are approaching physical limits of silicon, says Gregory Wright, a wireless communications expert at Nokia Bell Labs in Holmdel, New Jersey. Electrons are confined to patches of silicon just 100 atoms wide, he says, forcing complex designs that prevent electrons from leaking out and causing errors. "We're running out of room."

Moreover, only a handful of companies can afford the multibillion-dollar fabrication plants that make the chips, stifling innovation in a field once dominated by small startups, says Valeria Bertacco, a computer scientist at the University of Michigan in Ann Arbor. And some big companies are going down separate paths, designing specialized chips for specific tasks, Fuchs says. That has reduced the incentive for them to pay for shared, precompetitive basic research. The number of companies involved in the Semiconductor Research Corporation in Durham, North Carolina, which backs such work, dropped from 80 in 1996 to less than half that in 2013, according to a study by Fuchs and her colleagues.

DARPA is now trying to fill the gap, with grants to researchers such as Shulaker. He is fashioning 3D computer chips with transistors made of carbon nanotubes, which switch on and off faster and more efficiently than silicon transistors. Companies today already make 3D chips with silicon as a way to pack logic and memory functions closer together to speed up processing. But the chips are slowed down by bulky and sparse wiring that carries information between the chip layers. And because 2D silicon chip layers must be fabricated separately at more than 1000°C, there is no way to build up 3D chips in an integrated fabrication plan without melting the lower layers.

Carbon nanotube transistors, which can be made at nearly room temperature, offer a better path to dense, integrated 3D chips, Shulaker says. Even though his team's 3D chips will have features 10 times larger than state-of-the-art silicon devices, their speed and energy efficiency is expected to be 50 times better—a potential boon for power-hungry data centers.

The DARPA program is also supporting research into flexible chip architectures. Daniel Bliss, a wireless communications expert at Arizona State University in Tempe, and his colleagues want to improve wireless communications with chips that can be reconfigured on the fly to carry out specialized tasks. Bliss is working on radio chips that mix and filter signals using software rather than hardware—an advance that would allow a larger number of devices to transmit and receive signals without interference. This could improve mobile and satellite communications, as well as enable a rapid growth in the "internet of things," where myriad devices communicate with one another, he says.

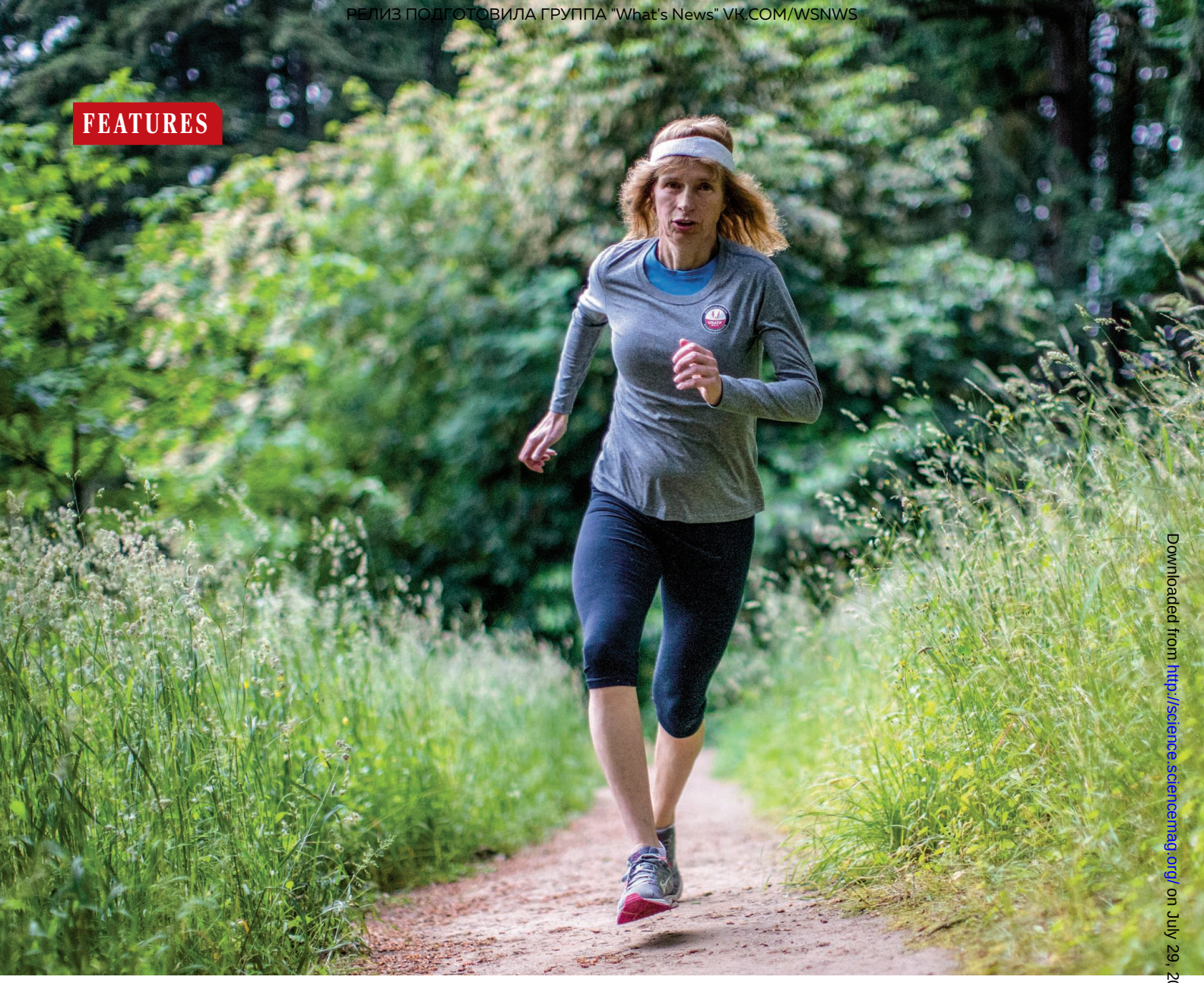


A wafer contains hundreds of computer chips made from carbon nanotubes, which switch on and off faster and more efficiently than transistors made from silicon.

PHOTO: STANFORD ENGINEERING

Another DARPA grant, for researchers at Stanford University in Palo Alto, California, will go to improving computer tools used in chipmaking. These tools verify novel chip designs with a form of artificial intelligence called machine learning. They would help automate the largely manual process of detecting design bugs in chips made up of billions of transistors, and could speed up the ability of companies to design, test, and fabricate new chip architectures.

If even a fraction of the new projects succeed, the DARPA project "will completely revolutionize how we design electronics," says Subhasish Mitra, a Stanford electrical and computer engineer, and a researcher on the 3D carbon nanotube and circuit validation projects. He says it will also encourage engineers to look beyond silicon, which has dominated research for decades. "When I was a student, life was boring," Mitra says. "It was clear that silicon would move forward along a known path. Now, it's absolutely clear that's not what the future is." ■



THE TRAILBLAZER

Joanna Harper is racing to reveal how a gender transition alters an athlete's physiology and performance

By **Katherine Kornei**, *in Portland, Oregon, and Phoenix*

PHOTO: BETH NAKAMURA

Published by AAAS

Joanna Harper's personal journey and love of running have shaped her research interests.

Joanna Harper swallowed a few pills in late August 2004, a day after running in the Hood to Coast relay race between Oregon's highest mountain and the Pacific Ocean. They delivered a small dose of estrogen and a testosterone blocker and set in motion changes that Harper, who was designated male at birth and raised as a boy, had imagined since childhood. Harper's timing was deliberate—the 47-year-old nationally ranked runner wanted one more race before disrupting her hormones because she knew she'd never run as fast again.

The testosterone that courses through a man's body after puberty triggers and maintains a slew of physical changes: Men, whose levels of the hormone are usually some 10 to 15 times those of women, typically have larger muscles, denser bones, and higher fractions of lean body mass than women. That hormone-fueled transformation confers certain athletic advantages, and men on average run faster, lift more weight, and throw harder and farther than women. Sporting events are therefore usually split into male and female categories to ensure fair competition. But this division of the sexes, which has existed for as long as women have competed as athletes, forces an important question: Who, at least from an athletic standpoint, is female?

Many people believe transgender women such as Harper have athletic advantages over non-transgender women—sometimes called cisgender women—because of their previous exposure to male levels of testosterone. But Harper, a medical physicist at a large medical center in Portland, Oregon, has been challenging that assumption with data. In 2015, she published the first study of transgender athletes' performances, finding that transgender women who received treatment to lower their testosterone levels did no better in a variety of races against female peers than they had previously done against male runners. Although Harper's study included only a few transgender women, Eric Vilain of The George Washington University in Washington, D.C., a geneticist who specializes in gender-based biology, calls it "groundbreaking."

That work helped make Harper an unpaid adviser for sporting bodies, such as the International Olympic Committee (IOC), that are wrestling with transgender issues and other matters of gender. Although Harper has just a master's degree, she is helping spearhead several studies documenting how the physiology and performance of transgender athletes change as they make their transition.

Harper may not have the traditional pedigree of a scientist, Vilain says, but "her approach is highly respected." Harper has made "very important" contributions to understanding gender and sports performance, anticipating the debate that now swirls around transgender and "intersex" athletes, adds Stéphane Bermon, a sports physician and exercise physiologist at the Monaco Institute of Sports Medicine and Surgery.

IN THE 1960s, elite female athletes had to prove their sex by showing sporting officials their genitalia. Those unpopular "nude parades" were soon replaced by chromosomal tests, which had their own problems. Biology does not always cleanly divide human beings into two sexes. Some people, often described as intersex, have unusual sex chromosome arrangements or physical attributes such as ambiguous genitalia. Others have an innate sense of gender differing from the sex they were assigned at birth; they often choose to hormonally and/or surgically transition to the gender they identify as.

Harper, who grew up in Parry Sound, a small town about 250 kilometers north of Toronto, Canada, is among the latter. The oldest child of a high school teacher and a nurse, Harper wore her mother's and sister's clothing in private from a young age. Leaving elementary school one day, Harper asked a male friend an earnest question—had he ever thought, as Harper had, about being a girl? The boy recoiled. "It was clear from the way that my friend looked at me that I was never going to ask anybody that again," Harper says.

In high school, Harper threw herself into sports. She was best known as a star basketball player, averaging 18 points per game as a guard on the boys' team her senior year. But Harper was short for the men's college game. Instead, she ran track and cross-country as an undergraduate at what was then the University of Western Ontario in London, Canada, where she majored in physics. In 1982, Harper earned a master's degree in medical physics, training that prepared her to determine radiation treatment dosages for cancer patients. The career choice was, in part, lifestyle-oriented. "It was clear medical physicists were not putting in huge hours, and it would allow me time to run," she says.

Harper moved to Oregon a few years later to be closer to family in the Canadian province of British Columbia and to take a job in Portland. But her gender dysphoria always lurked. By 2004, "I was very close to a mental breakdown," she says. She started to see a therapist, who asked her to list what she'd lose if she transitioned genders. Near the top of that list was competitive sports—transgender athletes had no official opportunities to compete at the time.

Barely 3 months later, in May 2004, IOC announced a landmark ruling: Transgender athletes could participate in the Olympics. Strict stipulations were in effect for transgender women competing in women's events—they needed sex reassignment surgery, legal recognition of their gender by "the appropriate official authorities," and at least 2 years of verifiable hormone replacement therapy—but it was "a light at the end of the tunnel" for Harper. She wasn't an Olympic-level athlete, but she believed that change would trickle down to other sports organizations and she would be able to compete again.

Harper started hormone replacement therapy in August 2004. Spironolactone, a testosterone blocker and diuretic, flushed the testosterone her body was producing, and estradiol, a form of estrogen, began to make her form more typically feminine. In just a few weeks, Harper noticed changes such as breast tenderness and a decrease in body hair. The transition was a "very, very,

very difficult time," Harper says. She lost most of her male friends, and her mother—her only living immediate family member—was not supportive. "When it became clear that I was going to go through with my transition, my mother said she never wanted to see me again," Harper recalls. Barbara Harper, who died in 2013, eventually relented, in a way. "When I visited, she didn't tell anyone who I was." "This is Joanna," was all her mother would say.

Today, the 61-year-old with collarbone-length red hair describes herself as a scientist, an athlete, and a transgender person—in that order. But being transgender can sometimes overshadow everything else. "We joke in the transgender world that if Hitler had been transgender, he would be described as 'that transgender dictator.' That's the category that everybody wants to put you in."

IN 1976, RENÉE RICHARDS entered a New Jersey professional tennis tournament. Richards, who had enrolled at Yale University as Richard Raskind and captained its men's tennis team before undergoing sex reassignment surgery, was met with open hostility—more than 20 female players boycotted the competition in protest of her perceived advantages. Indeed, some physical attributes such as hand size and height—like Richards's tall frame—remain largely unchanged after hormone therapy, Harper says.

Many people expect other physical advantages to linger, too. Men generally have blood with higher oxygen-carrying capacity because testosterone stimulates bone marrow to produce more red blood cells, says Siddhartha Angadi, a cardiovascular physiologist at Arizona State University in Phoenix. Male bodies are also generally leaner, and carry less body fat—"an obvious benefit when it comes to athletic performance," Angadi says.

Some people therefore insist that transgender women and many intersex athletes competing in women's events will always have an unfair edge. (Little controversy exists over transgender men in sports, as many expect them to be at a disadvantage.) Others believe athletes should be able to compete in their self-identified gender without regulations. Harper wants to address the question with data. "You have to go to science."

Before her own transition in 2004, Harper expected that her 10,000-meter race time might increase by "a minute or two" as her testosterone level dropped and she slowed. But in less than a year, Harper was running a full 5 minutes slower than her personal best. "It just blew me away, and it very much piqued my interest as a scientist."

In 2005, Harper realized her experience wasn't unique after reading an article in *Runner's World* about another transgender female runner who had also become significantly slower. But when Harper searched for studies about the physiology of transitioning, she found none. So on nights and weekends, she began to moonlight on a research project.

Harper searched for transgender female distance runners willing to share race times from before and after their transitions. The transgender population, even now, tends to be "small and secretive," Harper says, and it took 7 years of contacting athletes through Yahoo and Facebook groups to collect data from eight runners. All the women had undergone hormone therapy to bring their testosterone levels in line with typical female levels. In Harper's study, titled simply "Race Times for Transgender Athletes" and published in 2015 in the little-known *Journal of Sporting Cultures and Identities*, she showed that all but one person ran substantially slower after transitioning.

Harper also calculated each subject's age grade, a common metric in track and field and distance running that reflects an athlete's performance compared with the fastest known time by someone of the same age and sex. Harper showed that the athletes' age grades before and after hormone therapy remained nearly the same. That is, the women were as competitive with their age- and sex-matched peers as they had been when competing against men. They weren't, in other words, likely to dominate women's races. "No one had previously looked at actual performance of transgender athletes pre- and posttransition," Vilain says.

Harper has since shown similar results for a transgender rower, a cyclist, and a sprinter. Together, the findings make a case that previous exposure to male levels of testosterone does not confer an enduring athletic advantage.

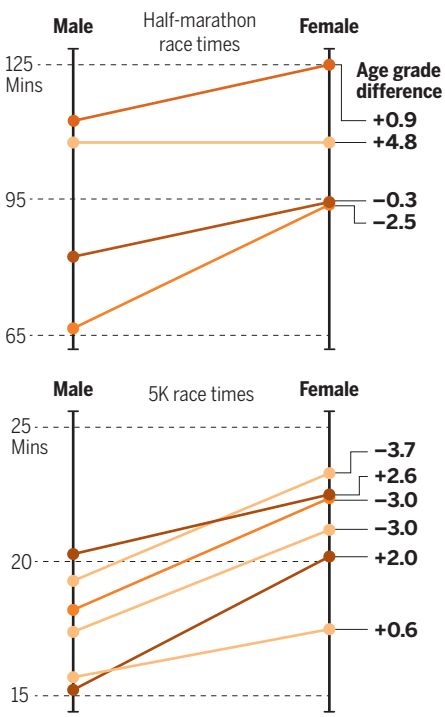
In 2015, IOC invited Harper to attend its Consensus Meeting on Sex Reassignment and Hyperandrogenism held in Lausanne, Switzerland. After 3 days, the panel of scientists and physicians converged on revised rules for transgender competitors, including at least 1 year of hormone replacement therapy for female competitors, rather than the 2 years previously required. That change was a nod to Harper's personal transition experience and to research published in 2004 in the *European Journal of Endocrinology* showing that the testosterone levels—and therefore performance—of 19 transgender women stabilized after 12 months of hormone therapy. The revised IOC policy also lifted the requirement for sex reassignment surgery. That decision was a long time coming, Harper says. "What your genitals are doesn't make a difference."

Less settled, however, is the debate about the appropriate upper limit of women's testosterone levels in elite athletic competition. The current IOC policy dictates that transgender women must have a testosterone level less than 10 nanomoles per liter, roughly the low end of typical male values. But because more than 99% of women have testosterone levels less than 3 nanomoles per liter, some researchers have suggested that limit is too high. Harper is among them. "If you're competing in the women's division, you should do so with women's hormone levels," she says. "I understand just how much difference they make."

South African runner Caster Semenya, who has always competed in women's races and won Olympic gold in Rio de Janeiro, Brazil, in 2016, recently refocused attention on the testosterone issue. In 2009, the International Association of Athletics Federations (IAAF)—the Stockholm-based, world-governing body for track and field—controversially required her to take a sex-verification test after she breezed past competitors in the 800-meter race at the IAAF World Championships. The results, leaked during the competition, allegedly

Different, but same

A study of transgender women found their race times slowed after transitioning, but their age grades, which compare people to the best runners of the same sex and age, hardly changed, suggesting they have no advantage over non-transgender women.



CREDITS: (GRAPHIC) N. DESAI/*SCIENCE*; (DATA) J. HARPER, *JOURNAL OF SPORTING CULTURES AND IDENTITIES* 6, 1 (2015)

revealed that Semenya was intersex and had three times the testosterone of a typical woman. Neither she nor IAAF has ever confirmed that publicly, however.

In April, IAAF issued a policy that many groups—including Athletics South Africa, the country's athletics federation—view as targeting Semenya. It applies to women competing in certain track and field events, including those that Semenya excels at, who have specific intersex conditions in which their bodies produce and are sensitive to higher levels of testosterone. Drawing on performance and hormone data from an IAAF-sponsored study of athletes competing at its recent World Championships, which was published last year in the *British Journal of Sports Medicine*, the policy requires testosterone levels below 5 nanomoles per liter, half the previous IAAF threshold. The goal is "leveling the playing field to ensure fair and meaningful competition," IAAF President Sebastian Coe said in a statement.

In June, Semenya called that policy discriminatory and said she would challenge it at the Court of Arbitration for Sport. "I just want to run naturally, the way I was born," the 27-year-old told media. "I am a woman and I am fast." And earlier this month, three researchers who analyzed a subset of data from the 2017 study claimed to have found "significant problems and anomalies" and called for its retraction. IAAF updated the paper, but told *The New York Times*, which first reported the flap, that "the conclusions remain the same."

More controversy may be on the way: Within the next few months, IAAF is expected to issue updated testosterone-based regulations for transgender women as well. IOC also plans to announce new testosterone limits for athletes in women's events, which will be in effect for the 2020 Summer Olympics in Tokyo.

Caster Semenya's impressive victories in women's races have ignited discussions on testosterone limits in athletes.

PHOTO: ANDY ASTFALCK/GETTY IMAGES

TO GET A CLEARER PICTURE of how changing hormone levels affect an athlete's body, Harper and others want to collect data from people during their transitions. "It's important to know more about lean body mass; hemoglobin concentration; and psychological, endocrine, and metabolic changes during transition," says Bermon, who is also a scientific consultant for IAAF.

Harper recruited the athletes for the first such studies. One focuses on a 28-year-old distance runner named Lauren. (The athletes undergoing testing requested that only their first names be used in this story.) Roughly once a month, Lauren makes the 45-minute drive from her home in Queen Creek, Arizona, to downtown Phoenix to undergo a battery of tests in Angadi's lab.

He and his team have been measuring, among other things, the elasticity of Lauren's arteries, her bone density and distribution of fat, and how the myocardial fibers of her heart twist and untwist with each beat. After she started estrogen injections last fall, Lauren's blood pressure dropped by about 10 points and her body fat increased, Angadi says. Those changes are small, he cautions, and more measurements are needed. Another year of data may reveal a decline in bone density. "Bone is a really slow-turnover organ," he says.

During her visit in April, Lauren pulled on a bright blue mask that covered her nose and mouth, and Andrew D'Lugos, an exercise physiologist, had her start to run on a treadmill. "Way to go, Lauren!" "Enjoy the flat!" and "Looking good!" he and other scientists called out. D'Lugos cranked the treadmill up to 13, then 14 kilometers per hour. A large screen displayed Lauren's heart rate and oxygen consumption, new points popping up every 15 seconds. D'Lugos gradually increased the treadmill's incline, effectively forcing Lauren to run up an ever-steeper hill at her marathon pace. When at last she could go no farther, she pushed down on the treadmill's side railings, lifting her thin frame above the spinning black belt, her head bent in exhaustion.

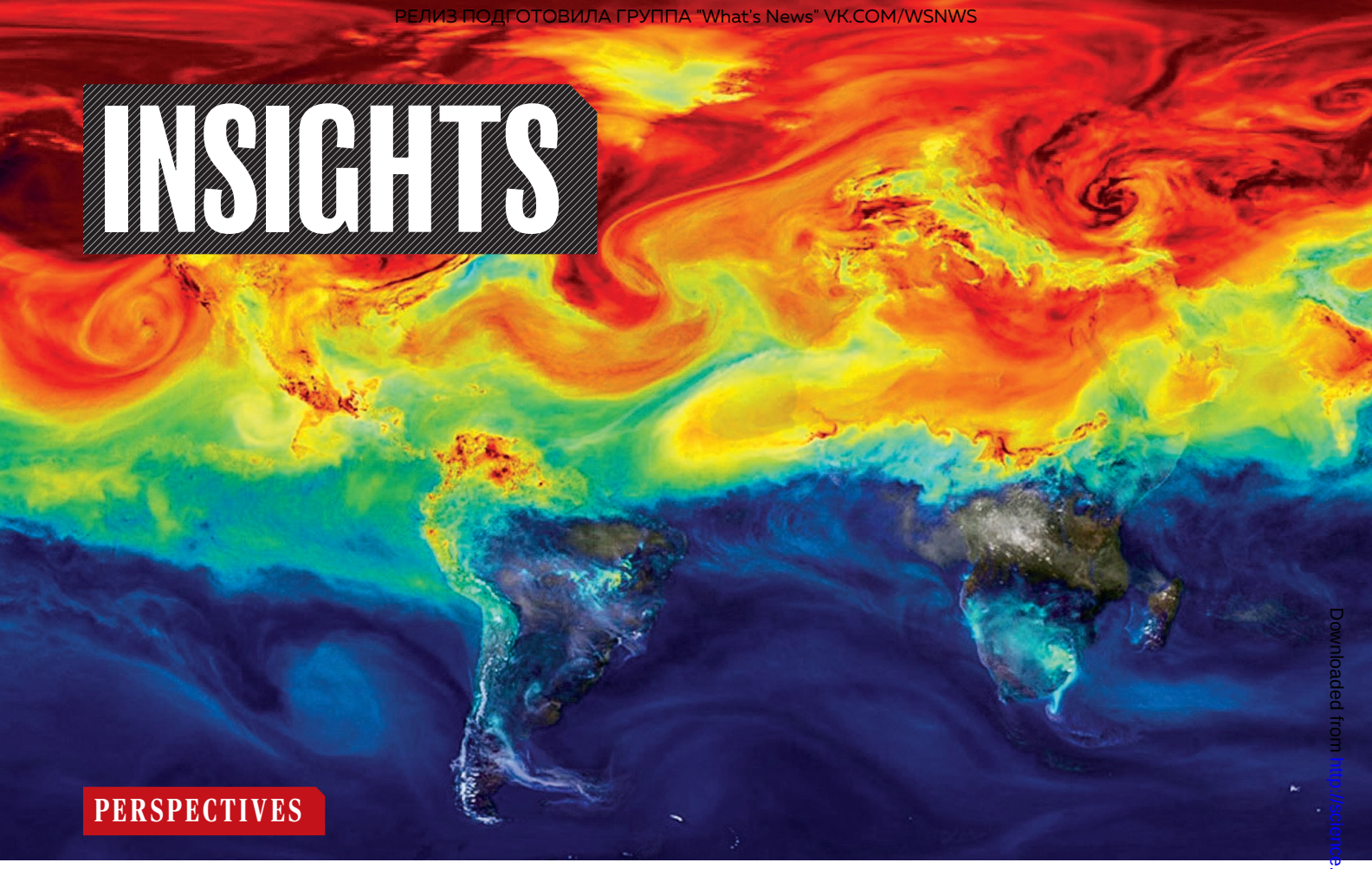
Within 6 weeks after she started estrogen injections, Lauren's peak oxygen consumption rate—a measure of fitness—fell by 17%, the researchers reported at the American College of Sports Medicine's meeting in Minneapolis, Minnesota, in June. "She lost a fair bit of performance really quickly," Angadi says. But Lauren's fitness probably hasn't stabilized yet, he notes, because she only recently started to take a testosterone blocker. Angadi's team will test Lauren into 2019, when she runs the Boston Marathon. Charissa, a triathlete living in Colorado who is taking part in a similar study, lost roughly 15% of her aerobic capacity in 9 months since beginning hormone therapy, Harper reported in March at a British Association of Sport and Exercise Medicine conference held in Doncaster, U.K.

There, Harper announced that she and Yannis Pitsiladis—an exercise physiologist at the University of Brighton in the United Kingdom best known for his so-far futile efforts to train a man to run a marathon in less than 2 hours—plan to monitor roughly 20 men and 20 women as they transition. The largest study of its kind, it will recruit subjects from a London-based gender clinic and enlist the expertise of endocrinologists, muscle physiologists, and mental health professionals, among others. The undertaking will be enormous, Harper admits. "We're going to need help," she told the Doncaster audience. Pending research funding from the U.K. government, the study will begin in 2019.

Harper still has her day job, and she's writing a book about gender variance in sports. Fortunately, a refuge from all the demands on her time—and the controversies that come with her research—is just a block and a half from her front door. Several times a week, Harper turns left from her house toward Mount Tabor Park, one of Portland's largest. There, she runs 95 kilometers a week, often with friends, on trails that weave among 100-year-old cedar, spruce, and redwood trees. Just like the paths that sometimes open up to reveal fleeting views of Mount Hood in the distance, Harper's life has had its share of twists and turns. But she doesn't regret the decision she made 14 years ago. "I became a much happier person." ■

Katherine Kornei is a science journalist in Portland, Oregon.

INSIGHTS



PERSPECTIVES

CLIMATE

Reducing uncertainties in climate models

Implementing accurate calculations of radiative forcing can improve climate projections

By **Brian J. Soden[1], William D. Collins[2,3], Daniel R. Feldman[2]**

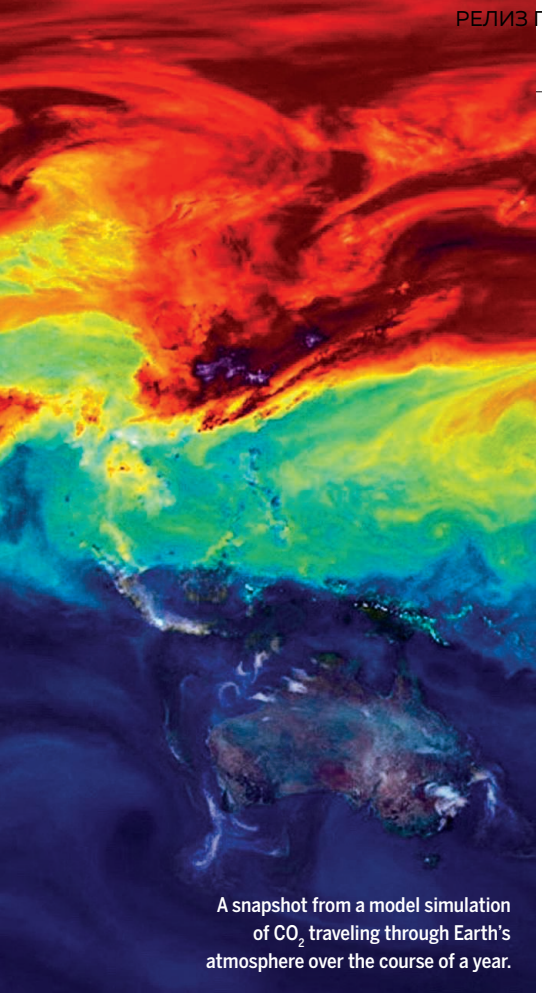
Radiative forcing is a fundamental quantity for understanding both anthropogenic and natural changes in climate. It measures the extent to which human activities [such as the emission of carbon dioxide (CO_2), see the image] and natural events (such as volcanic eruptions) perturb the flow of energy into and out of the climate system. This perturbation initiates all other changes of the climate in response to external forcings. Inconsistencies in the calculation of radiative forcing by CO_2 introduce uncertainties in model projections of climate change, a problem that has persisted for more than two decades. The explicit calculation of radiative forcing and a careful vetting of radiative transfer parameterizations provide a straightforward means to substantially reduce these uncertainties and improve the projections.

CO_2 is the main forcing agent in both 20th- and 21st-century emission scenarios (*1*). Twenty-five years ago, Cess *et al.* provided the first comprehensive assessment of the calculation of radiative forcing by CO_2 in global climate models (GCMs) (*2*). They found that when CO_2 was doubled, the radiative forcing differed substantially among 15 different GCMs, ranging from ~3.3 to 4.7 W/m^2 (see the graph; see the supplementary materials for further details). This spread mainly arose from intermodel differences in the parameterization of infrared absorption by CO_2. Other sources of differences, such as the parameterization of overlapping absorption by water vapor or differences in the cloud distributions, were shown to be small.

Thirteen years later, Collins *et al.* conducted a more extensive intercomparison of radiative forcing, using a newer generation of more than 20 different GCMs (*3*). They found a similar range in radiative forcing at the top of the atmosphere for a doubling of CO_2 (see the graph), which again was largely due to spread in the infrared component of CO_2 absorption. The authors also compared the radiative forcing computed using line-by-line (LBL) calculations; the latter solve the equation of radiative transfer for each absorption line individually, rather than parameterizing their absorption over spectrally integrated bands. The forcing calculations between several different LBL models were in much better agreement (see the graph). The LBL calculations have also been extensively validated by using both laboratory and field measurements (*4*), and the spectroscopic foundation for this radiative forcing is quite robust (*5*). The agreement among LBL models forms the basis for the narrow uncertainty range for CO_2 forcing noted in the Intergovernmental Panel on Climate Change (IPCC) reports (*1*). However, LBL calculations

[1]*Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33149, USA.* [2]*Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.* [3]*Department of Earth and Planetary Science, University of California at Berkeley, Berkeley, CA 94720, USA. Email: b.soden@miami.edu*

IMAGE: NASA'S GODDARD SPACE FLIGHT CENTER



A snapshot from a model simulation of CO_2 traveling through Earth's atmosphere over the course of a year.

GRAPHIC: ADAPTED BY N. CARY/*SCIENCE*

are computationally expensive, and parameterized models of radiative transfer must be used in GCMs. Unfortunately, substantial differences still exist in these parameterizations. Chung and Soden found that the spread in CO_2 forcing from the most recent generation of GCMs remains largely unchanged compared with that documented in previous generations (see the graph) (*6*).

The precise measure of radiative forcing differs slightly between these three studies (*7*). As a result, their absolute values of radiative forcing are not directly comparable. However, the relative spread in radiative forcing between models is meaningful and has shown little change compared with the true uncertainty in radiative transfer, as represented by the spread in the LBL calculations.

This lack of progress over the past 25 years is disconcerting. The spread in model calculations of CO_2 forcing does not represent an uncertainty in radiative transfer theory, but rather the failure to implement that theory consistently in radiative transfer parameterizations. This introduces unnecessary noise into the model experiments that is difficult to remove. Although the users of these models are largely unaware of this ongoing problem, the unsatisfactory implementation of CO_2 forcing propagates needlessly onto efforts to reduce uncertainty in projections of future climate change.

As noted by Cess *et al.*, the impact of this inconsistency in the calculations of radiative forcing on estimates of climate sensitivity "is nearly half of the often quoted range of uncertainty of 1.5° to 4.5°C." Thus, even if we could make all other aspects of the models perfect, the spread in projections of CO_2-induced climate change would only be reduced by 50% because of the remaining differences in radiative forcing.

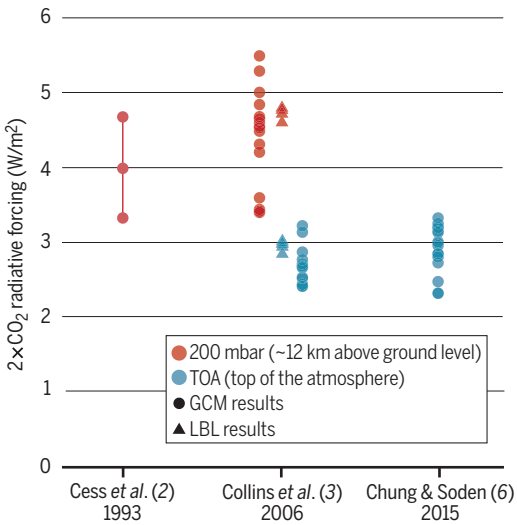
The contributions of erroneous CO_2 forcing to the persistent spread in climate projections undermines the utility of these models to answer fundamental questions of central societal importance. These errors add unnecessary confusion to the development of scientifically rigorous targets for atmospheric CO_2 concentrations—and therefore, emissions reductions—that are required to limit global temperature change. Constraining global warming to less than 2°C, as set by the Paris Climate Agreement, requires a limit to be set on the maximum globally averaged CO_2 concentration compatible with that constraint. This limit should be established by a multimodel ensemble, but the corresponding range of allowable CO_2 concentrations is unnecessarily large because the ensemble does not consistently incorporate known and established physics that relate rising CO_2 concentrations to radiative forcing.

Although some efforts are under way to better document these differences (*8*), there are two immediate solutions that could help. First, it is essential that radiative forcing be routinely computed and reported for models that participate in Coupled Model Intercomparison Projects (CMIP), a series of coordinated experiments performed in support of the IPCC assessments. For each experiment, model simulations are performed by using matching emission scenarios, with the intent of imposing identical forcings. However, radiative forcing is rarely reported explicitly by these models. Requiring models to do so for all emission scenarios would help to ensure transparency between the radiative forcing experienced by the models and the climate response that results. Cess *et al.* made a similar recommendation 25 years ago (*2*). The adoption of this recommendation is long overdue.

Second, the diversity of radiative transfer parameterizations used in GCMs should be reduced. Maintaining diversity in models is valuable for areas where there is substantial uncertainty in the underlying physics. For most aspects of radiative transfer of relevance to climate change, this is not the case. An effort to consolidate the number of radiative transfer parameterizations used and to implement only those that have been thoroughly vetted against LBL calculations would significantly reduce the spread in model projections. It would also reduce discrepancies in the parameterization of other key absorbers, such as water vapor, that also affect model calculations of climate sensitivity. Last, it would enable those researchers who focus on less well-known forcing agents and their radiative interactions to have a readily available, radiometrically accurate understanding of the direct radiative influence of the quantities they are measuring, and the processes they are studying, on Earth's climate system. ■

Reducing the uncertainty

Radiative forcing uncertainty in GCMs has remained high over the past 25 years. LBL calculations show that this uncertainty can be substantially reduced.



REFERENCES AND NOTES

1. G. Myhre *et al.*, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker *et al.*, Eds. (Cambridge Univ. Press, 2013), pp. 659–740
2. R. D. Cess *et al.*, *Science* **262**, 1252 (1993).
3. W. D. Collins *et al.*, *J. Geophys. Res.* **111**, D14317 (2006).
4. D. R. Feldman *et al.*, *Nature* **519**, 339 (2015).
5. M. G. Mlynczak *et al.*, *Geophys. Res. Lett.* **43**, 5318 (2016).
6. E.-S. Chung, B. J. Soden, *Environ. Res. Lett.* **10**, 074004 (2015).
7. The Cess *et al.* and Chung and Soden calculations represent global averages; however, the radiative forcing from Cess *et al.* is computed at 200 mbar, whereas those for Chung and Soden are computed at the top-of-atmosphere (TOA). Collins *et al.* performed calculations for both 200 mbar and the TOA; however, their results are for a standard mid-latitude summer profile under cloud-free conditions rather than global averages.
8. R. Pincus, P. M. Forster, B. Stevens, *Geosci. Model Dev.* **9**, 3447 (2016).

ACKNOWLEDGMENTS

Some of this material is based on work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Atmospheric System Research Program, and the U.S. National Aeronautics and Space Administration, CloudSat, and CALIPSO Science Team.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/326/suppl/DC1

10.1126/science.aau1864

NEUROSCIENCE

Neurotechnology to address big questions

Profiling of single neurons in tissue allows structure and function linkage in brain circuits

By **Thomas Knöpfel**

Emotions, cognition, and consciousness emerge with the processing of neuronal information provided by memory readouts and senses. How this actually works seemed for a long time to be a question too big to ask. Advances in neuroscience have always been driven by methodological inventions, and current efforts to develop neurotechnologies are motivated by experimental strategies, including analytical dissection (that is, inverse engineering), large-scale interrogation, and synthetic reconstruction of the mammalian cerebral cortex (which is important for higher cognitive functions) and connected brain structures (*1*, *2*). As part of this neurotechnological endeavor, mapping the transcriptome of neurons with single-cell resolution and with known three-dimensional tissue localization has been a long-sought enabling technology (*3*). On page 380 of this issue, Wang *et al.* (*4*) present the technology and workflow to access transcriptional states of more than 100 genes from up to 30,000 cells in a cubic millimeter of cortical tissue. This advanced methodology will facilitate studies that improve our understanding of the neuronal hardware and, when combined with other emerging neurotechnologies, will enable big questions in neuroscience to be addressed.

Wang *et al.* integrated knowledge of mouse cortex-specific gene expression, targeted signal amplification, in situ transcriptomics, and hydrogel-tissue chemistry (*5*) into a workflow for single-cell resolution spatial transcriptomics, called STARmap (spatially resolved transcript amplicon readout mapping). Targeted signal amplification was achieved by generating circular complementary DNAs (cDNAs) after recognition of a specific transcript [messenger RNA (mRNA)] by a probe from a library of gene-specific primers. These cDNAs were then rolling-circle-amplified so as to generate a DNA nanoball (amplicon) that contains multiple copies of the cDNA templates (*6*). Importantly, this process takes place in structurally conserved brain tissue. Formation of the circular cDNAs included a barcode as a distinct identifier, which was decoded with a newly designed error-reduced in situ sequencing technology to identify specific mRNAs (*5*).

Since the development of techniques that involve mRNA harvesting from single cells for sequencing—single-cell reverse transcription–polymerase chain reaction (scRT-PCR)—transcriptomics has been instrumental to classify the many functionally distinct cell types found in the cerebral cortex (*7*). scRT-PCR in combination with electrophysiological recordings has the advantage to directly link specific gene expression patterns and cellular functional characteristics (including information on how specific neurons integrate synaptic information and on their synaptic connectivity) but is not practical for more than 100 cells (*7*). Scaling this approach is required for a systematic analysis of cellular diversity and brain state–dependent transcription. STARmap can provide information about the transcriptome (up to 10,000 genes) of up to 30,000 cells in a cortical volume prepared from an individual mouse brain.

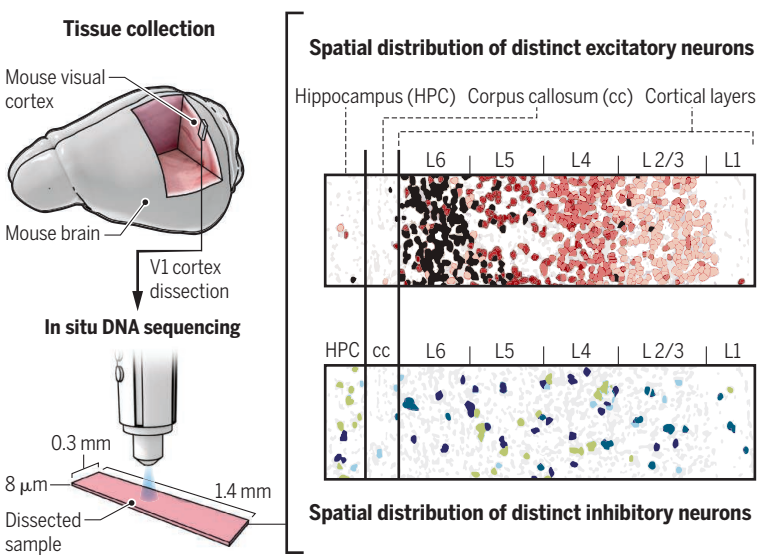
Hydrogel-tissue chemistry (*5*) was needed to fix the amplicons in the cubic millimeter tissue samples. Classification of cortical cell type based on anatomical location, gene expression pattern, and functional properties has been performed in the past by accumulating data across many experiments and many specimens (*7*). This knowledge served the authors to identify a set of 160 target genes, including 112 putative cell-type marker genes and 48 genes that are expressed in response to neuronal activity patterns. It is not known which of the selected marker genes are most indicative for specific cell types and which of the activity-regulated genes are recruited in which cell types and under specific patterns of neuronal activity. Therefore, Wang *et al.* used this set of target genes both to benchmark their methodology and workflow but also to contribute to efforts aimed at detecting and classifying cell types and corresponding tissue-organization principles in the neocortex of adult mouse brain (see the figure).

A weakness of STARmap is the lack of direct linking of transcriptomes with cellular protein expression and function. However, because several of the chosen marker genes encode proteins—including ion channels, receptors, and neurotransmitter-related enzymes that indicate functional properties—an indirect link between transcriptomics and function is already provided. In future implementations, a larger number of genes indicating functional properties may replace marker genes with low cell type–specificity to increase the possibilities for functional inference.

The high data content from a single tissue specimen provided by STARmap allows for correlation of gene activation patterns within an individual's specific brain state at the time

Profiling brain tissue

Wang *et al.* provide a method to determine the activity of marker genes within a sample of brain tissue. This allows identification and mapping, for example, of subtypes of excitatory or inhibitory neurons in the cortical layers, corpus callosum, and hippocampus.



GRAPHIC: C. BICKEL/*SCIENCE*

Department of Medicine and Centre for Neurotechnology, Imperial College London, UK. Email: tknopfel@knopfel-lab.net

of tissue preparation. As a proof of principle, the authors compared the transcriptome from two cohorts of mice: The first was exposed for 1 hour to light after 4 days of housing in darkness, whereas the second cohort was housed continuously in darkness before tissue collection. STARmap analysis confirmed and refined our knowledge about the recruitment of activity-dependent genes and their enhancer RNAs (which control transcription of their target genes). In the future, a larger range of physiological brain states as well as rodent models of brain diseases, and even human tissue, may be analyzed this way.

The STARmap data are a valuable source for data mining to identify patterns and higher-level structural and functional principles, many of which are likely still hidden behind the complex nature of cortical circuits and their function. The first applications of STARmap already revealed an intriguing finding: a short-range self-clustering organization of inhibitory neuron subtypes (*3*). These clusters may just be relicts of cell differentiation and cortical morphogenesis with little functional consequences, but more exciting are the possible functional implications. Among these is the increased opportunity of these inhibitory neuron subtypes to form direct electrical connections with cell bodies and the proximal processes of neighboring cells. These gap junctions are involved in the generation of high-frequency rhythmic circuit activities (*8*).

STARmap is well suited for further interfacing with other recent technologies that have been developed to address the big questions on brain functions. The technology closest in reach for combination with STARmap is functional in vivo optical imaging with activity indicators (*9*). This combination could be tremendously helpful for directly linking gene expression patterns with cell activity patterns and brain circuit functions. Another exciting possible application of STARmap could be the mapping of trans-synaptically distributed (*10*) and activity-dependent expressed barcoded optical activity reporters. This combination of cutting-edge technologies would allow efficient linking of cellular activity with connectivity, cell-type classification, and proteomic state. ■

REFERENCES

1. www.braininitiative.nih.gov
2. www.humanbrainproject.eu/en/follow-hbp/news/worlds-brain-initiatives-move-forward-together
3. E. Lein *et al.*, *Science* **358**, 64 (2017).
4. X. Wang *et al.*, *Science* **361**, eaat5691 (2018).
5. V. Gradinaru *et al.*, *Annu. Rev. Biophys.* **47**, 355 (2018).
6. R. Ke *et al.*, *Nat. Methods* **10**, 857 (2013).
7. J. R. Ecker *et al.*, *Neuron* **96**, 542 (2017).
8. B. W. Connors, *Dev. Neurobiol.* **77**, 610 (2017).
9. T. Knöpfel, *Nat. Rev. Neurosci.* **13**, 687 (2012).
10. S. Chatterjee *et al.*, *Nat. Neurosci.* **21**, 638 (2018).

10.1126/science.aau4705

GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

GENE EXPRESSION

Dynamic condensates activate transcription

Transcriptional components exhibit transient phase separation to drive gene activation

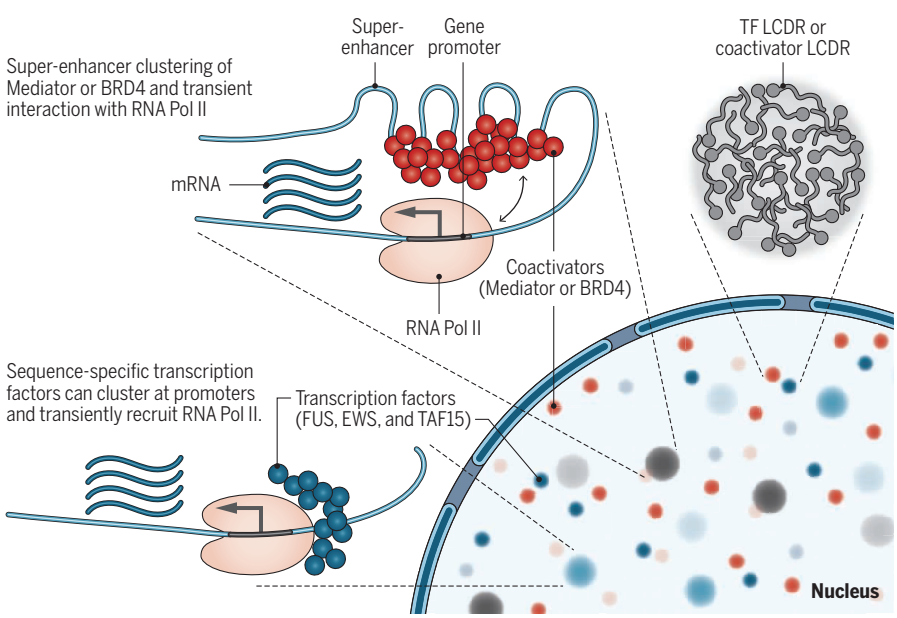
By **Aaron J. Plys**[1,2] *and* **Robert E. Kingston**[1,2]

Every aspect of human function, from proper cell differentiation and development to normal cellular maintenance, requires properly timed activation of the necessary genes. This requires transcription of genomic DNA into messenger RNA (mRNA), accomplished by RNA polymerase II (RNA Pol II), which initiates transcription at gene promoters. This highly regulated process requires hundreds of proteins that must go to the promoter in a coordinated manner. Although many of these proteins are already organized into large and stable protein complexes, and so travel as a group, the process still requires coordination of many individual proteins and preformed complexes so that they are all in the same place on genomic DNA at the same time. This problem has been appreciated for years and has led to models such as "transcription factories," where components are organized and ready to act on a gene that goes to the cellular location of the factory (*1*). On pages 378, 379, and 412 of this issue, Chong *et al.* (*2*), Sabari *et al.* (*3*), and Cho *et al.* (*4*), respectively, argue that special protein domains, which interact with each other to form fleeting or more persistent interactions, form biomolecular condensates that concentrate the transcription machinery. Some of these condensates might even form droplets, generating a liquid phase separated from the rest of the nucleus. Phase separation is a phenomenon familiar to anyone who has made a salad dressing: The oil and vinegar exist as two separated liquids. Phase separation in cells creates membraneless organelles that, in this case, provide the organization necessary for productive transcription (*5*).

Several distinct types of proteins are needed for transcriptional activation. Gene-specific transcription factors (TFs) bind to

Dynamic transcription machinery clustering during gene activation

Transcription factors (TFs) and coactivators condense into high-concentration clusters in the nucleus. Condensation is mediated by low-complexity disordered regions (LCDRs) in these proteins. These clusters can incorporate RNA Pol II through transient interactions to efficiently activate gene transcription.



specific regions of the genome. They then interact with large complexes needed for the transcription process, including a key complex called Mediator, which, in turn, interacts with RNA Pol II to increase transcription from a promoter (*6*). Enhancer sequences, which are physically removed from the promoter in the genome, also increase transcription and can be bound by other specialized proteins (*7*). Regions of proteins involved in these processes were analyzed in the three studies, with the notable finding that several different transcriptional regulatory components each contain protein domains that form condensates in cells. These condensates increase the effective concentration of components needed for transcriptional activation and allow organization of those components via numerous cooperative interactions within the condensates, thus providing an attractive mechanism for combining factors in a timely fashion to generate transcription.

"special protein domains ... form biomolecular condensates that concentrate the transcription machinery."

The visualization of these condensates required the use of imaging technologies to characterize the behavior of individual protein domains, which was compared to the behavior of entire complexes. The domains were fused to fluorescent proteins to allow visualization with lattice light-sheet imaging (*8*). This allowed sufficient spatial and temporal resolution to see condensates, which display as puncta of fluorescence, formed by these domains in living cells and to characterize the dynamics of these condensates (see the figure). Domains of the TFs, FET [composed of FUS, EWS, and TAF15 (TATA-box binding protein–associated factor 15)] and SP1 (specificity protein 1) form puncta. Thus, these protein domains cluster with each other instead of freely diffusing separately from each other. Similarly, the enhancer binding factors BRD4 (bromodomain-containing protein 4) and MED1 (mediator of RNA Pol II transcription subunit 1) were seen in discrete puncta. MED1 is one of more than 20 proteins that comprise the Mediator complex (*9*), and the entire Mediator complex can also be seen as puncta. RNA Pol II, itself composed of 12 subunits, also forms clearly delineated puncta. There is evidence for multiple molecules important for transcription being incorporated into the same phase-separated condensates (*10*). Chong *et al.* and Cho *et al.* found that RNA Pol II colocalized with TFs or Mediator puncta in live cells, respectively.

[1]Department of Molecular Biology and MGH Research Institute, Massachusetts General Hospital (MGH), Boston, MA, USA. [2]Department of Genetics, Harvard Medical School, Boston, MA, USA. Email: kingston@molbio.mgh.harvard.edu

These puncta are formed via interactions between domains that are called either low-complexity domains (LCDs) or intrinsically disordered regions (IDRs). These domains have limited types of amino acids and are characterized as disordered according to their predicted secondary structure (*11*). To unify the terminology, we refer to these domains as low-complexity disordered regions (LCDRs). The current hypothesis is that proteins with these domains form networks, based, in part, on hydrophobic interactions, that are individually short lived and that allow for dynamic interplay that can create liquid-like properties (*12*). All three studies offer support for this hypothesis by showing that 1,6-hexanediol, which impairs hydrophobic interactions, can disrupt the structures. They also all used fluorescence recovery after photobleaching (FRAP) to show that molecules move in and out of these puncta rapidly, indicating that the components that make up the puncta are dynamic and not solid aggregates. Finally, Sabari *et al.* and Cho *et al.* show, with live-cell movies, that the puncta can merge together, just as water droplets will form a bigger droplet when they interact on a glass surface. Thus, the puncta have liquid-like characteristics.

The studies explore the role for specific protein sequences in self-association. In Chong *et al.*, examination of the LCDRs of TFs shows that there is specificity in the interactions. Factors can self-associate (for example, FET LCDRs), but certain interactions between separate factors (for example, FET LCDRs and the SP1 LCDR) do not occur. This might be due to differences in the sequences of the LCDRs. Mutational analysis of the EWS LCDR demonstrated that 29 tyrosine residues were required for LCDR interactions. Similarly, Sabari *et al.* found that the MED1 LCDR sequence was dominated by serine residues, which were required for self-association and liquid droplet formation in vitro.

The extent to which phase separation is a necessary element of transcriptional activation is called into question by Chong *et al.*, who found that TF LCDR self-assembly is transient at physiological concentration, in the range of seconds, and thus not consistent with phase separation into isolated droplets. This raises the question of when phase separation is an important part of the mechanism as opposed to a side effect of more transient interactions. Transient interactions between LCDRs might play a critical role in organizing components without a need for a stably phase-separated state. One possibility is that some interactions need only be transient, whereas others require greater stability. Perhaps the "potency" of domains to phase-separate (that is, the concentration needed to achieve phase separation) varies depending on need. For example, in Cho *et al.*, Mediator and RNA Pol II both have properties consistent with being phase-separated at normal physiological concentration. This is presumed to help the two complexes interact, but the interaction between the two complexes is transient. Activation of transcription by necessity requires that some interactions be transient: RNA Pol II must initiate transcription and elongate the transcript by interacting with different complexes (*13*). The ability of the activation components to move onto the next round of transcription, and thus switch contacts to a new RNA Pol II, is likely to be important for genes that are being rapidly transcribed. Thus, in transcriptional activation, there are sound theoretical reasons to have interactions that not only increase effective local concentration (as transient interactions between LCDRs would) but that also allow those components to be integrated with, instead of separated from, other nuclear components.

By contrast, in other regulatory settings, phase separation might be helpful or even necessary for appropriate regulatory functions of transcription, for example, interactions involved in stable repression (*14*, *15*) or, perhaps, long-range enhancer interactions. The continued development of technologies and experimental strategies to determine the importance of phase separation will be an exciting area to follow. Are these condensates central to most nuclear functions and thus a general regulatory mechanism with multiple distinct specificities and temporal characteristics? How much of nuclear function occurs in phase-separated domains? Do long noncoding RNAs, which are prevalent in the nucleus, contribute to the potency of phase separation and/or the organization of phase-separated domains? When is phase separation essential to the processes that generate the regulatory organization needed for life? ■

REFERENCES

1. M. S. Buckley, J. T. Lis, *Curr. Opin. Genet. Dev.* **25**, 126 (2014).
2. S. Chong *et al.*, *Science* **361**, eaar2555 (2018).
3. B. R. Sabari *et al.*, *Science* **361**, eaar3958 (2018).
4. W.-K. Cho *et al.*, *Science* **361**, 412 (2018).
5. S. F. Banani *et al.*, *Nat. Rev. Mol. Cell Biol.* **18**, 285 (2017).
6. F. Spitz, E. E. M. Furlong, *Nat. Rev. Genet.* **13**, 613 (2012).
7. M. Levine *et al.*, *Cell* **157**, 13 (2014).
8. B.-C. Chen *et al.*, *Science* **346**, 1257998 (2014).
9. B. L. Allen, D. J. Taatjes, *Nat. Rev. Mol. Cell Biol.* **16**, 155 (2015).
10. A. A. Hyman, K. Simons, *Science* **337**, 1047 (2012).
11. C. J. Oldfield, A. K. Dunker, *Annu. Rev. Biochem.* **83**, 553 (2014).
12. M. Kato *et al.*, *Cell* **149**, 753 (2012).
13. I. Jonkers, J. T. Lis, *Nat. Rev. Mol. Cell Biol.* **16**, 167 (2015).
14. A. G. Larson *et al.*, *Nature* **547**, 236 (2017).
15. A. R. Strom *et al.*, *Nature* **547**, 241 (2017).

10.1126/science.aau4795

CHEMISTRY

Hydrocarbon synthesis with vinyl cations

Catalytic generation of vinyl cations enables synthesis of functionalized hydrocarbons

By **Sean H. Kennedy** *and* **Douglas A. Klumpp**

Hydrocarbons—substances composed of carbon and hydrogen—are central to many industrial processes, including fuel preparation, polymer manufacturing, and the synthesis of feedstock chemicals (*1*). Many biosynthetic pathways also involve hydrocarbon reactions. For example, geranyl, farnesyl, or geranylgeranyl diphosphates are transformed into thousands of different terpenoid natural products (*2*). Hydrocarbons are key components of extraterrestrial atmospheres, nebulae, and other regions of space (*3*). Reactions of simple or polycyclic hydrocarbons were probably among the prebiological conversions that led to more complex molecular structures and the emergence of life (*4*, *5*). On page 381 of this issue, Popov *et al.* (*6*) report a method for catalyzing hydrocarbon reactions that have been difficult to achieve with other methods.

Over the past century, great strides have been made in understanding hydrocarbon chemistry and using hydrocarbons in synthetic reactions. Nevertheless, carrying out useful chemical transformations on the least-reactive hydrocarbon substrates—namely, the saturated alkanes and cycloalkanes—remains challenging. As a result, there has been much effort to develop C–H bond activation chemistry for hydrocarbons and hydrocarbon groups (*7*). These compounds are relatively inert because of their strong C–H bonds and the absence of reactive frontier molecular orbitals. To develop new methods of synthesis with saturated hydrocarbons, chemists must devise techniques for generating activated reagents that can attack the strong C–C or C–H bonds of these substrates. Popov *et al.* accomplish this in their study of vinyl cations and their chemistry.

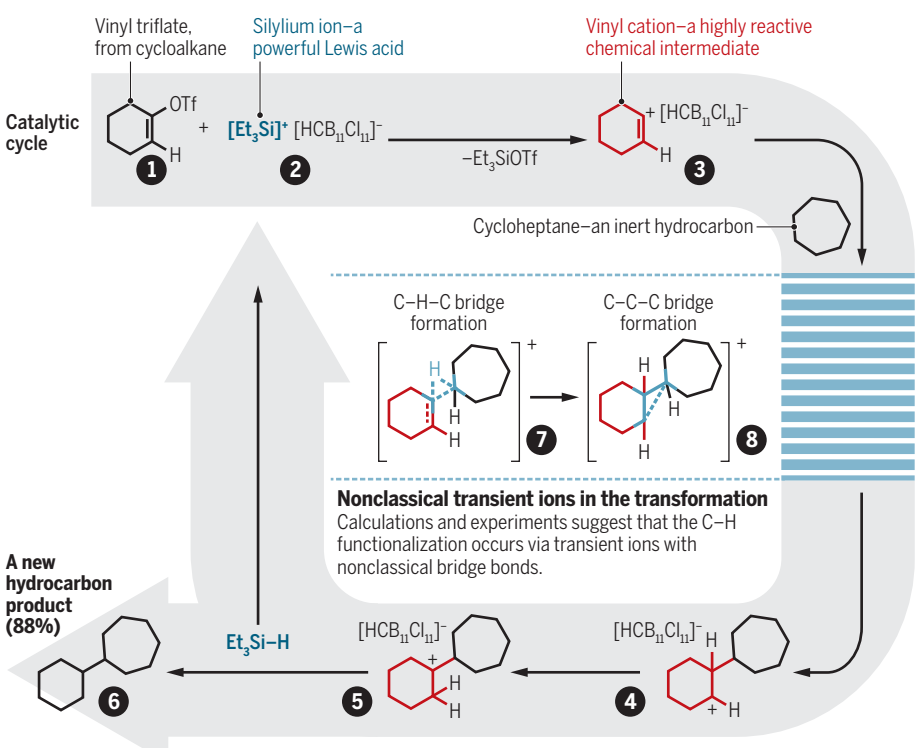
The authors show that use of a vinyl triflate, triethylsilane, and a trityl salt initiator leads to a catalytic reaction that converts alkanes and cycloalkanes into functionalized products. For example, the vinyl triflate **1** derived from cyclohexanone reacts with cycloheptane to provide the hydrocarbon **6** in 88% yield (see the figure). This remarkable transformation is accomplished under very mild conditions for a C–H functionalization; it is complete in 2 hours at just 30°C. A key part of the methodology is the use of carborane, a weakly coordinating anion. As an inert counter ion, the carborane helps to stabilize highly electrophilic cationic intermediates—the silylium ion **2** and vinyl cation ***3***—and enables the cations to exist as persistent species in the solution. This allows the authors to use the reactive vinyl cation in difficult hydrocarbon synthetic reactions. To demonstrate the synthetic utility of this method, the authors use it to functionalize *n*-alkanes, cycloalkanes, and arenes and to prepare a steroid derivative.

Like a skilled trapeze artist swinging high above a circus floor, the chemistry moves beautifully through high-energy intermediates (see the figure). The reaction sequence begins with an initiation step, in which the trityl cation abstracts hydride from triethylsilane. This step produces the first high-energy intermediate, the silylium ion **2**. Since the first silylium ions were generated in the 1990s, it has become clear that these species are very strong Lewis acids (*8*). As such, the silylium ion coordinates to the triflate leaving group, resulting in cleavage of the C–O bond. This step provides the second high-energy intermediate, the vinyl cation ***3***. This intermediate inserts into the C–H bond of cycloheptane to provide trivalent carbocation intermediates **4** and **5**. The final step involves hydride transfer from triethylsilane to the carbocation **5**, leading to the hydrocarbon product **6** and regenerating the high-energy silylium ion **2**.

When the authors investigated the key C–C bond–forming reaction step in a molecular dynamics simulation, they found that the C–H functionalization occurs through a nonclassical ionic structure (see the figure). Simulation of the reaction of vinyl cation ***3*** and cyclohexane suggests that the system proceeds from an ambimodal transition-state structure to the transient species **7**, with a bridging C–H–C structure. This is followed by C–C

Transforming saturated hydrocarbons

Use of a silylium catalyst leads to the formation of a reactive vinyl intermediate that facilitates the formation of new hydrocarbon products. Et, ethyl; OTf, trifluoromethanesulfonate.



GRAPHIC: A. KITTERMAN/*SCIENCE*

Department of Chemistry, Northern Illinois University, DeKalb, IL 60115, USA. Email: dklumpp@niu.edu

bond formation to give the transient species **8**, a structure with an unsymmetrical C–C–C bridge and nonclassical ion character. Elegantly designed mechanistic experiments with the vinyl triflates provide support for the nonclassical structures seen in the molecular dynamics simulations.

Nonclassical carbocation structures are often characterized by three-atom two-electron bonding patterns. Because these types of structures deviate considerably from the common valance bond model (with two-atom-two-electron bonding), their existence and role in organic chemistry has been a matter of considerable debate during the past 70 years. Strong evidence for nonclassical carbocations and related hypercoordinate carbon species has been reported—including spectroscopic, chemical, and x-ray crystallography evidence—and these structures are now universally recognized as a viable part of hydrocarbon chemistry (*9*, *10*).

The existence of nonclassical carbocations was once questioned by many in the chemical community, but it is now clear that these types of structures have broad importance in the chemistry of hydrocarbons (*9*). Nonclassical carbocation chemistry extends all the way from enzyme active sites in biosynthetic chemistry (*11*) to the catalytic cracking reactors in the petroleum industry (*1*). Popov *et al.*'s study demonstrates that a new generation of synthetic hydrocarbon reactions is also possible through the involvement of vinyl cations and transient nonclassical carbocation structures. ■

"Like a skilled trapeze artist swinging high above a circus floor, the chemistry moves beautifully through high-energy intermediates..."

REFERENCES

1. G. A. Olah, A. Molnar, *Hydrocarbon Chemistry* (Wiley, ed. 2, 2003).
2. A. Vattekkatte, S. Garms, W. Brandt, W. Boland, *Org. Biomol. Chem.* **16**, 348 (2018).
3. S. Kwok, *Nature Astronomy* **1**, 642 (2017).
4. R. I. Kaiser, D. S. N. Parker, A. M. Mebel, *Ann. Rev. Phys. Chem.* **66**, 43 (2015).
5. C. Sagan, W. R. Thompson, *Icarus* **59**, 133 (1984).
6. S. Popov *et al.*, *Science* **361**, 381 (2018).
7. J. F. Hartwig, *J. Am. Chem. Soc.* **138**, 2 (2016).
8. P. P. Gaspar, *Science* **297**, 785 (2002).
9. G. A. Olah, G. K. Surya Prakash, K. Wade, A. Molnar, R. E. Williams, *Hypercarbon Chemistry* (Wiley, ed. 2, 2011).
10. F. Scholz *et al.*, *Science* **341**, 62 (2013).
11. Y. J. Hong, J.-L. Giner, D. J. Tantillo, *J. Am. Chem. Soc.* **137**, 2085 (2015).

10.1126/science.aau0809

TUMORIGENESIS

Inflamed T cells and stroma drive gut tumors

Loss of a tumor suppressor in T cells and stromal cells drives gastrointestinal polyp growth

By **Pablo E. Hollstein** *and* **Reuben J. Shaw**

Inactivating somatic mutations causing loss of function in the tumor suppressor gene *STK11* (serine-threonine kinase 11), which encodes the protein LKB1 (liver kinase B1), frequently occur in several sporadic cancers, notably lung, pancreatic, and female reproductive tumors. Additionally, inherited heterozygous germline mutations in *STK11* cause Peutz-Jeghers syndrome (PJS), a cancer predisposition syndrome (*1*). A hallmark of PJS is the growth of numerous benign gastrointestinal (GI) hamartomatous polyps and an elevated risk of developing malignancies in several organs (*2*). Studies in mouse models of PJS have pointed to LKB1 deficiency in the stroma, not epithelium, as the contributing factor for the formation of the GI polyps (*3*). Stroma consists of the fibroblasts, smooth muscle, extracellular matrix, and basement membrane that support epithelial tissue. However, how LKB1 deficiency in stromal cells could trigger polyp formation was unclear. Two recent studies now shed light on the mechanism of PJS polyp formation. On page 406 of this issue, Poffenberger *et al.* (*4*) find that loss of a single allele of *Stk11* in T cells is sufficient to drive the formation of GI polyps. Similarly, a study by Ollila *et al.* (*5*) demonstrated that heterozygous loss of *Stk11* in gastric stromal cells was sufficient to drive GI polyp formation. Transcriptional profiling in both studies revealed an up-regulation of inflammatory cytokines involved in promoting the expansion and overgrowth of both stroma and normal GI epithelium. By finding that inflammation in the stromal compartment due to LKB1 loss in T cells (*4*) or stromal cells (*5*) drives polyp formation, these studies highlight how inflammatory signals can profoundly alter the microenvironment and fuel tumor formation.

GI polyps in PJS patients contain a mixture of cell types, all of which have inherited a heterozygous mutation in *STK11*, and are characterized by a smooth muscle core, an expanded stromal compartment, and hyperplastic epithelia (*2*). Poffenberger *et al.* used mouse models in which one allele of *Stk11* could be conditionally deleted in different cellular compartments to understand which cells contribute to polyp formation. They observed that heterozygous deletion of *Stk11* in GI epithelial cells did not promote polyp formation. Unexpectedly, they instead found that heterozygous loss of *Stk11* in the hematopoietic compartment was sufficient to drive growth of GI polyps that were histologically indistinguishable from those in *Stk11* heterozygous mice, and which also resembled polyps from PJS patients. Furthermore, they found that whereas heterozygous loss of *Stk11* in B cells did not contribute to polyp formation, heterozygous loss of *Stk11* in T cells was sufficient to promote intestinal polyposis. Complementing these findings, Ollila *et al.* showed that heterozygous loss of *Stk11* in gastric stromal cells was sufficient to induce PJS-like polyps.

How does loss of LKB1 lead to polyp formation? LKB1 activity exerts pleiotropic effects on cell polarity, growth, and metabolism, including responses to cellular energy stress that are mediated by phosphorylating and hence controlling the activation of adenosine monophosphate (AMP)–dependent protein kinase (AMPK) (*1*). AMPK negatively regulates the activity of mammalian target of rapamycin complex 1 (mTORC1), a central regulator of cell growth and proliferation (*1*). Consequently, loss of either LKB1 expression or AMPK activation results in up-regulation of mTORC1 signaling, which is thought to contribute to tumor growth in many contexts, including PJS polyposis, and in sporadic tumors with LKB1 loss (*1*). Previous studies found increased mTORC1 signaling in the epithelial compartment of PJS polyps (*6*), and remarkably, Poffenberger *et al.* also observed this event when LKB1 was only deleted in T cells. Surprisingly, mice with T cell–specific ablation of AMPK did not develop polyps. Moreover, simultaneous deletion of *Stk11* and *Mtor* in T cells still resulted in the development of GI polyps. Together, these results suggest that AMPK

Laboratory of Molecular and Cell Biology, The Salk Institute for Biological Studies, La Jolla, CA, USA. Email: shaw@salk.edu

and mTORC1 in T cells are dispensable for PJS polyp development in these models, and that alternative LKB1-dependent signaling pathways must be responsible for polyp growth. Similarly, Ollila *et al.* report that inactivation of AMPK in the GI stromal compartment did not induce polyposis.

What signals from LKB1-deficient T cells or gastric stromal cells are sufficient to drive polyp growth? Poffenberger *et al.* found that heterozygous loss of *Stk11* in T cells induced the activation of $CD4^+$ and $CD8^+$ T cells in mesenteric lymph nodes, with pronounced infiltration of those T cells, along with macrophages and neutrophils, into the polyp microenvironment. Heterozygous expression of *Stk11* in mice, or specifically in T cells (*4*) or stromal cells (*5*), resulted in the secretion of large amounts of inflammatory cytokines, including interleukin-6 (IL-6) and IL-11, which have previously been associated with gastric tumor development (*7*). This increased cytokine production was accompanied by hyperactivated JAK-STAT (Janus kinase–signal transducer and activator of transcription) signaling in the stromal compartment, which contributes to inflammation and cancer (*8*). Furthermore, treatment with AZD1480, an inhibitor of JAKs, quenched STAT3 activation in the stroma and significantly reduced polyp growth in *Stk11*-heterozygous mice. Similarly, Ollila *et al.* found that activated JAK-STAT3 signaling in stromal cells drives polyp formation, and the JAK inhibitor ruxolitinib (which is already clinically approved for myeloproliferative diseases) has therapeutic potential in mouse models of PJS. Taken together, a model emerges in which LKB1 deficiency in T cells or gastric stromal cells produces a proinflammatory microenvironment and subsequently gives rise to GI polyps (see the figure).

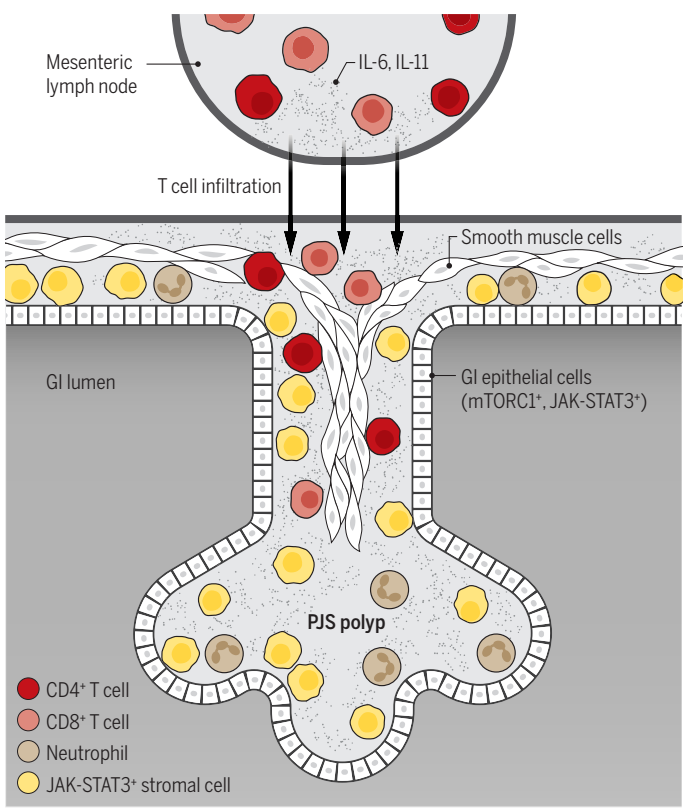
If not through reduced AMPK function, how does LKB1 deficiency lead to the production and promotion of inflammatory cytokines? In addition to activating AMPK, LKB1 also phosphorylates and activates a family of 12 additional AMPK-related kinases (*9*), the potential roles of which in T cell deregulation were not explored by Poffenberger *et al.* Notably, Ollila *et al.* reported that reduced expression of the AMPK family members SIK1 (salt-inducible kinase 1), MARK1 (MAP/microtubule affinity-regulating kinase 1), or MARK4 could induce production of IL-11 in fibroblasts, raising the intriguing possibility that one or more of these AMPK family members may mediate production of the proinflammatory cytokine storm observed in human and mouse PJS polyps.

The findings of Poffenberger *et al.* and Ollila *et al.* constitute an important example of how inflammation can be driven by haploinsufficiency of tumor suppressor genes involved in inherited familial cancer syndromes. Similar to LKB1 deficiency in PJS, heterozygous loss of other tumor suppressors can drive benign tumor growth containing mixed cell types, including epithelial, stromal, and immune cell components (*10*). Examples of heterozygous loss in mixed-lineage tumors include loss of APC (adenomatous polyposis coli) in familial adenomatous polyposis, SMAD4 in familial juvenile polyposis (FJP), PTEN (phosphatase and tensin homolog) in Cowden's disease, TSC1 (tuberous sclerosis complex 1) and TSC2 in tuberous sclerosis, VHL (von Hippel–Lindau tumor suppressor) in von Hippel–Lindau syndrome, and NF1 (neurofibromatosis 1) in neurofibromatosis type 1. The involvement of immune cell components in driving benign tumor formation as demonstrated by Poffenberger *et al.* echoes observations in mouse models of disorders such as FJP, in which loss of SMAD4 in T cells up-regulates inflammatory cytokines and results in gastrointestinal polyp growth (*11*, *12*), and in neurofibromatosis type 1, in which *Nf1*-heterozygous mast cells (a type of inflammatory immune cell) secrete proinflammatory cytokines into the nerve microenvironment that are required for and drive neurofibroma growth (*13*). It will be very interesting to see whether the involvement of a proinflammatory, deregulated immune system also contributes to benign tumor growth in other familial cancer syndromes besides PJS, FJP, and neurofibromatosis type 1 (*4*, *5*, *11*–*13*) or in sporadic solid tumors with LKB1 loss (*14*, *15*).

The findings of Poffenberger *et al.* and Ollila *et al.* also suggest that JAK inhibitors may be potential new therapeutic modalities for GI and other tumors arising in PJS patients for which there are few treatment options. Future exploration of the involvement and therapeutic potential of targeting JAK-STAT3 signaling in sporadic tumors with LKB1 inactivation is now also warranted. ■

Deregulation of the tumor microenvironment in polyposis

Heterozygous loss of LKB1 expression in T cells or stromal cells is sufficient to induce proinflammatory cytokines, including IL-6 and IL-11, which recruit neutrophils and other inflammatory immune cells. This inflammatory microenvironment drives JAK-STAT3 signaling in stromal and epithelial cells, concurrent with epithelial mTORC1 activation, and is sufficient to induce polyp growth.



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

REFERENCES

1. D. B. Shackelford, R. J. Shaw, *Nat. Rev. Cancer* **9**, 563 (2009).
2. A. Hemminki, *Cell Mol. Life Sci.* **55**, 735 (1999).
3. P. Katajisto *et al.*, *Nat. Genet.* **40**, 455 (2008).
4. M. C. Poffenberger *et al.*, *Science* **361**, 406 (2018).
5. S. Ollila *et al.*, *J. Clin. Invest.* **128**, 402 (2018).
6. R. J. Shaw *et al.*, *Cancer Cell* **6**, 91 (2004).
7. T. L. Putoczki *et al.*, *Cancer Cell* **24**, 257 (2013).
8. H. Yu *et al.*, *Nat. Rev. Cancer* **9**, 798 (2009).
9. J. M. Lizcano *et al.*, *EMBO J.* **23**, 833 (2004).
10. K. Inoki , M. N. Corradetti, K. L. Guan, *Nat. Genet.* **37**, 19 (2005).
11. B.-G. Kim *et al.*, *Nature* **441**, 1015 (2006).
12. J. N. Hahn *et al.*, *J. Clin. Invest.* **121**, 4030 (2011).
13. F.-C. Yang *et al.*, *Cell* **135**, 437 (2008).
14. S. Koyama *et al.*, *Cancer Res.* **76**, 999 (2016).
15. F. Skoulidis *et al.*, *Cancer Disc.* **8**, 822 (2018).

10.1126/science.aau4804

RETROSPECTIVE

Paul D. Boyer (1918–2018)

Pioneer of molecular machines and inspirational leader

By **David S. Eisenberg**

On 2 June, 2 months before his 100th birthday, Paul Boyer died peacefully, surrounded by family, including Lyda, his wife of nearly 79 years. Paul was deeply admired by his students, colleagues, and friends for his many talents. He was skilled in mechanics, science, diplomacy, administration, and athletics. Beyond his scientific success, culminating in a share of the 1997 Nobel Prize in Chemistry, Paul contributed to the community by serving as a role model, demonstrating the wisest path forward in tough situations.

Paul died at his home in Los Angeles, which he built, serving as his own general contractor. Recalling Paul's mechanical skills, his daughter Gail said that they had never had a repairman in the house. Earlier, while at the University of Minnesota, Paul and Lyda had designed their first home, for which Paul was the plumber, electrician, and cabinet maker. Paul was teaching his class at the university when the inspector arrived at their homesite to evaluate the plumbing. He lingered intently over each fitting and junction and then confronted Lyda with the question, "Who did this work?" Her heart fell as she told him it had been her husband. "He must have been trained in the old country," said the inspector, "they don't do fine work like this anymore."

In fact, Paul was raised in Provo, Utah, the fourth of seven children, a descendant of what he called "hardy Mormon pioneer stock." Although he rejected the religious tenets of his church, he attributed his scientific career to parental and community devotion to education. He attended Brigham Young University, a few blocks from his home, graduating with a degree in chemistry in 1939. That autumn, he left for graduate school at the University of Wisconsin–Madison armed with $150 in cash, a $400-per-year scholarship, and his new bride, Lyda. He soon developed a second love—for biochemistry. After earning his Ph.D. in the subject in 1943, he did war research on blood proteins at Stanford University, joined the faculty at the University of Minnesota, where he introduced chemical and isotopic methods for the study of enzyme mechanisms, and then settled for good at the University of California, Los Angeles (UCLA), in 1963. Starting at age 70, Paul and Lyda enjoyed an active retirement of tennis, golf, bridge, and travel throughout the American West, interrupted only by the award of Paul's Nobel Prize at age 79.

Paul's most profound discovery was that adenosine triphosphate (ATP), the energy currency of life, is synthesized in living cells by a rotary molecular engine, ATP synthase. The inference of this unprecedented mechanism from measurements of chemical rates could have been made only by a scientist with Paul's deep mechanical understanding. He applied mass spectrometry to monitor oxygen-18 exchange kinetics to understand the enzymatic mechanism of ATP synthase. His experiments suggested that the energy of oxidation is coupled to the release of ATP from the enzyme, rather than to ATP synthesis (his so-called "binding change mechanism"), and that the enzymatic reaction of ATP synthesis must involve more than one catalytic site. From these insights, he correctly proposed a rotary molecular engine, later gloriously illuminated by the crystal structure of ATP synthase by British chemist John E. Walker. For this work, he and Walker shared half of the Nobel Prize in Chemistry in 1997.

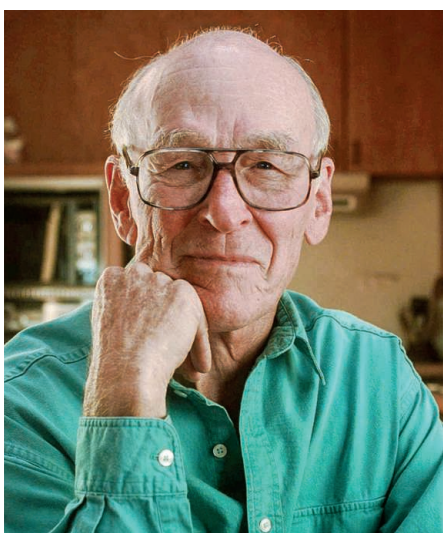


PHOTO: LACY ATKINS/AP PHOTO

Paul had the curiosity and confidence to grapple with big questions and big projects. Though he studied some 20 enzymes during the course of his career, he kept returning to ATP synthase. As he said in a biographical memoir, "Although the larger questions [about oxidative phosphorylation] were not likely to be answered, at least I wanted to try." He knew that "effort is difficult without optimism, and accomplishment is rare without effort."

Soon after moving from Minnesota to UCLA, Paul applied his optimism and effort to founding the Molecular Biology Institute (MBI). He immediately set out to recruit faculty (around 20 over time, including me) and students (some 400 have completed molecular biology Ph.D.s as of 2018). With phenomenal persistence, he put together a mosaic of funding sources for construction of the laboratory now known as Boyer Hall. However, not every project of Paul's came to fruition. In 1980, he envisioned a set of bicycle freeways emanating from UCLA that would change the character of Los Angeles. Clearly, that idea was before its time, but Paul could once again say "at least I wanted to try."

Active in the national leadership of the biochemical community, Paul served as president of the society now called the American Society for Biochemistry and Molecular Biology, as well as chair of the Division of Biological Chemistry of the American Chemical Society. With Lyda's assistance, he produced 18 volumes of the review series *The Enzymes*.

Paul's many achievements were made possible by his optimism, effort, ambition, and know-how, but there was something more: character. In my every encounter with Paul, from the first in 1967 to my last shortly before his death 50 years later, I sensed his drive to get at the truth, his generous judgments of others and their work, and his modesty. Together, these traits made him the most effective leader that many of us have ever personally known. Under his 18-year directorship of the MBI, faculty meetings were amicable and short. Paul had discussed the issues beforehand with each of us and had forged a consensus prior to the meeting. A generous author, Paul carefully noted the advances made by each of his graduate students and postdoctoral fellows in his several scientific autobiographies. In appreciation of these contributions, Paul donated to UCLA part of his Nobel Prize stipend for awards to his current UCLA postdoctoral fellows. Paul also invariably took care to credit the results from the labs of other scientists. Even when he disagreed, the tone was one of respect, sometimes noting where in hindsight his own interpretation had been wrong.

In the present era, when science and reason are under attack, facts are considered optional, lying by leaders is not unexpected, and bullying and falsely hogging credit are tolerated, Paul Boyer's life stands as a model of the proper path for all of us who have had the privilege of knowing him. ■

Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA. Email: david@mbi.ucla.edu

10.1126/science.aau6601

POLICY FORUM

BIOMEDICAL RESEARCH

The anticommons at 20: Concerns for research continue

Emerging trends in exclusionary rights may affect research

By **Jorge L. Contreras**

Fifty years ago, Garrett Hardin famously predicted that unrestrained use of a common resource could lead to its overuse and depletion—the so-called "tragedy of the commons." Thirty years later, Heller and Eisenberg introduced the term "anticommons" to the research-policy lexicon (*1*), postulating that unfettered exercise of individual property rights could lead to an equally tragic underutilization of resources. By drawing on analogies from land-use planning, they reasoned that if multiple holders of intellectual property (IP) rights, particularly patents, covering a biomedical technology can individually block others from conducting research on that technology, then overall research progress could be stifled. Though many observers now agree 20 years later that empirical evidence of a patent-based anticommons in biomedical research is inconclusive (*2*), if not wholly refuted (*3*, *4*), there are emerging areas beyond patent law in which the proliferation of exclusionary rights may threaten research in much the way that Heller and Eisenberg predicted.

Although a common resource such as a pasture or a river is held by multiple parties for the benefit and use of all, an anticommons is held not in common, but in separate, yet co-dependent, fragments, whereby any individual owner may block others' use of the whole. Heller and Eisenberg were particularly concerned with privatization of "upstream" research tools—basic scientific discoveries and techniques that can be used to develop a range of "downstream" diagnostics and therapeutics—arguing that "each upstream patent allows its owner to set up another tollbooth on the road to product development, adding to the cost and slowing the pace of downstream biomedical innovation" (*1*). For example, a proliferation of patents, sometimes called a patent thicket, covering specific protein receptors or basic DNA sequencing techniques could impede research on a broad range of applications. Only by limiting the potential hold that multiple patent holders may have over basic research tools, they argued, can the inefficiencies that are likely to emerge from anticommons be averted (*1*). They thus called on governmental actors such as the U.S. National Institutes of Health (NIH) to "ensure coherent boundaries of upstream patents and to minimize restrictive licensing practices that interfere with downstream product development."

In their conceptualization of the anticommons, Heller and Eisenberg challenged earlier theoretical work [for example, (*5*)] that argued that assets, including inventions and other intangibles, will be put to their highest and best use only if they are effected with private interests that incentivize their owners to manage and exploit them efficiently—a property-based solution to Hardin's tragedy of the commons. Heller and Eisenberg's work signaled a new wave of interest in the application of commons-based solutions to intellectual assets, building on the foundation laid by Elinor Ostrom and others in the area of tangible common pool resources. In addition to privatization strategies, tragedies of the commons can also be solved by collective management of common assets. In the regime of the anticommons, it is excessive privatization of resources that hinders, rather than helps, the productive use of assets.

> ***"In the regime of the anticommons, it is excessive privatization of resources that hinders, rather than helps, the productive use of assets."***

THE SEARCH FOR ANTICOMMONS

Soon after the Heller and Eisenberg article, researchers began to seek empirical evidence of a developing anticommons in biomedical research. The proliferation of patents on individual genes and key sequencing processes was of particular concern. One study reported that U.S. patents covered approximately 20% of known human genes (*6*), and another found that patents had a negative impact on the use and development of molecular diagnostic tests (*7*).

Others worried that patenting human DNA and other research tools could result in serious impediments to the development of multigene diagnostic panels and DNA chips that could simultaneously test for large numbers of genetic variants, as well as high-throughput genotyping and sequencing platforms and animal models (*8*). Another study showed that genetic variants contractually protected by the private firm Celera Genomics, when compared to public data on comparable variants from the Human Genome Project (HGP), "generated economically and statistically significant reductions in subsequent scientific research and product development" (*9*).

Findings like these set off alarm bells throughout the research community and seemed to validate the anticommons hypothesis. However, outside of genetics, studies found that the behavior of biomedical researchers was not substantially affected by the existence of patents. In many cases, academic researchers simply appeared to ignore patents in their research (*2*).

POOLING, AGGREGATION, COMMONS

If we have not seen the emergence of a pronounced patent anticommons in biomedical research, it is worth asking why. One oft-raised possibility is that rational private actors, when faced with the fragmentation and congestion threatened by an anticommons, will develop collective and mutually beneficial solutions to enable research to advance nevertheless. Such "pooling" approaches have arisen to address rights fragmentation in industries ranging from music to aircraft to radio and telecommunications (*3*).

Heller and Eisenberg recognized that the gridlock threatened by anticommons could be remedied through bargaining and the formation of IP pools. Yet they expressed reservations about the general viability of IP-pooling solutions. They identified several distinct features of the biotechnology and pharmaceutical industries that could make successful bargaining over patent rights less successful than in fields such as aviation and music, including transaction costs associated with accumulating sufficient rights to practice biotechnology inventions, the heterogeneous interests of patent holders, and cognitive biases that cause biotechnology patent holders to

S. J. Quinney College of Law and Department of Human Genetics, University of Utah, Salt Lake City, UT, USA.
Email: jorge.contreras@law.utah.edu

overestimate the value of their own technical contributions (*1*).

But, despite years of advocacy surrounding potential patent pools for HIV/AIDS and severe acute respiratory syndrome (SARS) research and the emergence of substantial pooling activity in other industries, relatively few patent pools—and none of commercial importance—have formed in the pharmaceutical or biotechnology sectors. Several factors could explain the absence of pooling in this arena: the need for at least some market exclusivity in an environment with extremely high costs of product development, clinical trials, and regulatory approval; patent holders' desire to retain control over their assets; and concerns over compromising commercial secrecy by collaborating with others (*4*).

Although the formal pooling of fragmented property interests in biomedical research may not have materialized in a meaningful way, this sector is also characterized by a remarkable, and arguably unique, outpouring of valuable discoveries to the public domain. The sum of these contributions may be moving us toward a comprehensive medical-information commons (*10*)—the very antithesis of the anticommons predicted by Heller and Eisenberg. The beginning of this trend is often traced to the 1996 Bermuda accord reached by leaders of the HGP, which required all HGP research groups to deposit their sequence data into public databases 24 hours after being generated. The ethos of rapid, public release of genomic and related data has now become the norm in many fields of biomedical research and has even expanded to the commercial sector (*11*). Voluntary data sharing by research institutions and corporations has also emerged in areas such as testing for mutations in the breast cancer gene *BRCA*, largely in response to the proprietary data approaches taken by firms such as Myriad Genetics (*11*, *12*).

Thus, just as collective action by affected stakeholders has been shown by Ostrom and others to avert tragedies of the commons involving scarce shared resources, the collective action of the biomedical research community—governments, institutions, and individual researchers—developed a powerful response to the potential fragmentation and propertization of the research environment (*11*). It is likely that the success of these public research commons has contributed to the dearth of observed anticommons effects.

There may be other reasons that patent anticommons did not take hold in the biomedical sciences to the degree envisioned by Heller and Eisenberg. For example, in line with their recommendation that upstream research tools be licensed broadly and nonexclusively, the NIH adopted a policy in 1999 urging its grant recipients to license patented research tools on a nonexclusive basis to promote their greatest utilization, commercialization, and public availability. In 2007, a group of 11 major U.S. research universities followed suit and committed, in a set of core principles known as the "Nine Points," that research tools should be made as broadly available as possible. Today, more than 100 research institutions around the world have voluntarily subscribed to the Nine Points.

"...it remains the case that the combination of extensive propertization with fragmentation of ownership can lead to transactional gridlock and underutilization of socially valuable assets."

Finally, over the past decade, patent laws in a number of countries have become weaker, not stronger, with respect to the protection of upstream biomedical innovations. Beginning in 2010, a series of U.S. Supreme Court decisions clarified that "products of nature," "mental processes," and "abstract ideas" are not eligible for patent protection. As a result, it has become increasingly difficult to patent basic biomedical discoveries in the United States, so much so that some have begun to ask whether the lack of patents in certain areas (for example, molecular diagnostics and personalized medicine) will itself impede future innovation and discovery (*13*).

ANTICOMMONS ON THE HORIZON?

Despite the absence of a serious patent anticommons in biomedical research today, Heller and Eisenberg's warning should not be ignored. There are several areas beyond patent law in which the proliferation of exclusionary rights could impede biomedical research and product development in ways that are similar to those that Heller and Eisenberg envisioned.

First, in the vacuum left by limitations on patenting human DNA, some firms have increasingly turned to trade-secret law to protect data that they collect from patients and test subjects (*12*). In most countries, trade-secret law gives an enforceable property-like right to the holder of commercially valuable information that is deemed to be confidential. And unlike patents, which expire after 20 years, trade-secret protection continues in perpetuity, so long as the relevant information remains secret. Trade secrecy also challenges research in ways different from patents. For example, a patent is an official document that publicly discloses the patented invention, thus enabling others to study and improve on its features and techniques. Trade secrets, by their nature, need never be disclosed to the world, thus limiting the opportunity for follow-on research and improvements.

If more data are treated as secret by researchers, there may be less overall growth in knowledge and the medical information commons may not grow as anticipated (*10*). In addition, the withholding of data by individual researchers may give rise to anticommons effects. In fields that are

ILLUSTRATION: ADAPTED FROM SHAW NIELSEN/THEISPOT

characterized by large bodies of interdependent observations, the withholding of individual research results may stymie the development of a full understanding of the field. This fragmentation effect is particularly salient in the area of molecular diagnostic testing, in which different testing labs may each collect genetic variant data from patients, but the true potential of these data will only be realized if they are combined and analyzed together. The inability of researchers to conduct cross-cutting analyses could result in less accurate diagnostics and fewer therapeutics. Thus, although holding data privately does not have the same exclusionary effects as obtaining a patent (that is, others are not precluded from independently generating the same data), the difficulties that emerge in consolidating data from different sources, coupled with the reduction in overall welfare arising from a lack of the full spectrum of results, fall close to Heller and Eisenberg's anticommons.

Eisenberg anticipated this issue in 2008, observing that with "practically excludable" resources such as data and biological samples (that is, where exclusion is based not on legal restrictions but instead on the need for one researcher to provide the resource to others), it is up to users to persuade owners to permit access (*2*). As such, a data-driven anticommons could emerge if enough researchers keep their data proprietary and share them only on condition of payment or not at all. There are few effective regulatory solutions to this issue under current law, but the continued encouragement (or requirement) of data sharing by research funders, health care payors, and leading journals could avert the worst effects of such an anticommons.

Second, despite the recent tightening of legal requirements for patent eligibility, patenting activity continues around the world with respect to emerging biomedical technologies such as CRISPR (clustered regularly interspaced short palindromic repeats) gene editing. The potential exists for substantial fragmentation of the CRISPR patent landscape—not through disaggregated ownership of CRISPR patents themselves (though that is also possible) but through parcelization of the CRISPR-patent estate via narrowly drawn licensing agreements controlled by a handful of private "surrogate" companies empowered by the academic research institutions that made foundational CRISPR discoveries (*14*). These companies may have incentives to license technology to others on a limited gene-by-gene, indication-by-indication basis that does not encourage the full breadth of potential research and product development. And although some CRISPR patent holders have granted favorable rights to academic researchers, this solution is neither universal nor binding nor permanent. As a result, policy-makers should continue to remind academic research institutions of their public commitments to work toward wide dissemination of the fruits of their research and to discourage the exclusive licensing of broadly applicable research tools. If such reminders prove to be insufficient, research-funding agencies could more definitively require the broad availability of funded research through licensing.

Finally, a new and potentially important form of fragmentation has recently emerged with respect to individual health information. A spate of recent legal disputes in the United States has led to increasing calls for personal ownership of genetic and other health information (*15*). The rationales for this privatization movement are varied, ranging from concerns over individual autonomy, privacy, and dignity, to offering a more palatable alternative than corporate ownership, to creating a basis for data-based market transactions, and to dissatisfaction with existing regulatory and administrative data protection frameworks in the United States and the European Union. Yet, despite the good intentions behind many of these proposals, granting individuals an enforceable property interest in information about themselves, including the right to receive compensation for its use, could pose considerable impediments to data-driven research, particularly in the coming era of megacohort studies involving a million or more individuals.

Thus, although Heller and Eisenberg worried that fragmented interests held by a few dozen or hundred patent owners could impede biomedical research, the possibility that millions of individual data subjects could demand clearance, oversight, or payment to use their data in successive studies, or withdraw their data from existing databases and ongoing analyses, has far more dramatic ramifications for research (*15*). To avert this version of the anticommons, policy-makers and courts should continue to resist calls for individual data ownership and look instead to meaningful regulatory and legal measures to protect individuals against invasions of privacy and abuse by researchers.

BROADER LESSONS

Though the patent-driven biomedical anticommons envisioned by Heller and Eisenberg does not appear to have emerged widely, researchers and policy-makers must remain vigilant as new sources of potential anticommons emerge. Admittedly, the precise parameters of exclusivity generated by new forms of property fragmentation vary, and they do not all exhibit the same features, or lend themselves to the same solutions, as patent-based anticommons. Nevertheless, it remains the case that the combination of extensive propertization with fragmentation of ownership can lead to transactional gridlock and underutilization of socially valuable assets.

Thus, it is worth looking beyond biotechnology patents to the more general distinction between fragmented and common ownership models. Most importantly, no matter how unobjectionable the legal and commercial rationales underlying propertization strategies may first appear, attention should be given to the downstream effects that potential anticommons could cause. If serious impediments to socially valuable activity, including biomedical research, are likely to result, then policies limiting the impact of such anticommons should be considered and weighed in the balance. ■

REFERENCES

1. M. A. Heller, R. S. Eisenberg, *Science* **280**, 698 (1998).
2. R. S. Eisenberg, *Hous. L. Rev.* **45**, 1059 (2008).
3. J. M. Barnett, *Harv. J. Law Technol.* **29**, 127 (2015).
4. G. Van Overwalle, in *Research Handbook on Intellectual Property and the Life Sciences*, D. Matthews, H. Zech, Eds. (Edward Elgar Publishing, 2017), pp. 304–334.
5. E. W. Kitch, *J. Law Econ.* **20**, 265 (1977).
6. K. Jensen, F. Murray, *Science* **310**, 239 (2005).
7. M. K. Cho, *J. Mol. Diagn.* **5**, 3 (2003).
8. B. Verbeure, in *Gene Patents and Collaborative Licensing Models: Patent Pools, Clearinghouses, Open Source Models and Liability Regimes* (Cambridge Univ. Press, 2009), pp. 3–32.
9. H. L. Williams, *J. Polit. Econ.* **121**, 1 (2013).
10. R. Cook-Deegan, A. L. McGuire, *Genome Res.* **27**, 897 (2017).
11. J. L. Contreras, B. M. Knoppers, *Ann. Rev. Genomics Hum. Genet.* 10.1146/annurev-genom-083117-021552 (2018).
12. C. J. Guerrini, A. L. McGuire, M. A. Majumder, *Science* **356**, 586 (2017).
13. K. Madigan, A. Mossoff, *George Mason Law Rev.* **24**, 939 (2017).
14. J. L. Contreras, J. S. Sherkow, *Science* **355**, 698 (2017).
15. J. L. Contreras, *Georgetown Law J.* **105**, 1 (2016).

10.1126/science.aat4684

Some researchers believe that beavers were critical to promoting ecosystem recovery in Yellowstone.

ECOLOGY

Nature's riverkeepers

Recognizing their role in maintaining healthy watersheds, "beaver believers" work to rehab the rodent's reputation

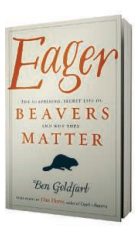
By **Sarah Boon**

Why should we care about beavers? Consider all they do. Beavers convert vegetation to marsh to wetland and back again. They facilitate water storage in ponds and recharge groundwater. Ponds and meadows sculpted by beavers concentrate nutrients such as nitrogen and phosphorus. Not only does this create fertile ground, it helps filter agricultural runoff. Beaver-dammed landscapes create habitats for other species, and their complexes can serve as wildfire breaks.

Researchers have calculated that between 15 million and 250 million beaver ponds once covered North America. A booming 19th-century trade in beaver pelts, along with a loss of habitat due to settlement, almost wiped them out. The landscape they helped to shape was a watery quagmire—and a classic example of shifting baselines. Because we didn't notice the initial profusion of beavers in the wild, we think current populations are normal.

Beaver populations recovered through the 1920s and 1930s, but never to their original numbers. *Eager*, by Ben Goldfarb, examines how they're getting along today.

Beavers still face obstacles when we attempt to reintroduce them into ecosystems in which they once flourished. Predators can eat a beaver for lunch, while cattle grazing removes vegetation and can alter a stream's configuration in such a way that it can no longer support beaver populations.



Eager
Ben Goldfarb
Chelsea Green, 2018.
302 pp.

Even wild ungulates can be a detriment to beavers. Goldfarb describes how the reintroduction of wolves into Yellowstone National Park led to the recovery of vegetation and rerouting of stream channels, which some researchers maintain was the result of an increase in the availability of streamside willows and alder for beavers. But one of the main barriers to beaver reintroduction is policy, especially in states such as New Mexico, where legislators fear beaver activity will harm the cattle industry.

Eager also highlights the problems of preconceived assumptions about beavers and beaver management. A common supposition among fish ecologists, for example, is that beavers are bad for salmon because their dams prevent the fish from swimming upstream. However, Goldfarb cites a comprehensive review of 108 papers that showed that beavers benefit fish populations more often than they cause negative consequences. A more egregious error occurred in California in 1937, when ecologist Joseph Grinnell declared that beavers had never inhabited large portions of the state. Although this was untrue, it went unchallenged and affected beaver recovery and ecosystem management across the state until it was disproven in 2012.

With drought in the southwest United States, there's talk of building new dams. Beavers could do that for us—and create a fully functioning ecosystem while they're at it. Goldfarb visits a property in Nevada, where a local rancher was able to extend his water availability by two months by allowing beavers to colonize one of his watersheds.

Later in the book, Goldfarb suggests that we could combine the heavy-duty approach of river restoration (such as use of backhoes and rip rap) with the work of beavers by installing beaver dam analogs (BDAs): two posts hammered into the ground with twigs woven between them (*Science*, 8 June, p. 1058). The BDAs, he argues, would speed up ecological recovery by encouraging beavers to colonize designated watersheds.

Goldfarb speaks largely with "beaver believers"—individuals who try to help humans and beavers coexist by mitigating the impact of beavers on the built landscape and by reintroducing them into stream systems that they can potentially restore. He lets his interviewees tell the majority of the story, recalling, for example, Councilman Mark Ross's interaction with a local businessman during a tense meeting of pro- and anti-beaver groups in Martinez, California (""This seventy-year-old guy is about to hit me! ...Do I hit back against a senior citizen or not?"") and Yellowstone scientist Dan Kotter's description of the effects of bison grazing on riparian vegetation (""like Jabba the Hutt eating a piece of pizza"").

Goldfarb ends the book with a trip to the United Kingdom, where beavers haven't been seen since the 17th (Scotland) and late 18th (England) centuries. Here, reintroduced beavers are a huge tourist draw, and beaver dams reduce the impacts of flooding—a big problem in the UK—although many farmers aren't convinced. As Goldfarb writes, "Everyone shares a goal; no one agrees on strategy."

One thing *Eager* was missing was a visit to Canada. The beaver is the country's national animal and graces its nickel. It has acres of landscape shaped by beavers, and Goldfarb cites a number of Canadian studies, but a firsthand experience would surely have enriched his otherwise excellent story. ■

The reviewer is a freelance science writer and editor and cofounder of Science Borealis, *Canada's science blog aggregator. Email: snowhydro1@gmail.com*

10.1126/science.aat7938

SOCIAL SCIENCE

In vivo we trust

An engaging synthesis highlights the value of field experiments in the social sciences

By **John A. List**

Social scientists once restricted their research to carefully controlled laboratory experiments. Over the past 25 years, however, they have increasingly made use of field experiments. Insights gained have spanned nearly every imaginable segment of our society, lending tests of theory, advice to policymakers, and guidance to nonprofit and for-profit firms alike.

In his new book, *Randomistas*, Andrew Leigh takes stock of a slice of this research in an even-tempered, scientific, and accessible way. The work reads like a stroll down memory lane, as Leigh digs into historically rich areas of research, ranging from the education production function to crime prevention and useful poverty interventions. Yet perhaps the most exciting aspect of the book concerns field experiments in politics and philanthropy.

Politics is interesting in its own right. Why people vote and how we can enhance voter participation remain first-order questions in well-functioning democracies. In chapter 9, Leigh cleverly illustrates how Barack Obama made keen use of field experiments to figure out what was working and why during his first presidential campaign. For instance, in 2007 Obama's team used a field experiment to determine the best image and slogan to use to motivate web visitors to subscribe to future campaign emails. The version that yielded the most email addresses—a black and white photo with the message "Learn More"—surprised even the most seasoned experts, who expected that a video accompanied by the message "Sign Up" would be the top performer, and served as a scientific basis for a successful campaign.

Later, Leigh describes how political scientists Alan Gerber and Donald Green determine the efficacy of get-out-the-vote interventions. As Leigh, a politician himself, points out: "I've met 'experts' who are convinced that partisan letters work best when paired with doorknocking, that telephone calls work best in the final week of the campaign, or that posters outside the election booth make a huge difference. But ask them about their evidence base and it's quickly apparent that their war stories lack a control group."

In their most effective intervention, Gerber and Green show that a letter revealing the recipient's turnout record, as well as their neighbor's, increases turnout by 8 percentage points. This remarkable effect likely reflects social-image concerns.

Charitable giving, meanwhile, is much more important than most people realize. The number of U.S. nonprofits registered with the Internal Revenue Service grew by nearly 60% from 1995 to 2005, and charitable gifts of money have more than doubled since 1990, now exceeding 2% of GDP.

The market for charitable giving primarily revolves around three major players: donors, who provide the resources to charities; charitable organizations, which develop strategies; and the government, which decides (among other issues) the tax treatment of individual contributions, the level of government grants given to various charities, and what public goods to provide itself.

Leigh catalogs interesting details of recent field experiments, which lend insights into all three actors, but he focuses most of his efforts on the relationships between charities and individuals. Here, he summarizes some of my own field experiments from the 1990s, which showed the importance of a "lead donor" and matching funds in raising charitable giving. (The mere mention of a lead donor raised giving by 50 to 100%.)

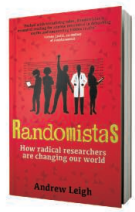
What is particularly appealing about *Randomistas* is that it does not stop at discussing the overarching literatures associated with the topics Leigh chooses to focus on. It also details how to build a better feedback loop and how organizations can get over hurdles that prevent them from running effective interventions. These include fairness concerns, replication needs, and the discomfort associated with admitting when you do not know something. (The latter is a challenging one for most managers I know.)

I do not have any qualms with what was written in this book, but I do feel that an important element of the field experiment equation was omitted. In the past few decades social scientists have done a superb job of developing methods with which to generate field data showing how the world works and detailing intervention effects. However, how we should use the data for policy purposes is often neglected. Do the results scale to a larger setting? What are the factors that affect that scaling? Without this information, empirical research can be quickly undermined in the eyes of the policy-maker, the broader public, and even within the scientific community. ■

10.1126/science.aau2825

The reviewer is at the Department of Economics, University of Chicago, Chicago, IL 60637, USA. Email: jlist@uchicago.edu

Randomistas
How Radical Researchers Are Changing Our World
Andrew Leigh
Yale University Press, 2018. 283 pp.



Bluffs at Albemarle Sound are rapidly eroding into the ocean, complicating efforts to find clues about the Lost Colony.

PHOTO: INGRID LEMME/WIKIMEDIA COMMONS

PODCAST

The Secret Token
Myth, Obsession, and the Search for the Lost Colony of Roanoke
Andrew Lawler
Doubleday, 2018. 410 pp.

Delayed for 3 years by an ongoing war, John White returned to the American colony of Roanoke on 18 August 1590 but found the settlement deserted. The only clue was the word "Croatoan" carved into a fence post. This week on the *Science* podcast, Andrew Lawler recounts the mystery of the Lost Colony and shares stories of those who've devoted their lives to finding it.

10.1126/science.aau6837

Negligence and habitat loss are among the threats that have rendered annual killifishes the most endangered fish in Brazil.

PHOTO: AUSCAPE/UIG VIA GETTY IMAGES

LETTERS

Edited by **Jennifer Sills**

Brazilian killifishes risk extinction

The annual killifishes are a diversified group of small fish in Africa and South America (*1*). Annual killifish are unique in their ability to survive in ephemeral pools; they have short, seasonal life cycles, laying eggs that lie dormant in the soil once the pools have evaporated and then resume development and hatch when the water returns (*2*, *3*). In Brazil, these fish are found in all biomes and have been recorded in the most variable seasonal freshwater environments (*2*). As a result of habitat loss, restricted distribution, and low dispersal of most species, annual killifishes face strong threats of extinction (*4*, *5*); they are the most endangered group of fish in Brazil (*5*, *6*). In a 2014 evaluation of endangered Brazilian fauna, according to the International Union for Conservation of Nature (IUCN) criteria (*5*), 102 species of annual killifishes were considered endangered, representing one-third of all threatened freshwater fish in Brazil. Some of these species are no longer found in nature and possibly extinct (*7*).

Several factors are responsible for the high degree of threat to killifish. The use of land for urbanization, agriculture, and hydroelectric dams has led to the loss and degradation of the pools in which the fish live (*5*–*7*). Many environmental agencies lack knowledge about the species and their habitats and sometimes have been negligent in their efforts to protect them (*8*). Because few killifish species live in protected areas, many of Brazil's conservation efforts have not benefited them (*8*). Consequently, annual killifish populations have declined rapidly in Brazil (*7*, *8*).

The problem is even greater because Brazil's policies hamper biodiversity conservation (*9*). The Brazilian Forest Act reduced the area of permanent preservation and overlooked ephemeral aquatic environments (*9*). Other policies, such as the New Law on Biodiversity, make access to biodiversity difficult and bureaucratic, thereby weakening universities and research institutions (*10*). Threats also include assaults on the environmental licensing system and disenfranchisement of environmental agencies (*11*). In a country where charismatic species are constantly suffering from population decline (*12*) and conservation research is under attack, there is little energy or attention left for small fish that inhabit extreme environments and are unknown to society.

Despite a national conservation action plan that has led to some small advances and greater visibility of killifishes in recent years (*4*), there is still much that needs to be done to mitigate these serious threats to Brazilian killifish biodiversity. Brazil must pass laws to protect temporary wetlands and facilitate public and private investments in scientific research and environmental education. Most urgently, the government must act to restore degraded areas and establish protected areas to save these species from extinction in Brazil.

Matheus Vieira Volcan* and Luis Esteban Krause Lanés
Instituto Pró-Pampa, Pelotas, RS 96010-630, Brazil.
*Corresponding author. Email: matheusvolcan@hotmail.com

REFERENCES

1. W. J. E. M. Costa, *Catalog of Aplocheiloid Killifishes of the World* (Reproarte, 2008).
2. W. J. E. M. Costa, *Peixes Anuais Brasileiros: Diversidade e Conservação* (UFPR, 2002) [in Portuguese].
3. N. Berois *et al*., in *Annual Fishes: Life History Strategy, Diversity, and Evolution*, N. Berois, G. García, R. O. de Sá, Eds. (CRC Press Taylor & Francis group, New York, 2015), pp. 33–45.
4. "Sumário Executivo do Plano de Ação Nacional de Conservação de Peixes Rivulídeos Ameaçados de Extinção" (ICMBio, 2013); www.icmbio.gov.br/portal/images/stories/docs-plano-de-acao/pan-rivulideos/sumario-executivo-rivulideos.pdf [in Portuguese].
5. "Portaria N° 445: Lista de espécies ameaçadas de extinção—peixes e invertebrados aquáticos" (2014); http://simat.mma.gov.br/acomweb/Media/Documentos/abbd1ad0-4aca-448b-a.pdf [in Portuguese].
6. R. Rosa, F. C. T. Llima, in *Livro Vermelho da Fauna Brasileira Ameaçada de Extinção*, A. B. N. Machado, C. M. Drummond, A. P. Paglia Eds. (Brasília e Belo Horizonte, MMA Fundação Biodiversitas, 2008), pp. 8–275. [in Portuguese].
7. W. J. E. M. Costa. *Biodivers. Conserv*. **21**, 2443 (2012).
8. M. V. Volcan *et al*., in *Annual Fishes: Life History Strategy, Diversity, and Evolution*, N. Berois, G. García, R. O. de Sá, Eds. (CRC Press Taylor & Francis group, New York, 2015), pp. 185–206.
9. R. Loyola, *Divers. Distrib*. **20**, 12 (2014).
10. F. A. Bockmann *et al*., *Science* **360**, 865 (2018).
11. S. Rodrigues, "MPF se posiciona contra nova lei de licenciamento ambiental," O Eco (2017); www.oeco.org.br/blogs/salada-verde/mpf-se-posiciona-contra-nova-versao-do-licenciamento-ambiental/ [in Portuguese].

12. A. G. Chiarello *et al.*, in *Livro Vermelho da Fauna Brasileira Ameaçada de Extinção*, A. B. N. Machado, C. M. Drummond, A. P. Paglia, Eds. (Brasília e Belo Horizonte, MMA Fundação Biodiversitas, 2008), pp. 680–880 [in Portuguese].

10.1126/science.aau5930

"Deadly mosquito" or "living freshwater"?

Organizations fighting disease often use catchy taglines such as "Which animal kills most people?," as the Gates Foundation did at its 2018 annual campaign against malaria (*1*). These questions pique our curiosity and anxiety and remind us that we can prevent disease by protecting ourselves from mosquitoes. However, casting these aquatic insects as villains could set back freshwater and invertebrate conservation. We, therefore, call for disease-focused organizations to adopt a message that puts human health more clearly in its environmental context.

Freshwater is life's and humankind's most critical resource, and shifts in precipitation and evaporation due to climate change directly affect aquatic habitats (*2*). As a result, it is also the environment with the highest proportion of threatened species (*3*). Invertebrates make up the vast majority of animal species, and insect populations are rapidly declining (*4*). Nonetheless, protecting this biodiversity is often an uphill battle, given that insects do not evoke the appreciation that benefits other ecosystems and species, such as forests and whales. Blood-sucking, disease-bearing insects are understandably unpopular, but exploiting our fear of them perpetuates the prejudice against their habitats.

Moreover, vilifying insects may detract from humans' own role in promoting mosquito-borne disease. Just a fraction of the 3500 known species of mosquitoes transmit the pathogens truly responsible for malaria, dengue, chikungunya, and other human diseases (*5*). These vectors have adapted to feed primarily on humans by developing in human-disturbed habitat that is often inhospitable to their natural predators. For example, we litter our yards with the rainwater-retaining plastic in which some of the most problematic mosquitoes thrive (*6*).

We encourage disease-focused organizations to work with ecologists and conservationists to expand their campaigns with the many positive messages that a healthy environment carries. Plastic-free yards would reduce the need for harmful pesticides, for instance (*6*). Wetlands are also home to attractive emissaries that eat mosquitoes, such as colorful dragonflies. The ultimate symbol of health, of course, is clean freshwater itself, undeniably humanity's most precious resource.

Klaas-Douwe B. Dijkstra,[1]* Maarten J. J. Schrama,[2] Erin E. Gorsich,[3] Axel Hochkirch[4]
[1]IUCN Freshwater Conservation Subcommittee, Naturalis Biodiversity Center, 2300 RA Leiden, Netherlands. [2]Institute of Environmental Sciences, Leiden University, 2333 CC Leiden, Netherlands. [3]Department of Biology, Colorado State University, Fort Collins, CO 80526, USA. [4]IUCN Invertebrate Conservation Subcommittee, Department of Biogeography, Trier University, D-54286 Trier, Germany.
*Corresponding author. Email: african.dragonflies@gmail.com

REFERENCES
1. N. Kristof, J. Ma, "The Deadliest Animal in the World," *New York Times* (2018); www.nytimes.com/interactive/2018/04/25/opinion/25kristof.html.
2. Intergovernmental Panel on Climate Change (IPCC), "Climate Change 2014: Synthesis Report" (IPCC, 2014).
3. B. Collen *et al.*, *Glob. Ecol. Biogeogr.* **23**, 40 (2014).
4. C. A. Hallmann *et al.*, *PLOS One* **12**, e0185809 (2017).
5. L. M. Rueda, *Hydrobiology* **595**, 477 (2008).
6. D. J. Gubler, G. G. Clark, *Acta Tropica* **61**, 169 (1996).

10.1126/science.aau5573

Where there is fire, there is smoke

In his In Depth News story "Scientists aim to smoke out wildfire impacts" (1 June, p. 948), W. Cornwall discusses wildfire-generated smoke and the associated large quantities of particulate matter. We agree that research on smoke demands more sustained scientific attention, particularly given its impacts on human communities [e.g., (*1*)]. However, not all fires are created equal in terms of their (natural or deliberate) origins, the amount of biomass they consume, and the amount of smoke they generate as a result.

High-intensity wildfires (those exceeding 200,000 kW/m^2) in very high biomass forests may consume 9 to 14% of the biomass (burning green parts of living trees but often leaving dense wood) or 40 to 58 tons per hectare (*2*). This is substantially less than half of the 140 to 450 tons of biomass per hectare consumed when logging slash (i.e., dense waste wood) is burned after logging operations in these same kinds of forest (*3*, *4*).

Moreover, such logging burns consume at least 10 times as much biomass as hazard reduction burns designed to reduce wildfire risk (*5*). The smoke generated from logging therefore represents a substantial form of industrial pollution, but it is not treated as such. Beyond improving efforts to quantify the chemical composition of smoke and its long-term impacts on human health, there is an urgent need to examine the problem in a more holistic manner. This demands far more careful consideration of the origins of smoke, the relative amounts of smoke originating from different kinds of fires, and the forest and land use policies that generate smoke (and might reduce it).

David B. Lindenmayer* and Chris Taylor
Fenner School of Environment and Society, The Australian National University, Canberra, ACT, 2601, Australia.
*Corresponding author. Email: david.lindenmayer@anu.edu.au

REFERENCES
1. W. E. Cascio, *Sci. Total Environ.* **624**, 586 (2018).
2. H. Keith *et al.*, *PLOS One* **9**, e107126 (2014).
3. R. J. Raison, R. O. Squire, "Forest management in Australia: Implications for carbon budgets" (Technical Report No.32, Australian Greenhouse Office, 2008).
4. A. Slijepcevic, *TasForests* **13**, 281 (2001).
5. K. G. Tolhurst, N. P. Cheney, "Synopsis of the knowledge used in prescribed burning in Victoria" (Department of Natural Resources and Environment, Melbourne, 1999).

10.1126/science.aau6672

TECHNICAL COMMENT ABSTRACTS

Comment on "An excess of massive stars in the local 30 Doradus starburst"

Will M. Farr and Ilya Mandel

Schneider *et al*. (Reports, 5 January 2018, p. 69) used an ad hoc statistical method in their calculation of the stellar initial mass function. Adopting an improved approach, we reanalyze their data and determine a power-law exponent of $2.05^{+0.14}_{-0.13}$. Alternative assumptions regarding dataset completeness and the star formation history model can shift the inferred exponent to $2.11^{+0.19}_{-0.17}$ and $2.15^{+0.13}_{-0.13}$, respectively.
Full text: dx.doi.org/10.1126/science.aat6506

Response to Comment on "An excess of massive stars in the local 30 Doradus starburst"

F. R. N. Schneider, H. Sana, C. J. Evans, J. M. Bestenlehner, N. Castro, L. Fossati, G. Gräfener, N. Langer, O. H. Ramírez-Agudelo, C. Sabín-Sanjulián, S. Simón-Díaz, F. Tramper, P. A. Crowther, A. de Koter, S. E. de Mink, P. L. Dufton, M. Garcia, M. Gieles, V. Hénault-Brunet, A. Herrero, R. G. Izzard, V. Kalari, D. J. Lennon, J. Maíz Apellániz, N. Markova, F. Najarro, Ph. Podsiadlowski, J. Puls, W. D. Taylor, J. Th. van Loon, J. S. Vink, C. Norman

Farr and Mandel reanalyze our data, finding initial mass function slopes for high-mass stars in 30 Doradus that agree with our results. However, their reanalysis appears to underpredict the observed number of massive stars. Their technique results in more precise slopes than in our work, strengthening our conclusion that there is an excess of massive stars (>30 solar masses) in 30 Doradus.
Full text: dx.doi.org/10.1126/science.aat7032

TECHNICAL COMMENT

STELLAR ASTROPHYSICS

Comment on "An excess of massive stars in the local 30 Doradus starburst"

Will M. Farr[1,2]* and Ilya Mandel[1,3]*

Schneider *et al*. (Reports, 5 January 2018, p. 69) used an ad hoc statistical method in their calculation of the stellar initial mass function. Adopting an improved approach, we reanalyze their data and determine a power-law exponent of $2.05^{+0.14}_{-0.13}$. Alternative assumptions regarding dataset completeness and the star formation history model can shift the inferred exponent to $2.11^{+0.19}_{-0.17}$ and $2.15^{+0.13}_{-0.13}$, respectively.

Schneider *et al*. (*1*) use spectroscopic observations of young massive stars in the 30 Doradus region of the Large Magellanic Cloud to infer a shallower-than-expected stellar initial mass function (IMF) with a power-law exponent of $\alpha = 1.90^{+0.37}_{-0.26}$, in contrast to the Salpeter exponent of 2.35 (*2*). They estimate the ages and masses of individual stars with the BONNSAI Bayesian code (*3*), then obtain an overall mass distribution by effectively adding together the posterior probability density functions of individual stars. There is no statistical meaning to a distribution obtained in this way, which does not represent the posterior probability density function on the mass distribution.

Hierarchical Bayesian inference provides the statistically justified solution to this problem (*4*). Mandel (*5*) specifically considered inference on a mass distribution given a sample of uncertain measurements, and we use a similar methodology here. We interpret the Schneider *et al*. inference on individual masses and ages as independent Gaussian likelihoods for the logarithm of the mass and the age, with parameters fixed by matching the mean parameter to the peak and the standard deviation parameter to the 68% width of the individual stellar distributions in the Schneider *et al*. data.

For our fiducial analysis, we model the star formation history as a truncated Gaussian distribution, and generally find a star formation history similar to that in Schneider *et al*., with the star formation rate in 30 Doradus peaking about 4 million years ago. We impose broad priors on the power-law exponent and the mean and standard deviation of the star formation Gaussian. We use the Hamiltonian Monte Carlo sampler STAN (*6*) to efficiently address the high-dimensional hierarchical problem with free parameters for each star's actual mass and age in addition to the IMF exponent and the mean and standard deviation of the star formation history.

Figure 1 shows the inferred power-law exponent of the IMF. We use the Schneider *et al*. fit to stellar lifetimes and assume that their dataset is complete above 15 solar masses ($M_\odot$); that is, we select only those stars whose observed mass is above $15M_\odot$ (*7*, *8*). We find an exponent of $\alpha = 2.05^{+0.14}_{-0.13}$ where the quoted value corresponds to the median of the posterior distribution and the range to the 16th and 84th percentiles (i.e., the symmetric 68% credible interval). Hierarchical Bayesian modeling steepens the preferred IMF slope; our median α value lies about 1σ above the preferred value from Schneider *et al*. This analysis narrows the uncertainty interval by more than a factor of 2.

The analysis above uses the same assumptions as Schneider *et al*. Below, we consider the impact of three additional assumptions: the stellar lifetime fit, the choice of the completeness limit, and the model for the star formation history.

We performed an independent fit to the main-sequence lifetimes τ_{MS} of nonrotating massive stars of mass M as modeled by Brott *et al*. (*9*) and Köhler *et al*. (*10*):

$$\ln\frac{\tau_{MS}(M)}{\mathrm{Myr}} = 9.1973 - 3.8955\ln\frac{M}{M_\odot} + 0.6107\left(\ln\frac{M}{M_\odot}\right)^2 - 0.0332\left(\ln\frac{M}{M_\odot}\right)^3 \quad (1)$$

Following Schneider *et al*., we increased the "observable" lifetime of a star by 10% beyond its main-sequence lifetime to account for helium burning. We find that this alternative fit does not affect the inferred IMF, yielding the same power-law exponent $\alpha = 2.05^{+0.14}_{-0.13}$.

The inferred power-law exponent is somewhat sensitive to the choice of the cutoff mass for survey completeness. The data of Schneider *et al*. show a relative scarcity of stars between $15M_\odot$ and $20M_\odot$; changing the mass cutoff from $15M_\odot$ to $20M_\odot$ further steepens the inferred exponent to $\alpha = 2.11^{+0.19}_{-0.17}$. However, these fluctuations are within the expected statistical variation based on the sample size, as confirmed with posterior predictive checking. In particular, there is no statistical evidence against the claim of Schneider *et al*. that the survey is complete for $M \geq 15M_\odot$.

Finally, we considered an alternative star formation history model—a double exponential with three free parameters: the time of the peak of the star formation rate, and the (possibly different) decay constants before and after the peak. This model allows for a sharper peak and longer tails than a Gaussian. This star formation rate history model is consistent with the data, as tested with posterior predictive checking (see below). However, it yields a power-law exponent $\alpha = 2.15^{+0.13}_{-0.13}$,

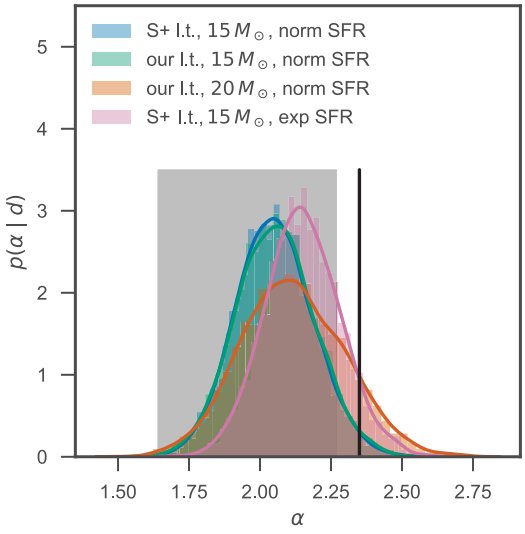


Fig. 1. The posterior on the IMF power-law exponent α is inferred from the observations *d*. See text for details on the four models. Blue: Schneider *et al*. [S+ (*1*)] stellar lifetimes (l.t.), survey completeness for $M \geq 15M_\odot$, and Gaussian star formation history model; green, same but with our lifetime fit; orange, same as green but with completeness for $M \geq 20M_\odot$; pink, same as blue but with a double-exponential star formation history model. The Salpeter power-law exponent is $-\alpha = -2.35$ (*2*), indicated by a vertical black line. The 68.3% range of power-law exponents derived by Schneider *et al*. is shaded in gray.

[1]Institute of Gravitational Wave Astronomy and School of Physics and Astronomy, University of Birmingham, Birmingham B15 2TT, UK. [2]Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA. [3]Monash Centre for Astrophysics, School of Physics and Astronomy, Monash University, Clayton, Victoria 3800, Australia.
*Corresponding author. Email: wfarr@star.sr.bham.ac.uk (W.M.F.); imandel@star.sr.bham.ac.uk (I.M.)

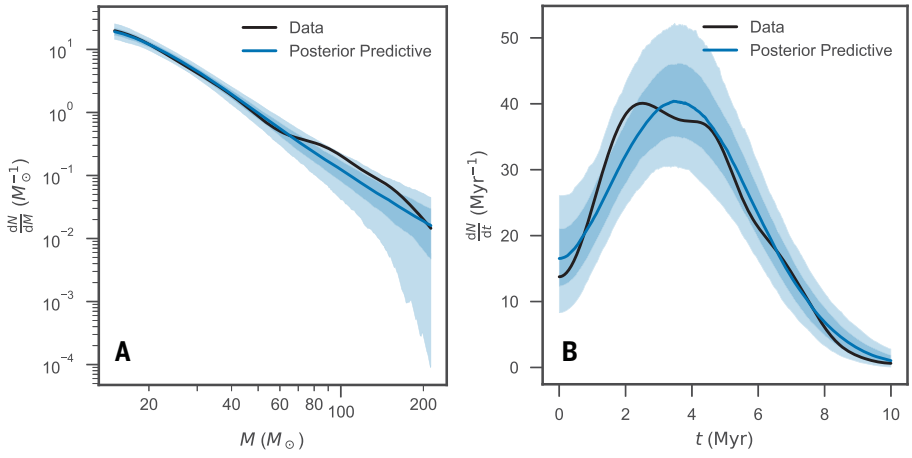


Fig. 2. Posterior predictive checking demonstrates that the observed data are consistent with being drawn from our model. The observed distribution of (maximum likelihood) masses M (**A**) and ages t (**B**) are shown as black curves; distributions of mass and age from synthetic data drawn from our fitted model (i.e., the posterior predictive distribution) are shown as blue curves (median) and shaded blue regions (68% and 95% credible intervals), respectively.

almost 1σ steeper than for our fiducial analysis. This indicates that the inferred IMF is sensitive to the systematics of the assumed star formation history model.

We also considered the possibility that the IMF power law has an additional break at higher masses, allowing for three free parameters: the mass at which the break happens, and the exponents below and above the break. However, we find that the data do not constrain the parameters of this more general model, and there is no statistical preference for a broken power-law model.

We confirmed the stability of our conclusions with posterior predictive checking. Figure 2 shows the distribution of observed masses and ages (i.e., the peak of the likelihood) from the Schneider *et al.* data overlain on the range of mass and age distributions that would be observed from a large number of datasets drawn according to our fitted fiducial IMF model. The data are consistent with being drawn from our model. We have also confirmed that all of our models yield predictions for the numbers of stars heavier than $30M_\odot$ and $60M_\odot$ that are consistent with observations.

We find that we can substantially reduce the statistical uncertainty in the IMF by applying an improved statistical analysis to the observations of young massive stars in 30 Doradus. However, the systematics from modeling uncertainties, such as the assumed star formation history model, can potentially shift the inferred power-law exponent by more than the statistical uncertainty. Furthermore, we adopted the mass and age posteriors for individual stars directly from Schneider *et al.* Imperfect stellar models or the inclusion of other complicating factors described by Schneider *et al.* (rotation, mass transfer, mergers, etc.) introduce further systematic uncertainty that could again shift the inferred IMF exponent. The combination of these factors makes it very challenging to infer the precise shape of the IMF even when a dataset as good as that obtained by Schneider *et al.* is available.

REFERENCES AND NOTES

1. F. R. N. Schneider *et al.*, *Science* **359**, 69–71 (2018).
2. E. E. Salpeter, *Astrophys. J.* **121**, 161 (1955).
3. F. R. N. Schneider, N. Castro, L. Fossati, N. Langer, A. de Koter, *Astron. Astrophys.* **598**, A60 (2017).
4. D. W. Hogg, A. D. Myers, J. Bovy, *Astrophys. J.* **725**, 2166–2175 (2010).
5. I. Mandel, *Phys. Rev. D* **81**, 084029 (2010).
6. B. Carpenter *et al.*, *J. Stat. Softw.* **76**, 1 (2017).
7. T. J. Loredo, *Am. Inst. Phys. Conf. Ser.* **735**, 195–206 (2004).
8. B. P. Abbott *et al.*, *Phys. Rev. X* **6**, 041015 (2016).
9. I. Brott *et al.*, *Astron. Astrophys.* **530**, A115 (2011).
10. K. Köhler *et al.*, *Astron. Astrophys.* **573**, A71 (2015).
11. T. P. Robitaille *et al.*, *Astron. Astrophys.* **558**, A33 (2013).
12. S. van der Walt, S. C. Colbert, G. Varoquaux, *Comput. Sci. Eng.* **13**, 22–30 (2011).
13. E. Jones *et al.*, SciPy: Open source scientific tools for Python; www.scipy.org/.
14. J. D. Hunter, *Comput. Sci. Eng.* **9**, 90–95 (2007).
15. M. Waskom *et al.*, mwaskom/seaborn: v0.8.1 (2017).

ACKNOWLEDGMENTS

We thank Schneider *et al.* (*1*) for making available for further study and analysis the data on which their conclusions are based, and F. Schneider personally for very useful discussions. This analysis made use of the PYSTAN (*6*), ASTROPY (*11*), NUMPY (*12*), SCIPY (*13*), MATPLOTLIB (*14*), and SEABORN (*15*) Python libraries. **Funding:** W.M.F. and I.M. are partially supported by the Science and Technology Facilities Council. **Author contributions:** W.M.F. and I.M. are jointly responsible for all aspects of this work. **Competing interests:** None. **Data and materials availability:** The code and LaTeX source used to prepare this document are publicly available under an open-source MIT license at https://github.com/farr/30DorIMF.

21 March 2018; accepted 4 June 2018
10.1126/science.aat6506

TECHNICAL RESPONSE

STELLAR ASTROPHYSICS

Response to Comment on "An excess of massive stars in the local 30 Doradus starburst"

Fabian R. N. Schneider[1]*, Hugues Sana[2], Christopher J. Evans[3], Joachim M. Bestenlehner[4,5], Norberto Castro[6], Luca Fossati[7], Götz Gräfener[8], Norbert Langer[8], Oscar H. Ramírez-Agudelo[3], Carolina Sabín-Sanjulián[9], Sergio Simón-Díaz[10,11], Frank Tramper[12], Paul A. Crowther[5], Alexander de Koter[2,13], Selma E. de Mink[13], Philip L. Dufton[14], Miriam Garcia[15], Mark Gieles[16], Vincent Hénault-Brunet[17,18], Artemio Herrero[10,11], Robert G. Izzard[16,19], Venu Kalari[20], Danny J. Lennon[12], Jesús Maíz Apellániz[21], Nevy Markova[22], Francisco Najarro[15], Philipp Podsiadlowski[1,8], Joachim Puls[23], William D. Taylor[3], Jacco Th. van Loon[24], Jorick S. Vink[25], Colin Norman[26,27]

Farr and Mandel reanalyze our data, finding initial mass function slopes for high-mass stars in 30 Doradus that agree with our results. However, their reanalysis appears to underpredict the observed number of massive stars. Their technique results in more precise slopes than in our work, strengthening our conclusion that there is an excess of massive stars (>30 solar masses) in 30 Doradus.

Farr and Mandel (*1*) reanalyzed the results of our study (*2*), in which we investigated the star formation history (SFH) and stellar initial mass function (IMF) of the local 30 Doradus (30 Dor) starburst in the Large Magellanic Cloud and found an overabundance of stars with initial mass exceeding 30 solar masses ($M_\odot$). They use an alternative and potentially more powerful statistical framework, hierarchical Bayesian inference, and infer IMF power-law indices for massive stars that are in agreement with our results (compare the IMF slope distributions in their figure 1 to the 1σ range inferred in our analysis). Their analysis allows them to infer the IMF slope with higher precision than was possible in our case, such that their inferred IMF slope for high-mass stars in 30 Dor is shallower than that of a Salpeter IMF (*3*) with an even larger confidence (more than 95.5%, versus 83% in our analysis). Their reanalysis therefore supports our main findings and conclusions about the IMF in 30 Dor.

Farr and Mandel's main criticism of our work is that "[t]here is no statistical meaning to [age and mass] distribution[s] obtained" by adding the posterior probability distributions of the ages and initial masses inferred for individual stars. It is true that such distributions are not posterior probability functions in a Bayesian framework. However, we caution that the IMF is historically defined as a histogram of stellar masses (*3*–*9*) and our procedure to add the posterior probability distributions of the initial masses of individual stars is the equivalent of computing a histogram for the mass distribution of a sample of stars, while taking into account the observational uncertainties of individual mass estimates. Virtually all IMFs inferred in the literature are constructed in this way, so Farr and Mandel's criticism implicitly applies to those as well. The VLT-FLAMES Tarantula Survey (VFTS) (*10*) has reached a completeness of about 73% with respect to a more complete census (*11*) of massive stars in 30 Dor (see figure S2 in our original work). For a complete stellar sample, the age distribution of stars obtained with our method would directly provide the SFH at the youngest ages where even the most massive stars did not yet end their nuclear burning lifetime—so there is also meaning to age distributions constructed as was done in our work.

We have tested our statistical analysis with mock data. To this end, we sampled a stellar population of 1000 stars more massive than $15M_\odot$ for a given Salpeter high-mass IMF with slope $\gamma = -2.35$ and a continuous SFH (constant star formation rate). In this way, we have obtained Gaussian distributions of the ages and masses of individual mock stars with 1σ uncertainties of 20% and 15% in age and mass, respectively. These uncertainties are characteristic of the age and mass uncertainties of stars in our sample of 30 Dor stars (*2*). We then used exactly the same analysis technique as in our original work to infer the IMF and SFH of the mock star sample. The results of this test are shown in Fig. 1 and demonstrate that our analysis method is able to reproduce the underlying SFH and IMF of the mock stars. For comparison, we show the distribution of initial masses for an IMF with slope $\gamma = -1.90$ to illustrate that our analysis technique can distinguish between a Salpeter IMF slope of $\gamma = -2.35$ and a shallower slope of $\gamma = -1.90$. This test further shows that both IMFs reproduce the mock data similarly well in the mass range 15 to $30M_\odot$ and that the high-mass end ($>30M_\odot$) of the distribution of mock masses has the largest power to constrain the high-mass IMF slope (Fig. 1C).

Our analysis of the VFTS data relies on two different techniques to infer the high-mass end of the IMF: (i) by fitting the observed distribution of stars in the mass range 15 to $200M_\odot$ and (ii) by fitting the number of stars more massive than 30 and $60M_\odot$. Both procedures give results that are in good agreement (*2*). From the inferred masses and corresponding uncertainties of our sample stars, we find $75.9^{+6.8}_{-7.0}$ stars above $30M_\odot$ and $22.2^{+4.0}_{-4.6}$ stars above $60M_\odot$ (*2*). Contrary to what Farr and Mandel write in their reanalysis, their online data (https://github.com/farr/30DorIMF, as accessed on 6 May 2018) suggest that their best-fitting SFH and IMF models underpredict the observed number of massive stars. They predict on average ~65 stars above $30M_\odot$ and ~18 stars above $60M_\odot$. Their ratio of the number of stars greater than $30M_\odot$ to the

[1]Department of Physics, University of Oxford, Oxford OX1 3RH, UK. [2]Institute of Astrophysics, KU Leuven, 3001 Leuven, Belgium. [3]UK Astronomy Technology Centre, Royal Observatory Edinburgh, Edinburgh EH9 3HJ, UK. [4]Max-Planck-Institut für Astronomie, 69117 Heidelberg, Germany. [5]Department of Physics and Astronomy, University of Sheffield, Sheffield S3 7RH, UK. [6]Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA. [7]Austrian Academy of Sciences, Space Research Institute, 8042 Graz, Austria. [8]Argelander-Institut für Astronomie der Universität Bonn, 53121 Bonn, Germany. [9]Departamento de Física y Astronomía, Universidad de La Serena, La Serena, Chile. [10]Instituto de Astrofísica de Canarias, E-38205 La Laguna, Tenerife, Spain. [11]Departamento de Astrofísica, Universidad de La Laguna, E-38206 La Laguna, Tenerife, Spain. [12]European Space Astronomy Centre, Mission Operations Division, 28691 Villanueva de la Cañada, Madrid, Spain. [13]Astronomical Institute Anton Pannekoek, Amsterdam University, 1098 XH Amsterdam, Netherlands. [14]Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK. [15]Centro de Astrobiología (CSIC-INTA), E-28850 Torrejón de Ardoz, Madrid, Spain. [16]Department of Physics, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK. [17]National Research Council, Herzberg Astronomy & Astrophysics, Victoria, BC V9E 2E7, Canada. [18]Department of Astrophysics/IMAPP, Radboud University, NL-6500 GL Nijmegen, Netherlands. [19]Institute of Astronomy, The Observatories, Cambridge CB3 0HA, UK. [20]Departamento de Astronomía, Universidad de Chile, Las Condes, Santiago, Casilla 36-D, Chile. [21]Centro de Astrobiología, CSIC-INTA, ESAC Campus, E-28 692 Villanueva de la Cañada, Spain. [22]Institute of Astronomy with National Astronomical Observatory, Bulgarian Academy of Sciences, 4700 Smoljan, Bulgaria. [23]Ludwig-Maximilians-Universität München, 81679 München, Germany. [24]Lennard-Jones Laboratories, Keele University, Staffordshire ST5 5BG, UK. [25]Armagh Observatory, College Hill, Armagh BT61 9DG, Northern Ireland, UK. [26]Johns Hopkins University, Homewood Campus, Baltimore, MD 21218, USA. [27]Space Telescope Science Institute, Baltimore, MD 21218, USA.
***Corresponding author. Email: fabian.schneider@physics.ox.ac.uk**

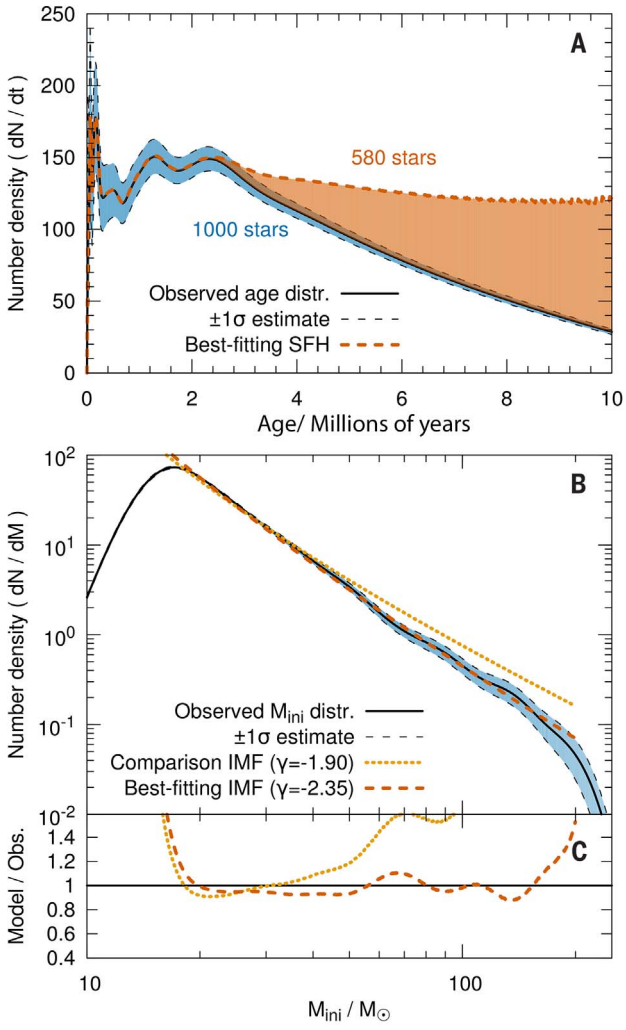


Fig. 1. Inference of the SFH and IMF of a mock stellar population. (**A** and **B**) Distributions of ages (A) and initial masses (B) of the mock stars (black lines) sampled from a Salpeter IMF with slope $\gamma = -2.35$ including bootstrapped 1σ estimates. The best-fitting IMF and SFH are indicated by the red dashed lines. For comparison, the predicted distribution of initial masses (M_{ini}) is shown for an IMF slope of $\gamma = -1.90$ (orange dotted line). (**C**) Ratio of the predicted model and "observed" mock initial mass distributions, showing that the two IMF models deviate from the mock data by more than the uncertainty only above $30M_{\odot}$.

number of stars greater than $60M_{\odot}$ (~3.6) is larger than what we have observed in 30 Dor (~3.4), which appears to be consistent with Farr and Mandel inferring slightly steeper IMF slopes than we did in our analysis. Indeed, using our SFH model and the results of our fitting method (ii), the numbers of massive stars above $30M_{\odot}$ and $60M_{\odot}$ as predicted by Farr and Mandel are found for an IMF slope of about $\gamma = -2.10$ [figure 2 in our original work (*2*)]. This is consistent with their best-fitting IMF slopes of $\gamma = -2.05$ to -2.15 for the different SFH models.

The reanalysis of Farr and Mandel gives systematically steeper IMF slopes than in our work and consequently seems to underpredict the observed number of massive stars in 30 Dor. We do not know the cause of this discrepancy. Our methodology appears to be robust, and the only other obvious difference in the two approaches—besides the statistical framework—is the assumption on the SFH. We directly infer the SFH from the data without making assumptions on its functional form. Farr and Mandel assume Gaussian and exponential SFH models that provide more degrees of freedom than in our case, and find IMF slope differences of $\Delta\gamma \approx 0.1$ depending on the assumed SFH model. This is a systematic uncertainty that we did not discuss in our original work, and that makes the inference of the IMF of composite stellar populations even more challenging.

REFERENCES AND NOTES

1. W. M. Farr, I. Mandel, *Science* **361**, eaat6506 (2018).
2. F. R. N. Schneider *et al.*, *Science* **359**, 69–71 (2018).
3. E. E. Salpeter, *Astrophys. J.* **121**, 161 (1955).
4. G. E. Miller, J. M. Scalo, *Astrophys. J. Suppl. Ser.* **41**, 513 (1979).
5. J. M. Scalo, *Fundam. Cosm. Phys.* **11**, 1 (1986).
6. P. Kroupa, C. A. Tout, G. Gilmore, *Mon. Not. R. Astron. Soc.* **262**, 545–587 (1993).
7. P. Kroupa, *Mon. Not. R. Astron. Soc.* **322**, 231–246 (2001).
8. G. Chabrier, *Publ. Astron. Soc. Pac.* **115**, 763–795 (2003).
9. N. Bastian, K. R. Covey, M. R. Meyer, *Annu. Rev. Astron. Astrophys.* **48**, 339–389 (2010).
10. C. J. Evans *et al.*, *Astron. Astrophys.* **530**, A108 (2011).
11. E. I. Doran *et al.*, *Astron. Astrophys.* **558**, A134 (2013).

ACKNOWLEDGMENTS

Funding: Supported by the Oxford Hintze Centre for Astrophysical Surveys, which is funded through generous support from the Hintze Family Charitable Foundation. Also supported by the FWO-Odysseus program under project G0F8H6N (H.S.); Deutsche Forschungsgemeinschaft grant GR 1717/5 (G.G.); the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement 665593 awarded to the Science and Technology Facilities Council (O.H.R.-A.); CONICYT-Chile through FONDECYT postdoctoral project 3170778 (C.S.-S.); Ministerio de Economía y Competitividad (MINECO), Spain, grants AYA2015-68012-C2-1 and SEV2015-0548 (S.S.-D. and A.H.); the European Union's Horizon 2020 research and innovation program from the European Commission under Marie Skłodowska-Curie grant agreement 661502 and European Research Council (ERC) grant agreement 715063 (S.E.d.M.); MINECO grants FIS2012-39162-C06-01 and ESP2015-65597-C4-1-R (M.Ga. and F.N.); the Royal Society (University Research Fellowship) and ERC (StG-335936, CLUSTERS) (M.Gi.); a Rutherford fellowship funded by STFC under grant ST/L003910/1 (R.G.I.); FONDECYT-Chile fellowship grant 3160117 (V.K.); MINECO grant AYA2016-75 931-C2-2-P (J.M.A.); and Bulgarian NSF grant DN08/1/13.12.2016 (N.M.). STScI is operated by AURA Inc. under NASA contract NAS5-26555. **Author contributions:** F.R.N.S. wrote the manuscript and all authors contributed to its discussion. **Competing interests:** None. **Data and materials access:** The stellar data are available in (*2*).

9 April 2018; accepted 4 June 2018
10.1126/science.aat7032

INSIDE

RELATED ITEMS

- EDITORIAL P. 313
- PERSPECTIVE P. 326

ILLUSTRATION: ANDY POTTS

MARVELOUS MODELS

By **Michael Funk, Colin Norman*, Keith T. Smith, Jelena Stajic,** *and* **Jake Yeston**

Scientists learn about the world by observing, manipulating, measuring, and abstracting. To make sure that they truly understand their system, and to gain insight beyond what experimental data can provide, many also turn to building mathematical models. Some models are based directly on fundamental physical laws, but most rely on approximations. The computational costs vary widely—from exactly solvable models to those that require all the computer power you can get. This power has become increasingly accessible; throw in the recent forays of artificial intelligence into scientific computing and the tantalizing promise of quantum computers, and it's clear that exciting times are ahead.

In this special issue, we highlight several areas of scientific computing that have seen recent progress. A feature article by *Science* reporter Paul Voosen focuses on an ambitious and controversial project, backed by high-tech philanthropists, to leverage breakthroughs in artificial intelligence, satellite imaging, and high-resolution simulations to create a new climate model. Kent and Kotliar reflect on two complementary approaches to predicting the properties of materials in which electrons are inextricably correlated with one another. Bottaro and Lindorff-Larsen report on the simulation of biological molecules, where recent progress has been driven by advances in theory and methods and tighter integration of experimental constraints. Sanchez-Lengeling and Aspuru-Guzik review approaches that use machine learning to design chemical structures that bear a desired characteristic. Finally, Brügmann describes how to simulate the complex equations that govern four-dimensional spacetime in Einstein's theory of general relativity, focusing on the example of gravitational waves produced by merging black holes.

The availability of computational resources enables the simulation of increasingly intricate models in many fields of science.

**Colin Norman is a former* Science *News Editor.*

THE EARTH MACHINE

With cash from tech philanthropists, science insurgents plot climate models driven by artificial intelligence

By **Paul Voosen**

Low stratocumulus clouds, like these above the northwest Pacific Ocean, are difficult to model.

Sometimes it seems the clouds over climate science just won't lift. Computer models of Earth's climate have multiplied in number, complexity, and computational power, yet they remain unable to answer more precisely some of the questions most on the public's mind: How high must we build sea walls to last until 2100? How bad will heat waves get in the next decade? What will Arctic shipping routes look like in 2030? Climate models all agree that global temperatures will continue to rise in response to humanity's greenhouse gas emissions, but uncertainties stubbornly persist over how quickly that will happen and how high temperatures will go.

Tapio Schneider, a German-born climate dynamicist at the California Institute of Technology (Caltech) in Pasadena, believes climate science can do better. And he's not alone. Later this summer, an academic consortium led by Schneider and backed by prominent technology philanthropists, including Microsoft co-founder Paul Allen, will launch an ambitious project to create a new climate model. Taking advantage of breakthroughs in artificial intelligence (AI), satellite imaging, and high-resolution simulation, that as-yet-unnamed model—the Earth Machine is one candidate—aims to change how climate models render small-scale phenomena such as sea ice and cloud formation that have long bedeviled efforts to forecast climate. A focus will be on the major source of uncertainty in current models: the decks of stratocumulus clouds that form off coastlines and populate the trade winds. A shift in their extent by just a few percentage points could turn the global thermostat up or down by a couple of degrees or more within this century—and current models can't predict which way they will go.

Within 5 years, the team hopes its AI-fortified model will drive out that uncertainty and others by learning on its own how clouds behave, from both actual observations and purpose-built cloud simulations. It's a lofty goal, Schneider admitted late one May afternoon in sun-soaked Pasadena, sitting outside with his newly assembled team. They had just wrapped up a workshop, the third he had convened in the past year, bringing together leading climate scientists and engineers to discuss the future of their field. "We're under no illusions," Schneider said. "This is not going to be a cakewalk."

There are reasons for skepticism. The United States already has many climate models, and some people question why it needs another, further dividing resources. Others question the technology and wonder whether the philanthropists backing the project have given it the scrutiny that an agency such as the National Science Foundation would provide. The team's unorthodox message and means won't make it easy to win people over, says David Randall, a climatologist at Colorado State University in Fort Collins. "I think the existing modeling centers will push back. If Tapio is getting funding, that in principle could have gone to someone else."

Tapio Schneider hopes his planned global climate model will reduce uncertainties over how high and how fast temperatures will rise.

PHOTOS: STEPHANIE DIANI FOR CALTECH; (OPPOSITE PAGE) STOCKTREK IMAGES, INC./ALAMY STOCK PHOTO

CLIMATE MODELERS have always followed two imperatives. First, they've folded ever more features of Earth into their simulations. Models once contained only the atmosphere and ocean; now, they have subroutines for ice sheets, land use, and the biosphere. Second, they've sought higher and higher resolutions—modeling interactions on smaller and smaller scales—riding the wave of Moore's law on government-owned supercomputers. By one estimate, the computing power those models use has increased by a factor of 100 million since the 1970s. As the models grew increasingly complex, they more fully reflected the vagaries of our planet—unknown unknowns turned to known unknowns. Yet the uncertainties remained.

At their most basic, all the models work the same way: They take the globe and chop it into a mesh, with cells some 25 kilometers to 50 kilometers on a side, and use a set of code called a dynamical core to simulate the behavior of the atmosphere and ocean over years and centuries. But much of what happens on the planet—cloud formation, for example—arises at scales smaller than those grids. Therefore, those phenomena have to be described indirectly—"parameterized" in the jargon of climate science—with rule-of-thumb equations. The modelers then adjust those various knobs to best represent the world as they know it—a process called tuning. "It's a mix of intuition and empiricism and some physically observed laws," says Isaac Held, Schneider's mentor and a scientist at the Geophysical Fluid Dynamics Laboratory, a prominent modeling center in Princeton, New Jersey.

Make no mistake: Current models do an admirable job of re-creating the world. But their shortcomings drive scientists bonkers. They struggle to re-create Arctic temperatures and melting sea ice. Their distribution of rainfall is off, biased against the extreme torrents that can cause flooding. "The rain is falling in the wrong place and at the wrong rate," says Paul O'Gorman, an atmospheric scientist at the Massachusetts Institute of Technology (MIT) in Cambridge, who formerly worked with Schneider. And, especially important, the models often fail to simulate those thick stratocumulus clouds, which typically form off the coasts of the western Americas and help cool the region.

Schneider, 46, has not always been fixated on clouds. Early in his career at Caltech, he focused on large-scale atmospheric flows, such as the Hadley cell. That atmospheric conveyor belt shifts air from the equator to the subtropics—the type of pattern that climate models can simulate using simple laws of physics. But while on an appointment at ETH Zurich in Switzerland, he became increasingly convinced that climate models could do a better job integrating new data on cloud behavior. He returned to Caltech in 2016 to seek a solution, adding a joint appointment at NASA's Jet Propulsion Laboratory (JPL) in Pasadena, where he had become a close collaborator with one of JPL's cloud gurus, João Teixeira.

That was the start of what is now a collaboration of about two dozen people. AI, particularly a variant called machine learning, was on the upswing, and Schneider and Teixeira mused that it might help with the cloud problem. Soon they recruited Andrew Stuart, a

soft-spoken computational mathematician at Caltech. The team found additional recruits at JPL, which has a vast archive of satellite data on clouds, and across the country at MIT, where researchers had built an ocean model infused with every possible satellite and buoy measurement of the seas.

The MIT group had ambitions to go bigger, and its members welcomed Schneider's overture. "Always the idea was to go to an Earth system model," says MIT physical oceanographer Raffaele Ferrari. "But the atmospheric community wasn't particularly primed to think the same way."

At first, the nascent collaboration was not set on creating a new climate model; the United States already has six prominent models. "It was more a question of how can we build a better model," Schneider says. But they wanted to be certain that a full climate model would incorporate their innovations. They decided the best way would be to build a new model, albeit one starting with existing code. Doing so meant they needed a computation whiz who could take their equations and make them run on a next-generation supercomputer.

A U.S. Navy expert reported for duty. Frank Giraldo, an applied mathematician at the Naval Postgraduate School in Monterey, California, is behind the Navy's new dynamical core, the mathematical engine at the heart of its next-generation weather and climate models. His core, the Non-hydrostatic Unified Model of the Atmosphere, is designed from the ground up for modern parallel computing. The core is also flexible and self-contained. It can solve equations to various degrees of accuracy in the same model, which should allow the Earth Machine to give a low-resolution overview of the planet while zooming in on clouds in real time.

CRUCIAL INPUT to the new model will come from simulations that have recently painted a much sharper picture of the low clouds and how they behave. Called large eddy simulations (LESs), those models trade the global scale and centuries-long time horizons of a climate model for narrow scope and high resolution. The models re-create several days in the life of small parcels of the atmosphere, with cells only 10 meters on a side. At such resolutions, key aspects of cloud formation—such as the convection that lofts sun-heated air upward until the water vapor it carries condenses into clouds—arise directly from physical laws. The results sometimes closely resemble reality, says Chris Bretherton, a leading cloud scientist at the University of Washington in Seattle.

Several years back, Bretherton led a project that used LESs to study how a 2°C temperature rise affected low ocean clouds. Two feedbacks emerged, both of which would exacerbate warming. First, higher temperatures appeared to allow more dry air to penetrate thin clouds from above, preventing them from thickening and reflecting more of the sun's energy. Second, increased carbon dioxide (CO_2) levels trap heat near the clouds' tops, preventing their cooling. Because such cooling drives the turbulence that forms clouds, the effect could impede cloud formation, fueling further warming. If CO_2 emissions continue unabated, Bretherton says, "It's possible that most of our low clouds in the tropics would melt away."

Other evidence, including actual cloud observations, also suggests "that the low-cloud feedback is positive and that low clouds will amplify climate warming," says Stephen Klein, an atmospheric scientist at Lawrence Livermore National Laboratory in California.

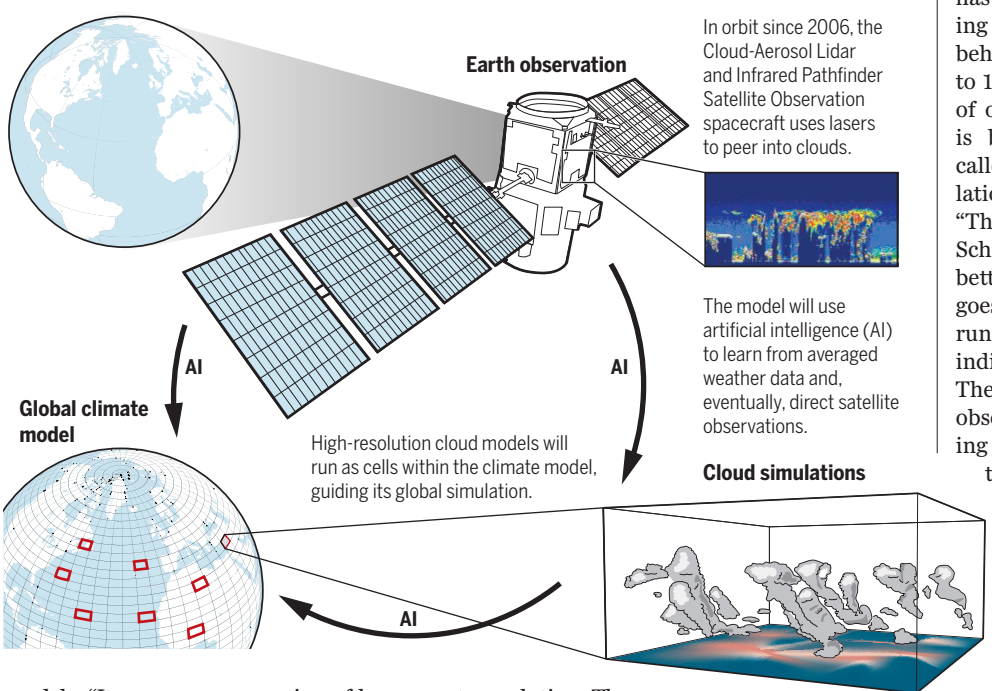
Those breakthroughs have not yet made their way into global models because no bridge, or technical way to get them there, has emerged. But Schneider's team is building one: an LES that can simulate cloud behavior over days within a domain of up to 100 kilometers on a side—about the size of one cell in a climate model. Their LES is based on a Caltech-developed model called the Python Cloud Large Eddy Simulation (PyCLES) that focuses on low clouds. "These simulations may not be perfect," Schneider says, "but they're much, much better than anything else we have." If all goes according to plan, Giraldo's code will run 1000 PyCLES-type models on the fly as individual cells inside the Earth Machine. The machine will also use AI to study the observed and simulated clouds, extrapolating what it learns to improve the rules of thumb it uses to simulate clouds across the globe. Soon, virtual cloud decks will sprout off the California coast.

INCORPORATING AI into climate modeling is a work in progress. Several researchers, including Bretherton's group and Michael Pritchard, a climate modeler at the University of California, Irvine, trained one form of AI, neural networks, on high-resolution simulations of the atmosphere. They then used the AI to replace several classic rules of thumb, such as how quickly the temperature and humidity change in rising air. "All of these are in the feeling-around type of phase," Bretherton says.

But neural networks and climate are an uneasy fit. The algorithms do best on problems such as classification—for example, learning from millions of labeled photos what a dog looks like. The code builds up an intricate model for what an object looks like that is often wholly inexplicable

Learning the climate

A new data-driven climate model will use satellite observations and high-resolution simulations to learn how best to render its clouds. Similar methods will also be applied to other, small-scale phenomena, such as sea ice and ocean eddies.



GRAPHIC: T. SCHNEIDER *ET AL.*, *GEOPHYSICAL RESEARCH LETTERS* **44**, 12,396 (2017). ADAPTED BY N. DESAI/*SCIENCE*

to human reasoning. The approach works for dogs—but may break down when it encounters something outside its training data—say, a camel. And for climate change, the future is a camel. For that reason, Stuart and Schneider are not banking on neural networks to guide the Earth Machine's AI. Instead, they seek a compromise, something between traditional rules of thumb and pure AI. They hope to develop code that can use hard-won knowledge of clouds and then fill in gaps with its learned intuition, essentially replacing the manual tuning typically done by modelers with a machine.

That learning won't be driven by individual cloud systems, which are imprinted with the atmospheric chaos that begat them. Rather, the AI will learn from seasonal or annual statistics on cloud coverage and other factors, wiping out the noise of weather. As Stuart and Schneider move each rule of thumb over to the AI's hands, they'll wire the model to calculate probabilities, allowing an overall reckoning of uncertainty not yet seen in current climate models. And, partly at the prompting of their engineering-minded funders, they'll develop metrics to gauge how accurately the model renders the world. They're betting on recent insights that, for some aspects of the climate system, short-term accuracy in a model indicates decades-long viability.

Success is far from guaranteed, the team agreed after the May workshop. "It could be that what we do ends up not improving the numbers, just to be completely scientifically honest," Stuart said. But even so, he added, the approach should spark new ideas across climate modeling. "I'd say that's the worst-case scenario," Schneider quickly interjected. "This is why I say we can't fail entirely. But I do hope we will do more than that."

AS SCHNEIDER ASSEMBLED his team and developed a general plan, he still faced a big question: Who would support their dream? That the U.S. government would finance yet another climate model seemed unlikely. Even before President Donald Trump's White House proposed cuts in climate science, former President Barack Obama's administration had explored whether the country needed to support so many models.

Fortunately for the Caltech team, the tech philanthropists—particularly Allen, who has already invested heavily in oceanography—were looking for something to make a splash. They sought a risky investment with a big potential payoff that could make climate forecasting more concrete. Schneider already had preliminary support for the Earth Machine from Charles Trimble, a Caltech alumnus who miniaturized the GPS receiver, and the Heising-Simons Foundation in Los Altos, California. But to reach their full ambitions, they needed more, some $5 million annually—a goal that now seems in sight, though the exact financing was still being finalized at press time.

The mix of ambition, metrics, and innovation embodied in the Earth Machine was exactly the type of work that Allen wants to fund, says Chris Emura, a Seattle, Washington-based computer engineer leading Allen's engagement with Schneider. Over the past year, Allen's team has been enmeshed in the modeling world, visiting leading centers to gauge what they can and can't do. Schneider's project, Emura says, has some audacity to it, with a high degree of "responsible risk." The team also has garnered interest from the Windward Fund, a conservation charity in Washington, D.C., that began an effort last month to support work that improves near-term practical climate predictions. And Schneider's endeavor has won praise from Schmidt Futures in New York City, the science-focused philanthropy of former Google CEO Eric Schmidt and Wendy Schmidt, president of the Schmidt Family Foundation in Palo Alto, California. "It's an attractive blend of conservative and bold approaches," says Stuart Feldman, the philanthropy's chief scientist.

As rumors of the Earth Machine have spread, the project has drawn a mix of support, envy, and skepticism. A new approach like that is desperately needed, says Trude Storelvmo, an atmospheric scientist at the University of Oslo. "This is a very welcome and innovative idea." She adds that it could bolster the case for expanded observations of clouds—necessary because NASA's current cloud satellites have worked nearly a decade longer than planned.

In contrast, Amy Clement, a cloud scientist at the University of Miami in Florida, laments the focus on building more complex models. "As a result, in my opinion, we are losing a lot of our ability to gain fundamental understanding of the climate system." However, she adds, given Schneider's acumen as a climate scientist, the model might lead to such understanding.

Bretherton, meanwhile, likes the group's ambitions but questions whether a new model is needed to realize them. "We already have too many climate models in the United States," he says. "It divides our resources and makes scientific progress slower."

Other people think the project is discounting rewards that will come when existing models are pushed to run globally at higher resolutions. Much of the climate science community in Europe, for example, is invested in a proposal called Extreme Earth, which would push models to a resolution of 1 kilometer per cell. Although such code would require a network of supercomputers and wouldn't run as long as traditional models, it would also eliminate many parameters that Schneider is seeking to improve with AI, replacing them with physics. "I'm so frustrated with the idea of parameterizing these things," says Bjorn Stevens, a climate scientist at the Max Planck Institute for Meteorology in Hamburg, Germany. "What I find more exciting is getting rid of those rules of thumb."

There's also a big assumption baked into the Earth Machine: that the cloud problem can even be solved, adds Joel Norris, a cloud scientist at the Scripps Institution of Oceanography in San Diego, California. Perhaps any sort of parameterization, even one tuned by AI, cannot crack clouds to a meaningful degree. "It may be the case you can't reduce the uncertainty," Norris says. Some satellite observations essential to rendering clouds, such as the exact location of water vapor in the lower atmosphere, simply don't exist. And Schneider's team could be shocked when it sees how apparently unconnected parts of the model go awry when clouds are tweaked, Held adds. "There's just a lot of connections."

Schneider's team is aware of all those concerns and shares many of them. But the members are ambitious and have grown impatient waiting for a breakthrough. They've lived with human-driven climate change, and its dogged uncertainties, as a reality for their entire adult lives. It's time for the clouds to lift. ■

Key uncertainties

Although climate models provide a fairly faithful rendering of the world, computational limits lead to persistent errors that have proved difficult to fix. Low stratocumulus clouds are a big problem. Some others include:

STORMS

A band of thunderstorms encircles the globe near the equator, shifting north or south depending on the season. Models have struggled to represent the intertropical convergence zone, as it's known, often producing two bands, one on each side of the equator.

ICE

Current models vary widely in their Arctic temperatures compared with reality, leading to large differences in their simulated sea ice extent.

RAIN

Models tend to simulate more light rain than seen in reality and underestimate the frequency of severe downpours and floods.

REVIEW

Toward a predictive theory of correlated materials

Paul R. C. Kent[1] and Gabriel Kotliar[2,3]*

Correlated electron materials display a rich variety of notable properties ranging from unconventional superconductivity to metal-insulator transitions. These properties are of interest from the point of view of applications but are hard to treat theoretically, as they result from multiple competing energy scales. Although possible in more weakly correlated materials, theoretical design and spectroscopy of strongly correlated electron materials have been a difficult challenge for many years. By treating all the relevant energy scales with sufficient accuracy, complementary advances in Green's functions and quantum Monte Carlo methods open a path to first-principles computational property predictions in this class of materials.

Predicting materials properties, starting from first principles, involves solving the quantum many-body problem in real materials. This has been a long-standing goal in theoretical and computational physics, chemistry, and materials science. The challenge was laid out by Paul Dirac in his famous 1929 quotation, "The underlying physical laws necessary for the mathematical theory of ... the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble," which was followed by a call for methodological developments: "Approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation" (*1*), a task which has kept researchers occupied for almost a century.

Two great advances toward meeting this goal were recognized by the 1998 Nobel Prize in Chemistry, awarded to John Pople and Walter Kohn. The first centered on constructing accurate approximations to the electronic wave functions from which one can construct directly all the observables of interest. The second focused on building functionals of electron density, which when minimized yield the physical density and the total energy of the ground state (*1*). Extensions of these methods have been successful in a large number of cases, but some long-standing difficulties remain. The wave function–based techniques, such as configuration interaction methods, have a computational cost that increases rapidly with system size. This makes it difficult to apply these methods to solids, which owe their emergent properties to their very large number of particles. Although density functional theory (DFT) is formally exact, improving the approximate functionals used in practice is difficult (*2*). The long-established and standard implementations of the density functional methods, such as the local-density approximation (LDA) or generalized gradient approximation (GGA), can treat the energetics of systems in the thermodynamic limit, but they do not capture the atomic character of the excitations in the so-called strongly correlated electron systems. A hallmark of strongly correlated materials is the presence of partially filled d or f electronic shells that are energetically situated to participate strongly in determining the material's electronic, magnetic, or structural properties. The strong and sensitive relationship between these properties requires them to be self-consistently determined, potentially amplifying any small errors or approximations made. Today, the standard approximations to DFT are not sufficiently robust, necessitating complementary approaches in these materials.

Substantial progress in methods development for strongly correlated electron systems has taken place over the last three decades, owing to the introduction of new concepts such as quantum embedding, novel algorithms, and the exponential increase in computational power. The solution of the quantum many-body problem has progressed along two complementary paths: the quantum Monte Carlo (QMC)– and Green's function–based methods. In the former, one focuses on describing the full wave function or density matrix of a given material, with the use of statistical methods to treat the large number of variables. In the latter, one computes

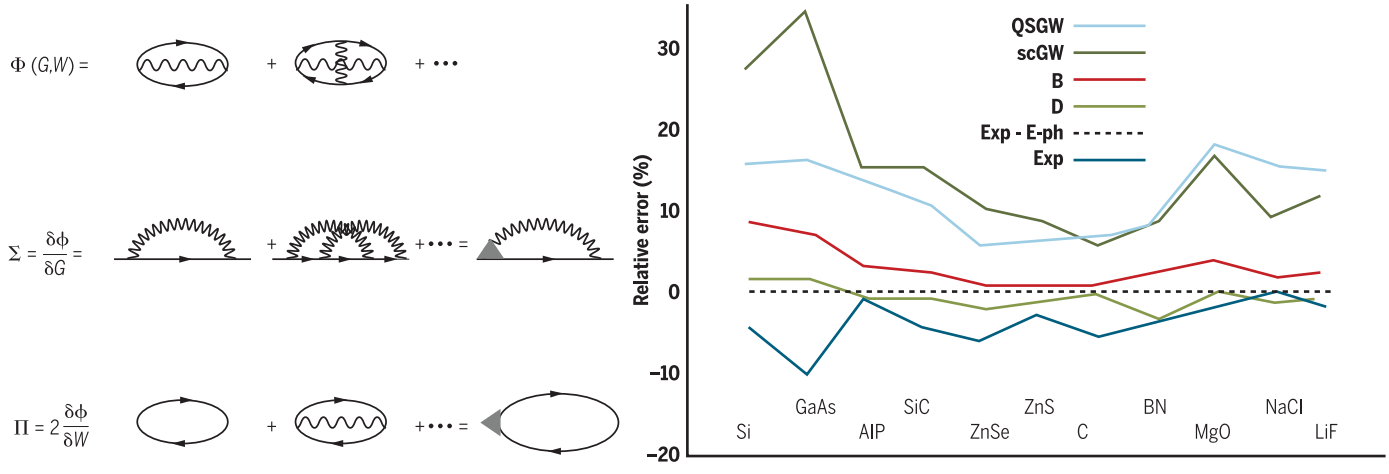


Fig. 1. Green's function diagrammatic schemes. (Left) Hedin's diagrammatic scheme for the free energy Φ, the self-energy Σ, and the polarization Π. Calculating only the first diagram in the expansion yields the self-consistent *GW* (sc*GW*) approximation. Inclusion of the second diagram is denoted by $GW\Gamma_1$. Other approximations, such as the quasiparticle self-consistent *GW* (QS*GW*) approximation, which is not fully self-consistent, or scheme D, which resumes the second-order vertex via the Bethe-Salpeter equation, work very well because of the cancellation of errors. (Right) Relative deviation of calculated semiconductor gaps from the experimental band gaps (excluding the electron-phonon renormalization). Different Green's function–based approximations were employed: QSGW, scGW, GW Γ_1 (scheme B), and scheme D. The experimental data (Exp) and the experimental data corrected by the effects of electron-phonon interactions (Exp - E-ph) are shown. These results are based on the approximations described in reference (*12*). Adapted with permission from (*12*), copyright (2017) by the American Physical Society.

IMAGE: ADAPTED BY N. DESAI

Green's function, a function of a smaller number of variables, from which one can extract observable information. Despite the different paths, these methods have now progressed sufficiently to inform each other. In this brief review article, we discuss these approaches critically, together with recent notable achievements.

Green's function approach

The quantum-mechanical Hamiltonian is an operator that represents the energy and governs the dynamics of the electrons in a solid. It can be written down explicitly; however, the evaluation of its ground-state wave function is impractical because it is a function of a very large number of coordinates (on the order of 10^{23} in a typical-size crystal) that cannot be realistically stored or computed. The Green's function approach shares the spirit of the DFT, as it formulates the exact solution of the quantum many-body problem in terms of a function, G, of a small number of variables. Furthermore, the approach works in the thermodynamic limit of a very large number of electrons. An advantage of this group of methods is that the form of the functional expressing the free energy in terms of G and the Coulomb interaction V_c is known explicitly as a sum of an infinite set of Feynman diagrams, which is a visual way to represent the physical processes of interacting electrons in a solid. The functional was first written down by Luttinger and Ward (*3*) and Baym and Kadanoff (*4*); its truncation gives rise to various approximate solutions to the quantum many-body problem (*5*).

Green's function $G(\omega)$ is an operator whose matrix elements $\langle r'|G(\omega)|r\rangle = G(\vec{r},\vec{r'},\omega)$ are defined as the Fourier transform with respect to time of the quantum-mechanical amplitude for an electron to propagate from point $\vec{r}$ to point $\vec{r'}$. It is expressed in terms of another operator, the self-energy $\Sigma(\omega)$.

$$G(\vec{r},\vec{r'},\omega) = \frac{1}{\omega + \nabla^2 + \mu - V_H - V_{cryst} - \Sigma(\omega)} \quad (1)$$

Here, V_H and V_{cryst} are operators representing the Hartree and crystalline potentials, respectively, and μ represents the chemical potential. Important concepts in this approach are the screened interaction in the solid, denoted by W, which represents the Coulomb interaction corrected by the effect of the motion of the electrons in the solid through their density-density response, $W^{-1} = V_c^{-1} - \Pi$, where Π is the polarization. A different concept in this field is the notion of a partially screened local Coulomb interaction, or Hubbard tensor U_{abcd}, which is the matrix element of a partially screened Coulomb interaction in a set of atomic centered orbitals $\chi_{\beta\vec{R'}}(\vec{r'})$.

The diagonal part of this tensor is called the Hubbard parameter U and describes the energetic cost of two electrons residing in the same atomic orbital, due to their average Coulomb repulsion. It first appeared in the Hubbard model, a simplified model that captures the physics of the competition between the kinetic energy, which delocalizes the electrons, and the Coulomb repulsion, which favors their spatial localization. The remaining parts contain the physics of the Hund's rule coupling J, which originates from the intra-atomic exchange interaction between electrons in different orbitals; J reduces the Coulomb repulsion energy when one has a configuration of electrons with parallel spins. Various methods to estimate the Hubbard tensor have been proposed (*6*, *7*), but a precise definition that lends itself to an accurate evaluation for use with embedding methods is an open problem and an active area of research (*8*–*10*).

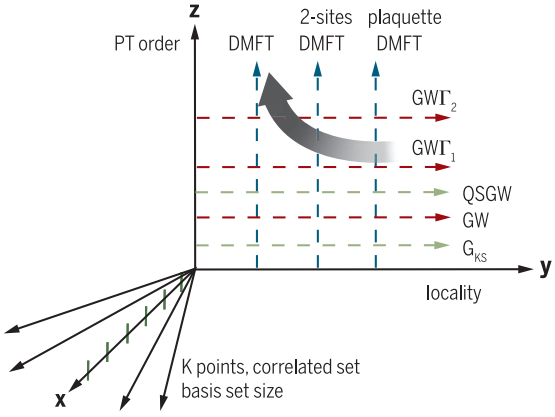


Fig. 2. The multidimensional computational complexity of Green's function approaches. Green's function calculations have to be converged along multiple axes to reach the full solution of the quantum many-body problem. The *z* axis denotes the order of Hedin's systematic PT in the screened Coulomb interaction *W* (blue dashed lines). The first-order PT leads to the *GW* approximation. The next perturbation order, GWΓ$_1$, has only recently been computed (*12*). Other approaches, such as the those exploiting the DFT Kohn-Sham Green function or the QS*GW* methods, are indicated with dashed green lines. An alternative approach is to sum all local diagrams to infinite order in PT, as done within the single-site DMFT. This local approximation can be improved through cluster extensions, as indicated schematically on the *y* axis (labeled by locality), which parametrizes the level of locality of the diagrams. Unlike model Hamiltonian studies, realistic electronic structure calculations involve numerous discretizations and many other parameters that need to be converged, such as the size of the basis set, the number of *k* points, and the dimension of the space of correlated orbitals. We represent the reduction of these discretization errors as increasing along the *x* axis. Future advances will involve different choices of locality and discretization errors at different orders in the PT. We represent this path as an arrow in the three-dimensional space. Adapted from (*94*).

The pioneering work of Hedin (*11*) outlined a systematic perturbation expansion in G and W, which is sketched in the left panel of Fig. 1. The lowest-order term for the self-energy gives rise to the so-called *GW* approximation. The full self-consistent evaluation of the next term has been achieved only recently (*12*). In principle, higher-order corrections are computable numerically. If for a given material the series is asymptotic, one can hope to obtain very accurate answers. In spite of the appeal of a systematic expansion, it turns out that for certain physical quantities it is better to avoid the self-consistency or achieve only partial self-consistency to obtain better agreement between theory and experiment (Fig. 1). Better agreement between the calculated and experimental band gaps of semiconductors can be obtained by using the DFT Kohn-Sham Green functions G_{KS} in the evaluation of the self-energy and the polarization (the one-shot or G_0W_0 method) (*13*) or a quasiparticle self-consistent Green function defined on the real (*14*) or imaginary (*15*) time axis. It is also sometimes necessary to carry out partial summations of diagrams to satisfy conservation laws. Therefore, implementing this formalism continues to be an art as well as a science; finding an appropriate path in the space described in Fig. 2 is important for materials at the edge of the research frontier, as it is not possible to check convergence simultaneously along the three axes of Fig. 2. With an increase in computational power and improved algorithms, we can envision a future where this convergence can be routinely checked systematically.

To understand the success of this perturbative approach, it is useful to return to the Fermi liquid theory, which justifies our thinking of the excitations of a solid in terms of quasiparticles. In many cases, the Kohn-Sham DFT is a good starting point for computing the quasiparticle properties in perturbation theory (PT) in the screened Coulomb interactions, and the lowest order in perturbation around the Kohn-Sham DFT is close to the experimental answers. On the other hand, there are materials where the effective Coulomb interaction is too large for this approach to be meaningful, as the excitations are better described in terms of atomic properties (such as local magnetic moments). A classic example are paramagnetic Mott insulators, which exhibit Curie-Weiss susceptibilities rather than Pauli paramagnetism. Other examples are mixed-valence and Kondo physics materials and unconventional superconductors. In these cases, a summation to all orders in PT is mandatory to obtain meaningful results.

A breakthrough in this area was the realization that this summation can be formulated and

[1]Computational Sciences and Engineering Division and Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [2]Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY 11973, USA. [3]Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.
*Corresponding author. Email: kotliar@physics.rutgers.edu

carried out nonperturbatively in terms of the solution of a self-consistent quantum impurity model, provided that one restricts the self-energy to be local. This defines the dynamical mean-field theory (DMFT) (*16*) approach. Originally formulated for simple model Hamiltonians where the summation becomes exact in the limit of large lattice coordination (*17*), DMFT was later incorporated into electronic structure methods, starting with the LDA+DMFT approach (*18*, *19*).

The LDA+DMFT approach introduces the notion of locality for the self-energy Σ by defining a basis of local orbitals $\chi_{\beta\vec{R}'}(\vec{r}')$, which are centered at sites defined by $\vec{R}'$. One then divides all orbitals into two sets. For the uncorrelated orbitals, the self-energy is approximated by the exchange correlation potential of DFT, whereas the self-energy of the correlated orbitals is evaluated by an impurity model [as $\Sigma_{imp}(\omega)$] in a self-consistent medium, as illustrated in Fig. 3,

$$\Sigma(\vec{r},\vec{r}',\omega) = \delta(r-r')V_{xc}(r) + \sum_{\alpha\vec{R},\beta} \chi^{*}_{\alpha\vec{R}}(\vec{r})(\Sigma_{imp}(\omega) - E_{dc})_{\alpha\vec{R},\beta\vec{R}}\chi_{\beta\vec{R}}(\vec{r}') \quad (2)$$

where χ^* indicates the complex correlate of χ, α and β correspond to orbital indexes, and δ is the delta function. E_{dc} accounts for the contribution to the self-energy that is included in the exchange correlation potential V_{xc} (*20*). Single-site DMFT is just one example of a quantum embedding scheme. The central idea of this approach is a judicious use of two steps, truncation and embedding. In the truncation scheme, one selects some orbitals to be treated at the DMFT level, whereas the rest of the system is replaced by a bath that enables fluctuations among different configurations. In the language of quantum chemistry, rather than focusing on configurations with a fixed number of electrons, the quantum embedding allows fluctuations among different configurations. This was a major step forward relative to more drastic truncations that describe a solid as a finite isolated cluster of atoms.

In DMFT one uses the system and bath to compute a local self-energy. For this purpose there have been notable advances in the capabilities for solving quantum impurity models, such as the introduction of continuous-time QMC methods (*21*, *22*). Further algorithmic advances will be needed to improve the efficiency of the sampling, to reach lower temperatures, and to treat a larger number of correlated orbitals. Another difficulty is the minus sign problem, which appears in the solution of the impurity model with off-diagonal hybridization, which occurs in sites with low crystal symmetry. There is a need for impurity solvers working on the real axis to avoid carrying analytic continuation of numerical data. Tailoring of Wilson renormalization group methods to the solution of self-consistent quantum impurity models has shown promising results (*23*), but major extensions are needed for tackling electronic structure problems. After the truncation step, the self-energy is embedded in a lattice Green function or the two-particle irreducible vertex in a two-particle response to obtain momentum- and frequency-dependent physical quantities (Fig. 3).

Other embedding methods exist that treat solid-state systems. The density matrix embedding theory (DMET), introduced by Knizia and Chan (*24*), was inspired by DMFT, but it gives rise to a simpler impurity model. The Gutzwiller approximation used in connection with LDA, called LDA+G (*25*, *26*), is also computationally less demanding than LDA+DMFT. The slave particle method is widely used for the study of model Hamiltonians and has been combined with LDA (*27*). The rotationally invariant slave boson (RISB) form (*28*, *29*) has been shown to be quite accurate for the estimation of total energies. Recently, simple connections among all these approaches have been uncovered (*30*, *31*), with LDA+G being equivalent to LDA+RISB and DMET appearing as a limiting case of RISB. For these simplified methods to reach first-principles status, a prescription for the determination of the Hubbard *U* tensor to be used is needed.

"Green's functions and Monte Carlo methods are largely complementary."

Single-site DMFT can be extended by taking as a reference system a cluster of sites, rather than a single site (moving to the right on the locality axis of Fig. 2). There are different cluster extensions of single-site DMFT (*5*, *32*) and many subtleties: for example, how to maintain causality (*33*). In practice, different materials require different types of approximations; hence, many-body theory of correlated systems requires some judicious choices, guided by physics and chemistry, deciding of the level of approximation for a given material in Fig. 2. For example, a material such as VO_2, which contains dimers in the unit cell, requires two-site DMFT (*34*), whereas in other systems such as V_2O_3, single-site DMFT is sufficient (*35*).

Representative applications: Theoretical spectroscopy

The goal of theoretical spectroscopy is to infer different correlation functions—for example, the one-electron correlation described in Eq. 1 or the correlation function between the electronic spins—with minimum or no input parameters. To extract this information experimentally, different large-scale facilities with different experimental tools, such as angle-resolved photoemission spectroscopy or neutron scattering, respectively, are needed. Theoretical spectroscopy also provides a framework to interpret the experimental results and advances the field by falsifying or verifying theories and indicating which aspects of the modeling need further attention. LDA+DMFT is now widely used in this area, with surprisingly good agreement with experimental data; for example, it has been used to explore the variation of the spin fluctuation spectra of the iron pnictides and chalcogenides (*36*, *37*) and the heavy fermion systems (*38*).

We provide a more detailed example of the power of this method in suggesting the solution of the mystery of Pu, a distinctive elemental solid. Pu metal is different from all other elements in the periodic table because it forms so many vastly different allotropes in its phase diagram in a narrow temperature range. Even its simplest face-centered cubic (fcc) phase has remarkable transport and thermodynamical properties that have for many years defied qualitative understanding.

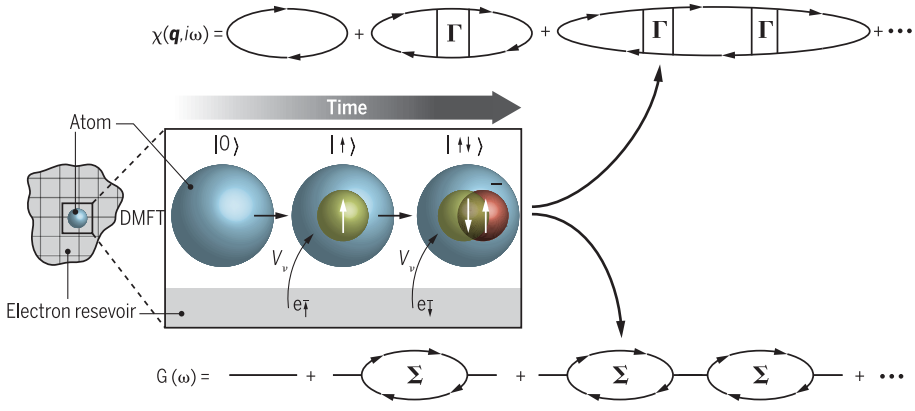


Fig. 3. DMFT as a quantum embedding method. A quantum truncation of a periodic system is an impurity model, an atom or a cluster of atoms in a bath (reservoir). From the impurity model, one calculates irreducible quantities such as the self-energy Σ or the particle hole irreducible vertex Γ, which are then embedded in lattice correlation functions. These are used to calculate frequency and momentum observables, such as the spin susceptibility χ or the one-particle Green function *G*. In realistic LDA+DMFT treatments of materials, a small subspace is correlated so that Σ is a small block, and the correlated self-energy is embedded in a large Hilbert space; thus, the lines in the equation for *G* represent the Kohn-Sham Green function from the DFT. The lines in the equation for χ represent the full Green function *G*. *q*, momentum; V_v, hybridization between the impurity and the bath of electrons.

IMAGE: ADAPTED BY N. DESAI

LDA+DMFT provided a comprehensive theory of this material. The anomalies in its properties can be traced to a mixed-valence condition, where the shell of electrons fluctuates between an f^5 and f^6 configuration with a characteristic mixed-valent scale that had been predicted to be about 80 meV (*39*).

The DMFT valence histogram describing the quantum-mechanical probability for a given f configuration is shown in Fig. 4. LDA+DMFT angle-integrated photoemission spectroscopy identified fingerprints of this mixed-valent state (Fig. 4, A and B). LDA+DMFT-based neutron theoretical spectroscopy was used to predict the powder average neutron scattering spectra of δ-Pu, as shown together with the experimental results in Fig. 4, D and E.

Wave function methods

To estimate the complete ground-state wave function in a solid, with its characteristically large number of electrons and its exponential scaling in the number of orbitals, seems at first sight a task that is impossible to carry out to useful accuracy. Nevertheless, stochastic algorithms offer a more scalable alternative to deterministic methods, and a quarter century of materials applications underscores their practicality. Stochastic algorithms were initially used for the homogeneous electron gas and then for weakly correlated solids and molecules (*40*), but algorithmic breakthroughs in real-space QMC (*41*–*43*) and the more recent development of related methods such as auxiliary-field QMC (AFQMC) (*44*) and full-configuration interaction QMC (FCIQMC) (*45*) have extended the reach of such methods from light elements to include solids with d- and f-electron atoms (*46*). These methods are complementary: the older and more established real-space methods scale with the second to fourth power of system size, depending on the measured quantity. AFQMC has similar scaling and gives simpler access to properties other than the energy but is generally believed to have a much larger cost prefactor. FCIQMC has exponential scaling, leaving a far smaller range of systems for which the method is applicable, but it is already near-exact in its current formulations. Although materials applications today may require the use of millions of supercomputer hours, these QMC methods all have the potential to treat a wide range of strongly correlated materials and potentially demonstrate systematic convergence of all errors. Although methodological improvements are needed, increased computer power is steadily increasing their range of application.

The central idea of QMC methods is to avoid explicit numerical integration of the many-body Schrödinger equation and to recast the problem statistically via use of Monte Carlo integration and stochastic projection. This allows the methods to operate on thousands of electrons, compared with the tens or low hundreds reached by today's most accurate deterministic algorithms, in exchange for a statistical uncertainty. Importance sampling via an input trial wave function can be

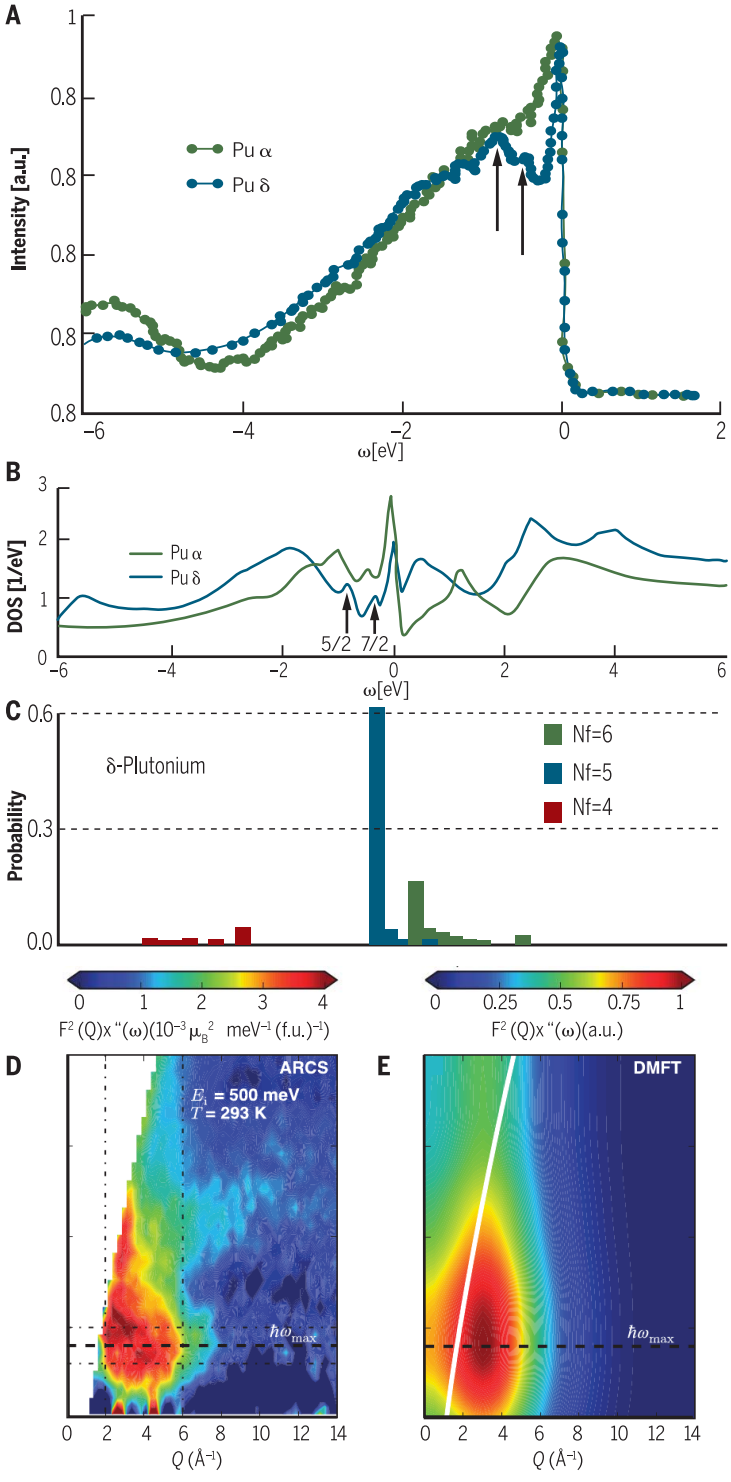


Fig. 4. Using LDA+DMFT to understand plutonium. (**A**) Experimental (*95*) and (**B**) theoretical (*96*) angle-integrated photoemission spectroscopy of δ-Pu. (A) was adapted with permission from (*95*), copyright (2000) by the American Physical Society. (B) was reproduced from (*96*). The arrows highlight two surprising features in addition to the narrow resonance and the broad Hubbard band, assigned to quasiparticle multiplets. For comparison, the experimental α phase is shown together with a compressed fcc phase. a.u., arbitrary units; DOS, density of states. (**C**) DMFT valence histogram of δ-Pu, reproduced from (*96*). The histogram displays the occupancies of two distinct valences, f^5 (blue bars) and f^6 (green bars), which are key to understanding the anomalous properties of Pu. The bars with the same color correspond to distinct values of total spin *J* of the Pu atom [for more details, see (*96*)]. (**D**) Experimental and (**E**) theoretical neutron spectroscopy on a powder sample. Reproduced from (*97*). *F*, magnetic form factor; *Q*, momentum transfer; μ_B, Bohr magneton; f.u., formula units; ARCS, angular-range chopper spectrometry; *T*, temperature; $\hbar$, Planck's constant *h* divided by 2π.

IMAGE: ADAPTED BY N. DESAI

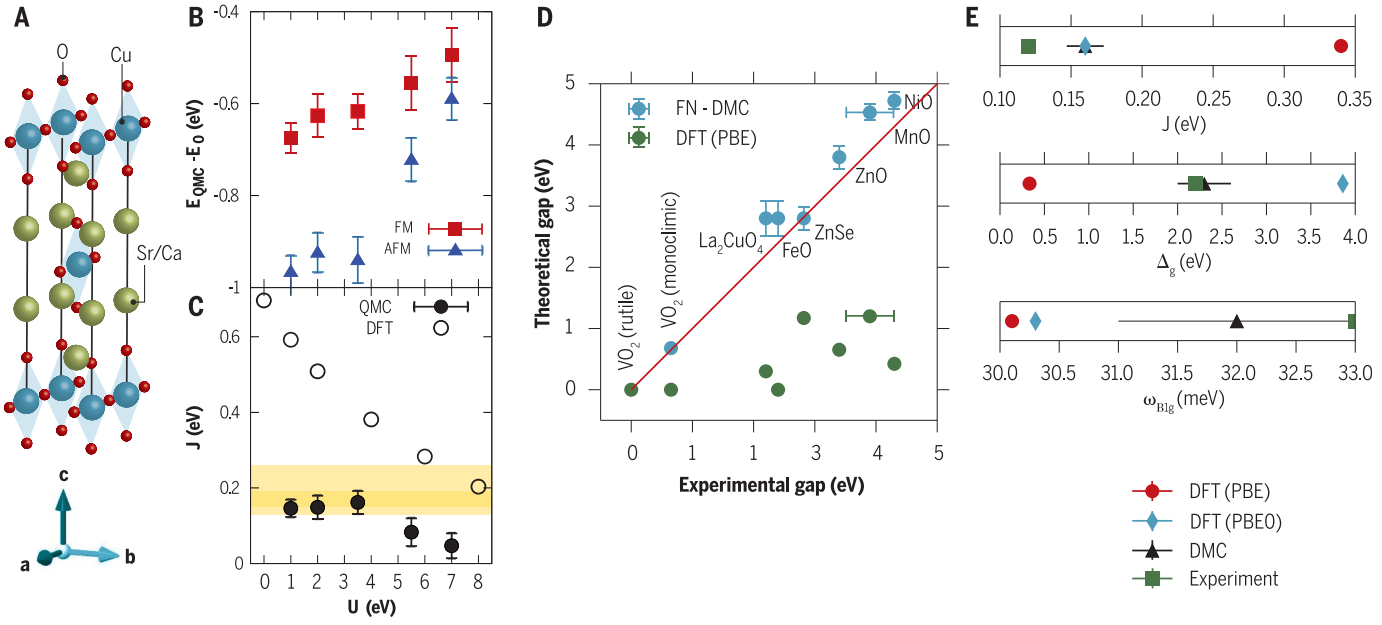


Fig. 5. Recent applications of QMC to strongly correlated materials. These applications were enabled by algorithmic advances and increased computational power. By using the atomic parameters and (**A**) the structure of the conventional unit cell of $(Sr/Ca)_2CuO_3$, (**B**) total energies (shifted by E_0 = 34,705 eV) for a ferromagnetic (FM) and an antiferromagnetic (AFM) state were evaluated to (**C**) extract the superexchange (J) by exploiting the variational principle and using trial wave functions from GGA+U. The best agreement with experimental data (shown in gold bands) is found at the lowest $E_{QMC} - E_0$ energies. Adapted from (*49*). (**D**) This method was also used in reference (*50*) to estimate phonon frequencies and gaps in isostructural La_2CuO_4; a correction of 0.5 eV was applied to the DMC values as the gap was evaluated at the Γ point. The agreement for gaps of correlated electron systems in fixed-node DMC (FN-DMC) was tested in several oxides, as also shown in (D). PBE, Perdew-Burke-Ernzerhof functional. (**E**) The comparison between calculated values within fixed-node DMC of the superexchange constant (J), gaps (Δ_g), and phonon frequencies (ω_{B1g}), with their values obtained by several other methods and experiment. PBE0, hybrid functional based on PBE. Reproduced with permission from (*41*), copyright (2016) by IOP Publishing.

used to improve the statistical efficiency, with higher efficiencies obtained by physically more accurate trials. The avoidance or minimization of numerical integration also allows for the use of more general wave functions, allowing more physics to be built in. As the trial wave function is improved and approaches the exact result, physical accuracy is increased and the number of Monte Carlo samples required for a given statistical accuracy is reduced, motivating developments of better trial wave functions.

The most fundamental challenge in QMC is the fermion sign problem: to sample an all-positive probability distribution derived from the wave function requires knowledge of where the wave function changes sign between positive and negative. To achieve this and determine the exact wave function formally has an exponential complexity. All the different QMC methods therefore aim to obtain a sufficiently accurate result by either introducing controlled approximations or reducing the prefactor of the exponential sufficiently that interesting materials and properties can be studied.

To guide the discussion, we will use a very general form of the many-body wave function in the solid as given by

$$\Psi\left(\{\vec{R}\}\right) = \sum_{i=1}^{M} c_i D_i^{\uparrow} D_i^{\downarrow} e^{J(\{\vec{R}\})} \qquad (3)$$

where $\{\vec{R}\}$ is the complete set of electron positions $\{\vec{r}\}$, $D_i^{\uparrow}$ and $D_i^{\downarrow}$ are Slater determinants filled with various single-particle orbitals, and c_i are expansion coefficients, multiplied by an optional Jastrow factor J that correlates electron positions. M is the number of electrons. The single-particle orbitals are usually obtained from DFT or quantum chemistry and the remaining components through numerical optimization. The trial wave function is therefore not exact and inherits some of the error made by the other methods.

The most accurate and widely applied QMC method for materials is diffusion Monte Carlo (DMC). Within this method, a reformulation of the time-dependent Schrödinger equation is used to project out the ground-state wave function Φ_0 from an input trial wave function Ψ_T. The use of a projection algorithm minimizes dependence on inputs, but with the complication of computing the mixed-probability density $|\Phi_0\Psi_T|$. The total energy is easily computed from this mixed distribution, but this method is more difficult to use to compute properties that do not commute with the Hamiltonian. To ensure that the solution has the correct symmetries of a fermionic state in DMC, the nodal surface of the solution is constrained to that of a trial wave function. This introduces a variational error in the computed energy, but this variational nature allows different options to be tested and the lowest-energy input to be rigorously selected. The challenge of applying DMC in correlated materials has therefore been to find a source of trial wave functions with sufficiently accurate nodal surfaces. Ideally these inputs would be obtained and optimized within QMC for a fully internally consistent theory, but this has not yet been realized in the solid state, although it has been achieved in molecules (*42*, *43*).

Starting in the mid-2000s, the QMC community began to realize that modern hybrid density functionals and simple variants of them could be a sufficiently accurate source of trial wave functions for transition metal oxides, with the variational principle used to select the best for each application. In these materials, the predictive power of DFTs is questionable. Accurate QMC results for materials such as MnO (*47*) and FeO (*48*) eventually spurred calculations for the challenging parent phases of the copper oxide–based superconductors (*49*–*51*). Here the calculations that are known to be most accurate through the variational principle were shown to predict magnetic exchange interactions and band gaps in best agreement with experimental data. Although inexact because of the limited flexibility in the input trial wave functions, these methods can now link fully atomistic structural models either to experimental data or to Hubbard model calculations, without empirical parameters. Similar techniques have been applied to the iron pnictide superconductors (*52*), the volume collapse in elemental cerium (*46*), and the phase transition in vanadium dioxide (*53*, *54*). Promisingly, these results have been obtained by using simple trial wave functions, leaving open the possibility of substantial improvements with more sophisticated and physically motivated choices.

Whereas the above-discussed techniques can be considered to work in real space, there are newer alternative approaches. FCIQMC (*45*) performs a stochastic sampling of the entire wave function on the basis of determinants. With near-exact but exponential scaling with increasing system size, current algorithms enable the study of simple solids (*55*), albeit with a very strong finite basis and finite size effects. AFQMC also makes use of a trial wave function, as does DMC, but uses a mathematically transformed Hamiltonian to enable sampling in terms of independent particles interacting with auxiliary fields. In materials, this sampling usually involves single-particle orbitals obtained from DFT. Compared with DMC, this method has the advantages of increased simplicity and allowing substantial reuse of the computational machinery from mean-field methods, as there is no Jastrow factor. This also allows easier access to materials properties. Although there are not yet head-to-head comparisons in correlated materials, the performance of AFQMC compared with that of other QMC methods is high in model systems (*56*). As with DMC, improving trial wave functions is key because they act as an initial constraint.

Although improved treatment of the fermion sign problem in all of these methods has long been recognized as an important challenge, they are also nearly exclusively formulated at zero temperature. Dynamics are little explored other than for the lightest elements of hydrogen and helium because of the high computational cost. Finite temperature properties of materials currently have to be accessed via other theories.

Representative applications: Total energies, magnetic exchange, and electron-phonon coupling

Accurate calculations of total energies, free energies, and enthalpies are essential for designing new materials. They enable evaluations of lattice constants, phonon energies, and formation energies. Evaluating them with methods that go beyond the more standard implementations of DFT has been the focus of multiple efforts. The accuracy and the relatively weak dependence on input parameters of QMC methods makes them an ideal tool for this task, in particular for light elements, where they have been used as a benchmark for different van der Waals density functionals (*57*–*59*).

QMC methods can also be used to infer information about excitations by arranging the excited-states wave function to be orthogonal to the ground-state wave function either by fixing different quantum numbers, such as spin, in the linear combination of Slater determinants in Eq. 3 or by enforcing a symmetry different from the ground state. Calculations for excitations must be performed individually, and in general complete spectra over a wide energy range are inaccessible. We illustrate recent successes of this method in correlated oxides in Fig. 5.

Though the evaluation of total energies and free energies has long been a strength of QMC methods, progress has been made in the use of Green's function methods for these analyses. Variants of the *GW* approach (*60*, *61*), LDA+G (*30*, *62*) and LDA +DMFT (*61*, *63*–*66*), have been used. The latter gives access to the effects of electronic entropies on the structural properties. This effect can be very important in strongly correlated materials. The variants of LDA+G, in turn, have provided zero-temperature phase diagrams of several actinides (*30*, *62*).

Another important quantity is the electron-phonon coupling, which can be related to the rate of change of electronic excitations with changes in the atomic positions. In the past decade, its evaluation in linear response theory implementations of the DFT has been carried out, the results of which have been successful in numerous compounds. For example, the dc resistivity as a function of temperature of the full elemental transition metal series has been explained without adjustable parameters (*67*).

Still, this approach sometimes fails in notable cases. One example is the high-temperature superconductor $Ba_xK_{(1-x)}BiO_3$, for which calculations gave a value of $\lambda \approx 0.3$, which cannot account for an observed superconducting transition temperature of 35 K (*68*). Recently, Yin *et al.* (*69*) have argued that the long-range Coulomb interactions are important in some high-temperature superconductors and that the induced static correlations increase the electronic bandwidth and enhance the electron-phonon coupling. This correlation-enhanced electron-phonon coupling provides a natural explanation for the superconductivity in this and related systems. Recent photoemission experiments have confirmed the main predictions of the theory (*70*).

"A hallmark of strongly correlated materials is the presence of partially filled d or f electronic shells that ...participate strongly in determining the material's... properties."

The iron pnictides and chalcogenides are Hund's metals, a different class of strongly correlated materials where the correlations derive from Hund's coupling rather than the Hubbard *U*. LDA+DMFT calculations for FeSe predicted a large enhancement (on the order of a factor of 10), relative to its LDA value, of the coupling between some Fe-3d correlated states and one phonon mode, which corresponds to the symmetric modulation of the distance of the chalcogenide atoms from the iron planes (*71*). This LDA+DMFT prediction has recently been confirmed experimentally by using a novel technique at the x-ray free-electron laser (*72*).

Outlook

Buoyed by advances in concepts, algorithms, and supercomputing, we can contemplate an accurate fully ab initio approach to compute the properties of correlated materials with only controlled approximations. If realized, such a scheme would enable highly robust predictions for strongly correlated quantum-mechanical phenomena. This requires accuracies that are out of reach today. The scheme could also lead to improvements in more practical faster methods (*73*), thus opening up correlated materials for materials design and optimization in the same way that weakly correlated materials are today. Some of the challenges, such as simply computing a sufficiently large number of atoms to capture the required physics, are greatly helped by the rapid advances in computational power, whereas other fundamental challenges require improved theories.

Green's functions and Monte Carlo methods are largely complementary. Green's function methods are ideally suited to describe spectroscopies, as they target exactly the quantities that are probed in experiments (Fig. 4), whereas Monte Carlo methods target total energies, as they have weak parametric dependence (Fig. 5).

The ranges of application of the different methods have recently begun to overlap, creating exciting opportunities. Combinations of methods advancing along the locality and PT axes of Fig. 2 were suggested very early on, as in the *GW*+DMFT method (*74*, *75*), but only recently have self-consistent implementations been carried out, as in the MQSGW (Matsubara quasiparticle self-consistent *GW*)+DMFT (*76*) method and the multitier self consistent *GW*+EDMFT (extended dynamical mean-field theory) (*77*).

Locality can be exploited in higher-order vertex functions and can be used to generate one-particle Green functions in methods such as dual fermions, $d\Gamma A$ (*78*), and trilex (*79*). Incorporating these ideas into realistic electronic structure is another direction to advance in the $y-z$ plane of Fig. 2 and is in the very early stages (*80*). The successrepresented in Fig. 1 in systematically improving the estimation of semiconductor gaps, together with the early suggestion of Zein *et al.* (*81*) that higher-order graphs are increasingly local, bodes well for this research direction. QMC methods have excellent access to the electron density, potentially allowing the inputs and outputs of DMFT calculations to be validated or improved. QMC methods now also have better access to excited states and are becoming capable of diagnosing when *GW* methods are best applied (*82*). Determining more accurately the effective Hubbard parameters that enter LDA+DMFT is another outstanding challenge to which QMC may contribute in the future (*83*).

Green's function methods that exploit locality are strongly dependent on the orbitals and projectors that are used to carry out the infinite-order resummations, as well as on the Hubbard interaction tensor. Investigations of criteria for optimizing these orbitals, with the concomitant determination of the local Hubbard interaction

matrix, are needed. Although QMC approaches are less sensitive on parametric choices, there is also now a common need, within all the QMC approaches, for improved trial wave functions to treat stronger correlations and to treat more complex systems. To be applied near coupled electronic, magnetic, or structural phase transitions—all hallmarks of strongly correlated materials—the methods must become independent of the mean-field theories and DFTs that are used as the foundation today. Current methods are expected to break down in these regimes, and cross-validation of new methods and new applications will be necessary (*56*). Encouragingly, the emergence of methods to generate multideterminant trial wave functions provides a practical mechanism to achieve this: selected configuration interaction can be used to produce systematically more accurate trial wave functions for DMC or AFQMC calculations; analysis of the results may be used for physical insight and for further tailoring of the configuration interaction expansions for larger systems, while maximizing accuracy. Although perturbative methods improve the trial wave functions markedly faster than a naïve configuration interaction expansion, physical insight to improve these methods specifically for the solid state will be highly beneficial. This will open up applications in elemental solids through to complex oxides, at ambient conditions through to conditions of high applied field or pressure.

We have not mentioned, among wave function methods, parallel progress within the field of quantum chemistry, where there are a range of complementary methods for treating electronic correlations in molecular systems [(*84*) and references therein]. Although traditionally implemented only for finite or isolated systems, these methods can be recast in the language of periodic boundary conditions needed to study materials in the solid state. This also applies to methods such as the coupled cluster, which have also recently been extended to periodic systems (*55*, *85*). These approaches stand to bring a half century of theoretical chemistry developments and analysis techniques to correlated materials (*86*). Conversely, ideas from DMFT are beginning to make their way into the realm of quantum chemistry (*87*–*90*). An important challenge for all of these methods is the use of frozen atomic core or pseudopotential approaches: Calculations actively using all electrons are too costly for general materials, whereas calculations that approximate the treatment of core electrons risk introducing significant errors.

The need and potential effects for society of scientific investment in computational first-principles approaches for correlated electron materials can hardly be overestimated. For weakly correlated electron systems, it took many years after the development of the DFT to reach the point where practical implementations of several algorithms using this method as an engine in user-friendly codes guided the discovery of the highest-temperature superconducting critical temperature of a material, H_3S under high pressure near 190 K (*91*, *92*). For a historical account of the research leading to this breakthrough, see reference (*93*). It is our hope that some of the developments reviewed in this article will help bring this new research paradigm of theory-assisted materials design to the realm of strongly correlated solids.

REFERENCES AND NOTES

1. M. Head-Gordon, E. Artacho, *Phys. Today* **61**, 58–63 (2008).
2. M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, K. A. Lyssenko, *Science* **355**, 49–52 (2017).
3. J. M. Luttinger, J. C. Ward, *Phys. Rev.* **118**, 1417–1427 (1960).
4. G. Baym, L. P. Kadanoff, *Phys. Rev.* **124**, 287–299 (1961).
5. G. Kotliar *et al*., *Rev. Mod. Phys.* **78**, 865–951 (2006).
6. V. I. Anisimov, O. Gunnarsson, *Phys. Rev. B* **43**, 7570–7574 (1991).
7. F. Aryasetiawan *et al*., *Phys. Rev. B* **70**, 195104 (2004).
8. A. Kutepov, K. Haule, S. Y. Savrasov, G. Kotliar, *Phys. Rev. B* **82**, 045105 (2010).
9. C. Honerkamp, *Phys. Rev. B* **85**, 195129 (2012).
10. Y. Nomura *et al*., *Phys. Rev. B* **86**, 085117 (2012).
11. L. Hedin, *Phys. Rev.* **139** (3A), A796–A823 (1965).
12. A. L. Kutepov, *Phys. Rev. B* **95**, 195120 (2017).
13. H. Jiang, P. Blaha, *Phys. Rev. B* **93**, 115203 (2016).
14. M. van Schilfgaarde, T. Kotani, S. Faleev, *Phys. Rev. Lett.* **96**, 226402 (2006).
15. A. Kutepov, K. Haule, S. Y. Savrasov, G. Kotliar, *Phys. Rev. B* **85**, 155129 (2012).
16. A. Georges, G. Kotliar, W. Krauth, M. J. Rozenberg, *Rev. Mod. Phys.* **68**, 13–125 (1996).
17. W. Metzner, D. Vollhardt, *Phys. Rev. Lett.* **62**, 324–327 (1989).
18. V. Anisimov, A. Poteryaev, M. Korotin, A. Anokhin, G. Kotliar, *J. Phys. Condens. Matter* **9**, 7359–7367 (1997).
19. A. I. Lichtenstein, M. I. Katsnelson, *Phys. Rev. B* **57**, 6884–6895 (1998).
20. K. Haule, *Phys. Rev. Lett.* **115**, 196403 (2015).
21. P. Werner, A. Comanac, L. de' Medici, M. Troyer, A. J. Millis, *Phys. Rev. Lett.* **97**, 076405 (2006).
22. E. Gull *et al*., *Rev. Mod. Phys.* **83**, 349–404 (2011).
23. K. M. Stadler, Z. P. Yin, J. von Delft, G. Kotliar, A. Weichselbaum, *Phys. Rev. Lett.* **115**, 136401 (2015).
24. G. Knizia, G. K.-L. Chan, *Phys. Rev. Lett.* **109**, 186404 (2012).
25. X. Deng, L. Wang, X. Dai, Z. Fang, *Phys. Rev. B* **79**, 075114 (2009).
26. K. M. Ho, J. Schmalian, C. Z. Wang, *Phys. Rev. B* **77**, 073101 (2008).
27. L. de' Medici, G. Giovannetti, M. Capone, *Phys. Rev. Lett.* **112**, 177001 (2014).
28. T. Li, P. Wölfle, P. J. Hirschfeld, *Phys. Rev. B* **40**, 6817–6821 (1989).
29. F. Lechermann, A. Georges, G. Kotliar, O. Parcollet, *Phys. Rev. B* **76**, 155102 (2007).
30. N. Lanatà, Y. Yao, C.-Z. Wang, K.-M. Ho, G. Kotliar, *Phys. Rev. X* **5**, 011008 (2015).
31. T. Ayral, T.-H. Lee, G. Kotliar, *Phys. Rev. B* **96**, 235139 (2017).
32. T. Maier, M. Jarrell, T. Pruschke, M. H. Hettler, *Rev. Mod. Phys.* **77**, 1027–1080 (2005).
33. G. Biroli, O. Parcollet, G. Kotliar, *Phys. Rev. B* **69**, 205108 (2004).
34. W. H. Brito, M. C. O. Aguiar, K. Haule, G. Kotliar, *Phys. Rev. Lett.* **117**, 056402 (2016).
35. I. Leonov, V. I. Anisimov, D. Vollhardt, *Phys. Rev. B* **91**, 195115 (2015).
36. Z. P. Yin, K. Haule, G. Kotliar, *Nat. Phys.* **10**, 845–850 (2014).
37. H. Park, K. Haule, G. Kotliar, *Phys. Rev. Lett.* **107**, 137007 (2011).
38. E. A. Goremychkin *et al*., *Science* **359**, 186–191 (2018).
39. C. A. Marianetti, K. Haule, G. Kotliar, M. J. Fluss, *Phys. Rev. Lett.* **101**, 056403 (2008).
40. R. M. Martin, L. Reining, D. M. Ceperley, *Interacting Electrons: Theory and Computational Approaches* (Cambridge University Press, ed. 1, 2016).
41. L. K. Wagner, D. M. Ceperley, *Rep. Prog. Phys.* **79**, 094501 (2016).
42. C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, R. G. Hennig, *Phys. Rev. Lett.* **98**, 110201 (2007).
43. M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory Comput.* **8**, 2181–2188 (2012).
44. S. Zhang, H. Krakauer, *Phys. Rev. Lett.* **90**, 136401 (2003).
45. G. H. Booth, A. J. W. Thom, A. Alavi, *J. Chem. Phys.* **131**, 054106 (2009).
46. N. Devaux, M. Casula, F. Decremps, S. Sorella, *Phys. Rev. B* **91**, 081101 (2015).
47. J. Kolorenč, L. Mitas, *Phys. Rev. B* **75**, 235118 (2007).
48. J. Kolorenč, L. Mitas, *Phys. Rev. Lett.* **101**, 185502 (2008).
49. K. Foyevtsova *et al*., *Phys. Rev. X* **4**, 031003 (2014).
50. L. K. Wagner, P. Abbamonte, *Phys. Rev. B* **90**, 125129 (2014).
51. L. K. Wagner, *Phys. Rev. B* **92**, 161116 (2015).
52. B. Busemeyer, M. Dagrada, S. Sorella, M. Casula, L. K. Wagner, *Phys. Rev. B* **94**, 035108 (2016).
53. H. Zheng, L. K. Wagner, *Phys. Rev. Lett.* **114**, 176401 (2015).
54. I. Kylänpää *et al*., *Phys. Rev. Mater.* **1**, 065408 (2017).
55. G. H. Booth, A. Grüneis, G. Kresse, A. Alavi, *Nature* **493**, 365–370 (2013).
56. M. Motta *et al*., *Phys. Rev. X* **7**, 031059 (2017).
57. R. C. Clay *et al*., *Phys. Rev. B* **89**, 184106 (2014).
58. P. Ganesh *et al*., *J. Chem. Theory Comput.* **10**, 5318–5323 (2014).
59. L. Shulenburger, A. D. Baczewski, Z. Zhu, J. Guan, D. Tománek, *Nano Lett.* **15**, 8170–8175 (2015).
60. B. Ramberger, T. Schäfer, G. Kresse, *Phys. Rev. Lett.* **118**, 106403 (2017).
61. A. Kutepov, S. Y. Savrasov, G. Kotliar, *Phys. Rev. B* **80**, 041103 (2009).
62. N. Lanatà *et al*., *Phys. Rev. Lett.* **111**, 196801 (2013).
63. S. Y. Savrasov, G. Kotliar, E. Abrahams, *Nature* **410**, 793–795 (2001).
64. K. Haule, T. Birol, *Phys. Rev. Lett.* **115**, 256402 (2015).
65. I. Leonov *et al*., *Phys. Rev. Lett.* **101**, 096405 (2008).
66. I. Leonov, V. I. Anisimov, D. Vollhardt, *Phys. Rev. Lett.* **112**, 146401 (2014).
67. S. Y. Savrasov, D. Y. Savrasov, *Phys. Rev. B* **54**, 16487–16501 (1996).
68. V. Meregalli, S. Y. Savrasov, *Phys. Rev. B* **57**, 14453–14469 (1998).
69. Z. P. Yin, A. Kutepov, G. Kotliar, *Phys. Rev. X* **3**, 021011 (2013).
70. C. H. P. Wen, H. C. Xu, Q. Yao, R. Peng, X. H. Niu, Q. Y. Chen, Z. T. Liu, D. W. Shen, Q. Song, X. Lou, Y. F. Fang, X. S. Liu, Y. H. Song, Y. J. Jiao, T. F. Duan, H. H. Wen, P. Dudin, G. Kotliar, Z. P. Yin, D. L. Feng, Unveiling the superconducting mechanism of $Ba_{0.51}K_{0.49}BiO_3$. arXiv:1802.10507 [cond-mat.supr-con] (28 February 2018).
71. S. Mandal, R. E. Cohen, K. Haule, *Phys. Rev. B* **89**, 220502 (2014).
72. S. Gerber *et al*., *Science* **357**, 71–75 (2017).
73. J. Kim *et al*., *J. Phys. Condens. Matter* **30**, 195901 (2018).
74. P. Sun, G. Kotliar, *Phys. Rev. B* **66**, 085120 (2002).
75. S. Biermann, F. Aryasetiawan, A. Georges, *Phys. Rev. Lett.* **90**, 086402 (2003).
76. S. Choi, A. Kutepov, K. Haule, M. van Schilfgaarde, G. Kotliar, *NPJ Quantum Mater.* **1**, 16001 (2016).
77. F. Nilsson, L. Boehnke, P. Werner, F. Aryasetiawan, *Phys. Rev. Materials* **1**, 043803 (2017).
78. G. Rohringer *et al*., *Rev. Mod. Phys.* **90**, 025003 (2018).
79. T. Ayral, O. Parcollet, *Phys. Rev. B* **93**, 235124 (2016).
80. A. Galler, P. Thunström, P. Gunacker, J. M. Tomczak, K. Held, *Phys. Rev. B* **95**, 115107 (2017).
81. N. E. Zein, S. Y. Savrasov, G. Kotliar, *Phys. Rev. Lett.* **96**, 226403 (2006).
82. L. Zhao, E. Neuscamman, A variational approach to optical band gaps. arXiv:1804.09663 [cond-mat.str-el] (25 April 2018).
83. H. J. Changlani, H. Zheng, L. K. Wagner, *J. Chem. Phys.* **143**, 102814 (2015).
84. A. Scemama, Y. Garniron, M. Caffarel, P.-F. Loos, *J. Chem. Theory Comput.* **14**, 1395–1402 (2018).
85. A. Grüneis, *Phys. Rev. Lett.* **115**, 066402 (2015).
86. A. Grüneis, S. Hirata, Y. Y. Ohnishi, S. Ten-No, *J. Chem. Phys.* **146**, 080901 (2017).
87. D. Jacob, K. Haule, G. Kotliar, *Phys. Rev. B* **82**, 195115 (2010).
88. N. Lin, C. A. Marianetti, A. J. Millis, D. R. Reichman, *Phys. Rev. Lett.* **106**, 096402 (2011).
89. D. Zgid, G. K.-L. Chan, *J. Chem. Phys.* **134**, 094115 (2011).
90. C. Weber *et al*., *Phys. Rev. Lett.* **110**, 106402 (2013).
91. D. Duan *et al*., *Sci. Rep.* **4**, 6968 (2014).
92. A. P. Drozdov, M. I. Eremets, I. A. Troyan, V. Ksenofontov, S. I. Shylin, *Nature* **525**, 73–76 (2015).
93. W. E. Pickett, Room temperature superconductivity revolution: Foreshadowed by Victorians, enabled by millennials. arXiv:1801.00165 [cond-mat.supr-con] (30 December 2017).
94. R. Adler, C.-J. Kang, C.-H. Yee, G. Kotliar, Correlated material design: Prospects and challenges. arXiv:1807.00398 [cond-mat.str-el] (1 July 2018).
95. A. J. Arko *et al*., *Phys. Rev. B* **62**, 1773–1779 (2000).
96. J. H. Shim, K. Haule, G. Kotliar, *Nature* **446**, 513–516 (2007).
97. M. Janoschek *et al*., *Sci. Adv.* **1**, e1500188 (2015).

ACKNOWLEDGMENTS

G.K. is grateful to his collaborators, K Haule, A. Kutepov, and Z. Yin, who made the work reviewed in this article possible. **Funding:** G.K. and P.R.C.K. were supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, as part of the Computational Materials Sciences Program. G.K.'s work on Pu was supported by DOE DOE BES DE-FG02-99ER45761. **Competing interests:** The authors declare that they have no competing financial interests.

10.1126/science.aat5975

REVIEW

Biophysical experiments and biomolecular simulations: A perfect match?

Sandro Bottaro and Kresten Lindorff-Larsen*

A fundamental challenge in biological research is achieving an atomic-level description and mechanistic understanding of the function of biomolecules. Techniques for biomolecular simulations have undergone substantial developments, and their accuracy and scope have expanded considerably. Progress has been made through an increasingly tight integration of experiments and simulations, with experiments being used to refine simulations and simulations used to interpret experiments. Here we review the underpinnings of this progress, including methods for more efficient conformational sampling, accuracy of the physical models used, and theoretical approaches to integrate experiments and simulations. These developments are enabling detailed studies of complex biomolecular assemblies.

In modern biological research, a key goal is to understand the functional consequences of structure, dynamics, and interactions of biological macromolecules. Proteins, lipids, carbohydrates, and nucleic acids interact, rearrange, and modify their shape while effecting their various functions. Experimentalists face the daunting task of characterizing thermodynamic and kinetic properties of macromolecules in a complex environment. Computational simulation plays a role in these efforts, as modeling approaches can aid in understanding experimental data and designing and predicting the outcome of future experiments.

Here we review the progress and current challenges in computational modeling of biomolecules, focusing on the topic of atomistic biomolecular simulations and the relationship between experiments and simulations. We highlight recent technological and theoretical advances in the field and consider whether there is a perfect match between experiments and simulations. Disagreement between computation and experiment provides useful insights to further our understanding, and their complementary use yields a clearer picture than either does alone.

Biomolecular simulations across length and time scales

Experimentalists often collect data that must then be synthesized into a coherent model through inverse problem-solving. Computational modelers deal with the forward problem: constructing a microscopic molecular model that can be compared with observed data (Fig. 1). In the best-case scenario, computational models are fully predictive and widely applicable. This would correspond to a model that correctly predicts the physical behavior of a system—for example, the catalytic power and stability of an enzyme as a function of pH and temperature. Such perfect models do not exist, so variants are developed with distinct strengths and areas of application.

Detailed information for reaction mechanisms and transition states in chemical reactions can be obtained via quantum mechanical (QM) calculations. These allow for simulation of the electronic properties of a subset of atoms within a macromolecule, which can be used to investigate bond cleavage and formation, distribution of charge and spin, and reaction mechanisms. Simulation of electronic properties of molecules requires a great deal of computational power, and thus the applicability of QM methods is, in general, limited to small systems or short time scales (*1*). Molecular dynamics (MD) simulations with empirical molecular mechanics force fields treat atoms as classical particles rather than considering their electronic structure: This approximation makes it possible to study the structure and dynamics of larger systems for longer periods of time, such as small proteins at the millisecond time scale. There are many relevant biological processes, however, that involve much larger biomolecular assemblies. The computational complexity of these problems can be decreased by grouping atoms together into single particles called beads. Such coarse-grained (CG) models range in resolution from one or a few beads per amino acid to one bead per hundreds or thousands of DNA bases. Despite their intrinsic approximations, such models are essential for tackling important problems in structural biology, including understanding complex formation between intrinsically disordered proteins (*2*) and rationalizing chromosome conformation–capture experimental data, thereby gaining insights into the internal chromosome organization (*3*). There also exist mixed, hybrid, and intermediate models that bridge together different resolutions. We focus below on the use of classical MD simulations to study biomolecules.

Challenges and opportunities for biomolecular simulations

The successful use of MD simulations hinges on solving two distinct, yet related, problems: the "sampling problem" and the "force-field problem" (Fig. 2). The sampling problem refers to our ability to sample the relevant biomolecular configurations and to determine their relative populations. Exhaustive sampling is difficult to achieve, because it is not possible to know in advance the required amount of sampling needed to calculate precise statistical averages. It is even difficult to assess whether a simulation is converged, because one may never know whether there are motions occurring on time scales beyond those sampled and robust and generally applicable tools to monitor convergence are needed (*4*, *5*). Thus, active areas of development are theories, algorithms, and technological improvements to increase the precision of the simulations.

The force-field problem refers to the construction of the energy function that describes the physical interactions between atoms. Improvements in force fields thus increase the accuracy of simulations by providing a more realistic description of the molecular interactions. Although progress on solving these two problems requires distinct approaches, they are tightly related. Only after taking into account all relevant conformations, that is, those that contribute to thermodynamic averages, is it meaningful to calculate average quantities and compare them to experiments. Hence, our ability to improve force fields is tightly connected to improvements in sampling.

Challenge 1: Improving physical models

The first fundamental challenge in biomolecular modeling is the construction of the physical model itself. Trade-off between computer power and spatial or temporal resolution requires a choice of model, ranging from all-atom representations to CG (*6*) and ultra-CG models in which multiple residues or nucleotides are represented by a single site (*7*). A long-standing goal of the field is to construct hybrid models that smoothly couple together different components at different resolutions (*8*).

In atomistic MD simulations, interactions between particles are modeled by "physics-based" terms that take into account chain connectivity, electrostatic interactions, London dispersion forces, and so on. Parameterized pairwise interaction terms are fitted against QM calculations and experimental data to generate a force field that describes interactions between individual particles. Parameters like the equilibrium distance between two covalently bonded atoms are known with high accuracy. Other parameters, such as partial charges, are difficult to establish, as they do not correspond to physical observables that can be directly probed through experiments.

Accurate force-field parameterization for proteins has benefited from benchmarking and direct optimization of MD simulations with experimental

Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science and Integrative Structural Biology at University of Copenhagen (ISBUC), Department of Biology, University of Copenhagen, Copenhagen, Denmark.
*Corresponding author. Email: lindorff@bio.ku.dk

nuclear magnetic resonance (NMR) data on partially structured peptides (*9*, *10*). Simulations of peptides that are 10 to 40 residues long are possible to converge, yet can capture cooperative phenomena, such as helix-coil transitions or formation of small hydrophobic cores, which are difficult to parameterize from smaller molecules. Solution NMR experiments can provide residue-level information and are sensitive to the relative energies of conformations that correspond to local minima and have sizable populations. By optimizing the backbone potential to match the experimentally measured helicity of a 15-residue peptide, as measured by NMR, a small change of about 1 kJ mol^{-1} was found to be sufficient to balance the secondary-structure populations (*10*). This small change in energy leads to a substantial improvement in accuracy, as the forces are accumulated over multiple residues and the populations scale to energies exponentially. Corrections to force fields obtained from examining short peptides are transferable among different structural classes of proteins and have improved models of folding processes for small globular and fast-folding proteins (*11*, *12*).

Unfolded states and intrinsically disordered proteins (IDPs) have long appeared to be more compact and structured when observed in MD simulations than when observed through experiments. This discrepancy suggests that important physical effects were not modeled correctly in the simulations. Proposed solutions include improving the description of water or of protein-water interactions (*13*). These modifications have improved the accuracy of simulations of IDPs (*14*), which play important roles in biology and disease.

A side effect of increasing protein-water interactions is, however, destabilization of folded proteins. In practice, one might thus have to choose between one family of force fields for simulations of folded proteins and a separate set for disordered systems, complicating studies of partially folded systems or of order-disorder transitions such as folding upon binding. To tackle this problem, it is necessary to consider simultaneously proteins that span from fully ordered to completely disordered and to test and optimize parameters at the same time on all of these systems. A comprehensive study of model systems with diverse properties has recently produced a force field capable of handling both fully folded proteins and IDPs (*15*).

In parallel to the development of models to study the structure and dynamics of proteins, there has been a growing interest in modeling nucleic acids, particularly RNA because of its catalytic and regulatory activities. Although important improvements have been made, state-of-the-art RNA force fields remain less accurate that those for polypeptides (*16*). Here, too, artifacts of MD simulations have been uncovered by direct comparison against solution NMR data on small model systems (*17*). Similar to the case of IDPs, promising results have been obtained by balancing water-RNA and RNA-RNA interactions (*18*).

Systematic benchmarking of force fields against experiments has revealed a comforting trend: Force fields are getting better (*12*, *15*, *19*). It is worthwhile to note that these improvements have been possible even without substantial changes to the underlying model or mathematical function used in the force fields. Thus, despite the inherent simplicity and lack of, for example, taking polarization into account, it has been possible to improve force fields dramatically. Indeed, it is surprising that it is possible to parameterize force fields that work well across many different proteins and problems (*20*), and eventually, progress will require models that are more complex.

Improvements of force fields generally rely on ab initio QM calculations. Machine-learning approaches, particularly neural networks, make it possible to train simple potentials with QM-level accuracy (*21*). Training is typically done on small molecules, and encouraging results have been obtained by transferring these potentials to the study of larger organic molecules (*22*). Force fields that explicitly include polarization effects are also likely to benefit from automated methods for integrated parameterization from experiments and QM calculations and from improvements in software and algorithms from sampling with these potentials. Here, Bayesian methods for optimizing force fields against experimental data and QM calculations are expected to play an even larger role, by enabling systematic balancing of different sources of information (*23*–*25*).

Challenge 2: Accessing long time scales

Atomistic biomolecular MD simulations are inherently costly, owing to the need to model forces

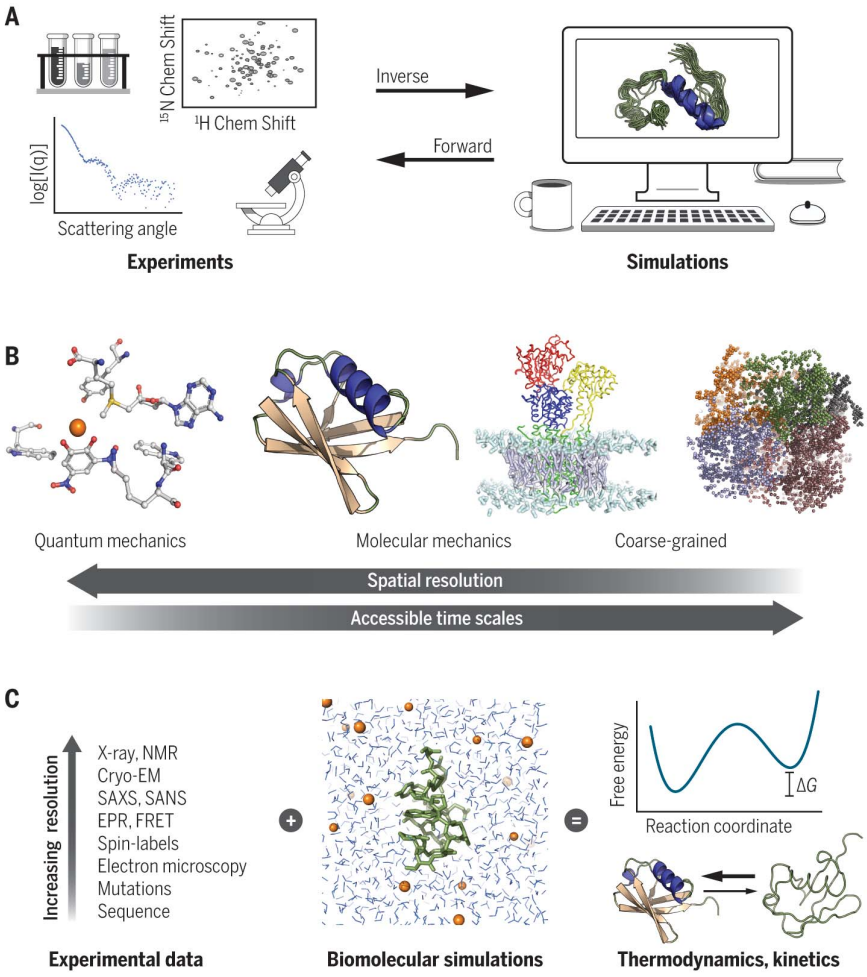


Fig. 1. Simulations and experiments are complementary. (**A**) Solving an inverse problem aims to describe causal factors that produce a set of observations. Molecular simulations, conversely, can be used to construct a set of microscopic molecular conformations that can be compared with experimental observations through the use of a forward model. (**B**) Computational approaches to studying biomolecules range from detailed quantum mechanical models to atomistic molecular mechanics to coarse-grained models, where several atoms are grouped together. The decreased computational complexity granted by progressive coarse-graining makes it possible to access longer time scales and greater length scales. (**C**) Experimental data can be combined with physical models to provide a thermodynamic and kinetic description of a system. As the model quality improves, it becomes possible to describe more complex phenomena with less experimental data. SANS, small-angle neutron scattering; EPR, electron paramagnetic resonance; FRET, fluorescence resonance energy transfer; Δ*G*, Gibb's free energy.

between tens or hundreds of thousands of individual atoms or more. These forces are evaluated every few femtoseconds of simulation time, requiring about a billion steps to simulate a molecule for a microsecond. Although the speed of a simulation depends strongly on the size of the biomolecular system and the available computational resources, it is not uncommon to require weeks or months of computer time with hundreds or thousands of processors working simultaneously to obtain microsecond-length simulations.

Conceptually, the most straightforward means to increase the speed and throughput of molecular simulations is perhaps "simply" to decrease the time it takes to perform a single iteration of the simulation. Widely used software packages designed for biomolecular simulations, such as GROMACS (*26*), NAMD (*27*), Desmond (*28*), AceMD (*29*), and AMBER (*30*), use different levels of parallelization by taking advantage of multicore processors and high-performance computing facilities. Speedups can be achieved by off-loading calculations to graphics processing units, which provide high performance at reasonable cost. A different route to improve efficiency is to build hardware specifically adapted to molecular simulations such as MDGRAPE (*31*) and Anton (*32*). For example, Anton is a massively parallel supercomputer designed to perform fast and accurate simulations of biomolecules by simultaneously considering all parts of the calculations, including MD-specific integrated circuits for calculating the costly parts of the force-field interactions, a specialized communication network tailored to match the periodic boundary conditions used in simulation, and special parallelization algorithms developed for this architecture. Anton enabled the first millisecond-length all-atom MD simulation of a globular protein (*32*). Its successor, Anton 2, is optimized for larger biomolecular systems and can perform multi-microsecond simulations in a single day for systems such as a small virus or a solvated ribosome with more than 1 million atoms (*33*).

Massive parallelization has been exploited in the folding@home project, which utilizes hundreds of thousands of "stand-by" machines all over the globe (*34*). Such distributed computing studies may now reach multiple milliseconds of aggregate simulation time and consist of hundreds or thousands of simulations ranging from hundreds of nanoseconds to a few microseconds (*35*). Because each simulation may be much shorter than the time scales of interest, a key problem is how to extract information about slow, long–time scale processes from a combined analysis of many short simulations. One possible solution to this problem is to build a Markov state model (MSM) (*35*, *36*), which enables one to construct a "memoryless transition network" describing the populations and kinetics of interconversion between different conformational states. In recent years, MSMs have gained widespread use, thanks to improved algorithms and software (*37*, *38*) and several successful applications to biomolecular processes, including protein folding, ligand binding, and protein-protein association (*35*). Path-based methods such as transition path sampling (*39*) and milestoning (*40*) also use many short simulations to study kinetics and mechanisms of long–time scale processes. These and related methods exploit the fact that many conformational transitions are "rare events," for which the time it takes to cross the barrier is substantially shorter than the waiting time between such events.

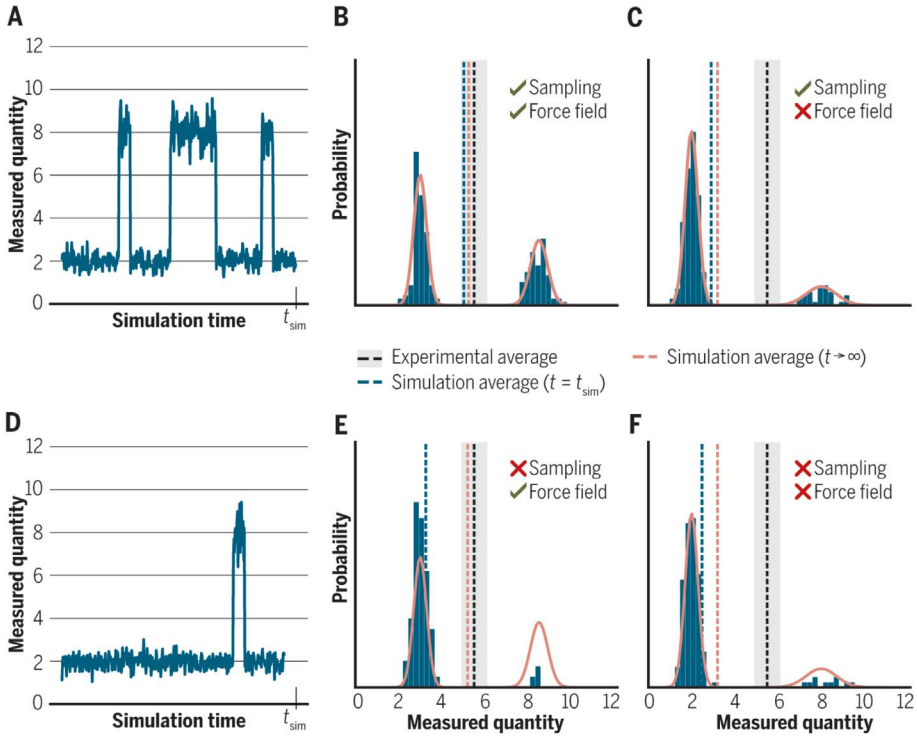


Fig. 2. Sampling and accuracy in molecular simulations. An MD simulation samples the temporal evolution of molecular configurations, but sampling is, in practice, limited to a finite time ($t = t_{sim}$). (**A**) When the simulation time is much longer than the slowest time scales of motions, many transitions are observed between the relevant conformational states. (**B**) When such simulations are performed with an accurate force field, statistical averages are converged and will be close to experimental values, and the averages approach the infinite time-scale average. (**C**) By contrast, when the force field is inaccurate, converged simulations give rise to precise, but inaccurate, results. (**D**) When the simulation time is too short compared to the time scales of the system, it is difficult to calculate precise quantities. (**E**) In this case, one may get disagreement between experiment and simulation even when the force field is accurate. (**F**) The worst situation, when sampling is insufficient and the force field is inaccurate.

Sampling may also be enhanced by changing simulation parameters. Increasing the temperature increases the kinetic energy, making barrier-crossing events faster (*41*). This idea is at the basis of parallel tempering, perhaps the most widely used enhanced sampling approach. These alterations can be viewed as enhancing simulations along a progress variable, also known as a reaction coordinate or collective variable (CV), in this case related to the energy of the system. For some problems, rapid fluctuations of the energy, and similar energies in different distinct conformational states, mean that increased temperature does not transfer into efficient sampling. This problem can be exacerbated by the fact that the available conformational space at high temperatures is larger, so that increased rates of sampling are more than offset by the increased volume of conformational space. Accelerated MD may instead be used to "boost" the energy along internal degrees of freedom, such as the backbone dihedral angles, thus enhancing the ability to cross local barriers (*42*).

Enhancing sampling along one or more prespecified CVs that describe the process of interest is another widely used strategy (*43*). In a protein-folding simulation, the number of native contacts formed or the progress along an initial guess of the folding path might be used to guide the simulation, even if the path is imperfect, and thus provide detailed insight into the folding free-energy landscape. Metadynamics (*44*) uses a time-dependent potential to simultaneously enhance sampling and construct a free-energy profile along such CVs and is widely used both because of its applicability to a range of problems (e.g., biomolecular processes, molecular docking, chemical reactions, crystal growth, and proton diffusion) and the availability of efficient and easy-to-use software (*45*).

Long unbiased simulations performed with Anton represent a useful reference to benchmark

CREDIT: ADAPTED BY V. ALTOUNIAN/*SCIENCE*

and validate enhanced sampling methods and kinetic models. In such applications, one may compare a specific protocol for enhanced sampling or constructing a kinetic model with the results from an unbiased simulation with the same force field to focus on benchmarking the algorithms and avoiding complications from force-field uncertainty (*46*, *47*).

Approaches based on CVs are very powerful, but their optimal choice is a critical and nontrivial step. For complex biomolecular rearrangements, it is difficult to identify CVs that correspond to the relevant, slowly varying degrees of freedom. In this respect, deep learning approaches have recently been used to identify improved CVs (*48*). CVs are not only useful to enhance sampling: They are essential to rationalize the large amount of complex data generated in MD simulations. New approaches to create better low-dimensional representations of high-dimensional data are also useful to construct improved MSMs (*35*), and we expect the advances in machine-learning methods [e.g., low-dimensional embedding and clustering (*49*, *50*)] to play an increasingly important role in this field.

Challenge 3: Integrating experiments and simulations

Although simulations and statistical mechanical theories are important and very powerful in their own right, direct integration of experimental data with molecular simulations can provide a rich description of the structure and dynamics of biomolecules. This field—also called integrative structural biology (*51*)—has benefited from recent technological advances in cryo–electron microscopy (cryo-EM) and is particularly important for studies of complex, dynamic systems for which multiple structural techniques provide complementary information. Formally, the problem consists of determining the three-dimensional structure or, more generally, an ensemble of molecular conformations and their associated weights, which are compatible with a set of experimental observations.

One strategy is to modify the simulation to match experimental data (Fig. 3). In this case, the force field is not considered a fixed, immutable model but instead a fitting function to be adjusted by experimental data. Indeed, this "pseudoenergy" approach underlies most structure-determination algorithms in which a physical energy function (often a simplified force field) is combined with an "experimental energy function" that measures the deviation between experiment and simulation (*52*). These integrative approaches enable accurate protein-structure determination when using chemical shifts (*53*) (Fig. 3A) or when using sparse, uncertain, and ambiguous experimental data (*54*). Similar approaches have been developed with the aim of providing molecular models of large molecular complexes constructed by using diverse sets of experimental data, including cross-linking, small-angle x-ray scattering (SAXS), and cryo-EM images (*55*, *56*). The availability of different sources of complementary experimental data is, in this context, important, as it allows one to cross-validate results and avoid overfitting.

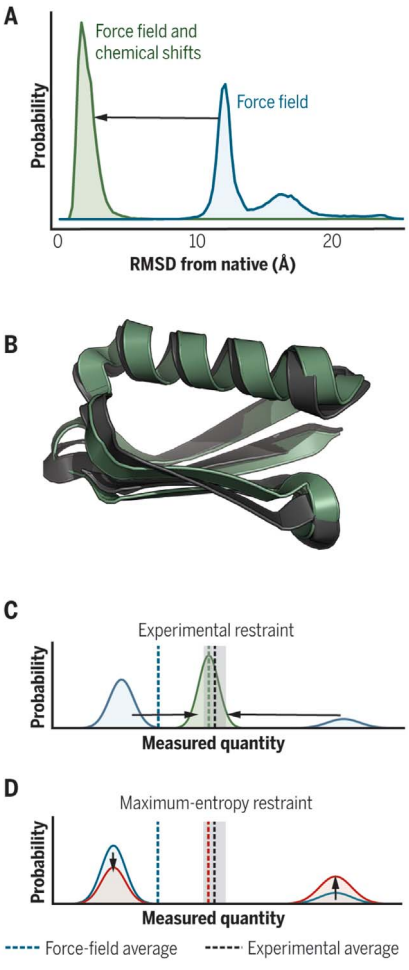


Fig. 3. Experimentally driven simulations. (**A**) Probability distribution of the structural similarity to the native structure of a protein determined by using a simplified force field (blue) or when the same force field is combined with NMR chemical-shift restraints (green) (*53*). RMSD, root mean square deviation. (**B**) A representative three-dimensional structure from the restrained simulation (green) and a reference structure (black). (**C**) In a conventional restrained simulation, the probability distribution of a measured quantity obtained by sampling using the force field alone (blue) is modified by adding an additional energy term that enforces the agreement with experimental data (green). In the resulting ensemble, all individual molecular conformations are close to the experimental average. (**D**) When heterogeneous conformations give rise to the measured average value (e.g., scalar couplings for different rotameric states), adding the experimental restraints to push individual conformation close to the experimental value is not correct, as this forces the simulation to structures that may not represent any of the relevant states. In maximum-entropy approaches, the experimental data are satisfied by introducing a minimal perturbation to the simulation ensemble. In this simplified example, the solution is a small shift in the populations of the two states, which results in a calculated average (red dashed line) compatible with the experiment.

Structural experiments such as SAXS, NMR, and x-ray diffraction report on quantities averaged over many molecules and long periods of time. For rigid molecules, the error may be small when interpreting ensemble-averaged quantities for individual structures. However, dynamical averaging is crucial when studying flexible molecules, such as IDPs or single-stranded RNA, because the structural interpretation of experimental data must be addressed by considering the coexistence of multiple conformational states (Fig. 3). One theoretical approach for dealing with the averaging problem is based on the maximum-entropy principle (*52*). The basic idea is to introduce a perturbation to the conformational ensemble generated by simulations in order to match a set of experimental data. The perturbation should be as small as possible: Mathematically, this is achieved by maximizing a quantity called relative Shannon entropy, hence the name maximum entropy. Thus, a minimal modification to the simulations to match the experimental data results in the least-biased combination of the force field and the experimental measurements. In practice, these approaches can remove much of the uncertainty associated with the choice of force fields so that conformational ensembles derived by combining experiments and simulations are more similar than ensembles derived solely from simulations (*5*, *57*).

Although the maximum-entropy principle provides a coherent framework to obtain conformational ensembles that combine force fields and experimental data, the basic formalism does not take sources of error into account. Another important development has thus been theory that considers not only experimental measurements but also the associated uncertainty. When combining data from multiple experimental techniques, uncertainties are essential to set the correct weights among them. For some sources of experimental data—for example, chemical shifts from NMR spectroscopy—the measurement itself is extremely precise, but our ability to relate the experimental quantity to molecular structure (i.e., the forward model that is used to calculate experimental quantities from three-dimensional structures) is associated with substantial uncertainty. Both experimental and modeling uncertainty can be treated by using Bayesian approaches such as those used in inferential structural determination protocols, leading to improved precision and a rigorous approach to integrate multiple sources of experimental data (*58*).

Combined Bayesian–maximum entropy integrative methods that consider uncertainty and averaging offer a promising route to reconstruct the conformational variability of complex biomolecular systems (*59*, *60*). These methods can be used with all-atom simulations or with CG models for larger assemblies. For instance, the structure and allosteric mechanism of a protein kinase were revealed by reweighting CG simulations using SAXS experimental data (*61*). An alternative

approach is to construct an MSM that has also been biased by using experimental data (*62*).

An important challenge is how the information gleaned from these studies may be fed back into improved force fields—for example, by systematically analyzing differences between the experimentally restrained ensembles and those obtained from the models alone. For instance, we recently identified a specific dihedral angle in the RNA backbone whose distribution in MD simulations was markedly different from that found when reweighting the same simulations using a Bayesian–maximum entropy approach (*63*). This observation suggested that force-field errors for this specific term could explain part of the disagreement between experiment and simulations, and, indeed, parallel work on improving RNA force fields resulted in distributions for this dihedral angle that were in much better agreement with the experimentally derived results (*18*, *63*).

The discussion above pertains to experimental data that can be related to equilibrium properties and that can be represented by population-weighted averages over individual conformations in the ensemble. For example, distances probed via nuclear Overhauser effect (NOE) NMR experiments are typically calculated from the average of the inverse sixth power of the distances in each individual structure (*58*). In reality, NOEs and many other experimental quantities depend on kinetic properties that need to be taken into account for the most accurate calculations. Recent theoretical and practical advances make it possible to construct conformational ensembles also based on such information (*62*, *64*–*66*) and thus extend applications to new sources of experimental data.

Conclusions and outlook

The complexity of biological systems often mandates the combined use of multiple techniques, including biomolecular simulations. Clearly, simulations are not ordinary experiments and often require a detailed knowledge of algorithms, underlying assumptions, and tricks that can be difficult to access and understand for nonspecialists. Much progress has been made on making these tools more user-friendly and accessible, though analyzing simulations often requires specialist knowledge. With a wide range of tools available, it is important to balance precision and accuracy when deciding on a simulation strategy (sampling method, force field, and level of resolution): What level of detail is relevant to the problem at hand, what are the relevant time scales, and can I address imperfections in the model by, for example, experimental restraints?

Substantial improvements in force fields have been made possible by using data from experimental studies on systems that are large enough to capture complex behavior yet simple enough to converge simulations. Future progress requires that experimentalist and computational chemists continue to work together to design experiments that are best suited to optimize force fields and to isolate properties that current models fail to describe (*9*, *17*). By testing and optimizing models broadly across different classes of problems and molecules, it will be possible to create force fields that are more transferable. Eventually, we will have to go beyond the current simple functional forms (*67*, *68*), but a surprising observation has been how much force fields could be improved by careful parameter optimization on an increasingly broad set of QM and experimental data. When reading the simulation literature, one should thus check whether a carefully validated force field has been used. Judging this is helped by the increased availability of systematic comparisons on a broad range of systems (*12*, *15*, *19*, *69*).

Further, as it remains difficult to sample conformational space sufficiently, particularly for complex systems, one should check whether convergence has been assessed and whether quantitative differences are backed up by sufficient sampling. This is inherently difficult because it is much easier to prove lack of convergence than the opposite (*70*). Nevertheless, useful questions to ask include (i) whether the same events are observed multiple times, (ii) the simulations are longer than the correlation times and the statistical analyses take time correlation into account, and (iii) whether the observed effects are greater than the statistical uncertainty.

We must, however, also be pragmatic in the way simulations are used. Like experiments, simulations are not perfect, and we will continue to live with uncertainty in sampling and force fields. Here the integration between experiment and simulations can help alleviate problems in both accuracy and sampling. We envision that these methods will play an increasingly important role in studying the relationship between structure and dynamics of large biomolecular assemblies or highly flexible molecules. The link between molecular simulations and cryo-EM, inherently a single-molecule technique, might be particularly fruitful for looking at conformational dynamics at high spatial resolution (*71*, *72*). Much can also be gained by carefully choosing systems that are amenable to both experimental and computational analysis. Recent examples include elucidating the molecular details that underlie the alternating access mechanism in a minimal sugar transporter (*73*) and an atomic-level description of interactions that lead to barrier roughness in protein folding (*74*).

The overwhelming growth of sequence data also presents new opportunities for computational chemists seeking to understand macromolecular structure and function. Evolution is, after all, governed by the same physical forces that simulations are constructed to model. One point of convergence has been the use of evolutionary records to construct statistical models of amino acid sequences (*75*, *76*). Conversely, computational biophysics can guide interpretations of what mutations do to proteins when analyzing exome sequencing for patient diagnosis (*76*–*78*). Finally, large-scale deep mutational scanning experiments can provide comprehensive maps of the mutational effects on protein stability across entire proteins (*79*) and enable a deeper understanding and benchmarking of our ability to predict the consequences of mutations (*76*, *80*).

We thus anticipate that simulations will eventually be commonplace when studying the effect of drugs and mutations and will play an essential role in the future of bioengineering in the same way that computer modeling is used today in computational prototyping of cars and buildings.

REFERENCES AND NOTES

1. R. E. Amaro, A. J. Mulholland, *Nat. Rev. Chem.* **2**, 0148 (2018).
2. A. Borgia *et al.*, *Nature* **555**, 61–66 (2018).
3. T. J. Stevens *et al.*, *Nature* **544**, 59–64 (2017).
4. J. D. Chodera, *J. Chem. Theory Comput.* **12**, 1799–1805 (2016).
5. M. Tiberti, E. Papaleo, T. Bengtsen, W. Boomsma, K. Lindorff-Larsen, *PLOS Comput. Biol.* **11**, e1004415 (2015).
6. S. J. Marrink, D. P. Tieleman, *Chem. Soc. Rev.* **42**, 6801–6822 (2013).
7. J. F. Dama *et al.*, *J. Chem. Theory Comput.* **9**, 2466–2480 (2013).
8. M. Praprotnik, L. D. Site, K. Kremer, *Annu. Rev. Phys. Chem.* **59**, 545–571 (2008).
9. J. Graf, P. H. Nguyen, G. Stock, H. Schwalbe, *J. Am. Chem. Soc.* **129**, 1179–1189 (2007).
10. R. B. Best, G. Hummer, *J. Phys. Chem. B* **113**, 9004–9015 (2009).
11. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, *Science* **334**, 517–520 (2011).
12. K. Lindorff-Larsen *et al.*, *PLOS ONE* **7**, e32131 (2012).
13. P. S. Nerenberg, T. Head-Gordon, *Curr. Opin. Struct. Biol.* **49**, 129–138 (2018).
14. J. Huang *et al.*, *Nat. Methods* **14**, 71–73 (2017).
15. P. Robustelli, S. Piana, D. E. Shaw, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4758–E4766 (2018).
16. J. Šponer *et al.*, *Chem. Rev.* **118**, 4177–4338 (2018).
17. J. D. Tubbs *et al.*, *Biochemistry* **52**, 996–1010 (2013).
18. D. Tan, S. Piana, R. M. Dirks, D. E. Shaw, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1346–E1355 (2018).
19. K. A. Beauchamp, Y.-S. Lin, R. Das, V. S. Pande, *J. Chem. Theory Comput.* **8**, 1409–1414 (2012).
20. J. C. Faver *et al.*, *PLOS ONE* **6**, e18868 (2011).
21. R. T. McGibbon *et al.*, *J. Chem. Phys.* **147**, 161725 (2017).
22. T. Bereau, R. A. DiStasio Jr., A. Tkatchenko, O. A. von Lilienfeld, *J. Chem. Phys.* **148**, 241706 (2018).
23. A. B. Norgaard, J. Ferkinghoff-Borg, K. Lindorff-Larsen, *Biophys. J.* **94**, 182–192 (2008).
24. L.-P. Wang, T. J. Martinez, V. S. Pande, *J. Phys. Chem. Lett.* **5**, 1885–1891 (2014).
25. J. Chen, J. Chen, G. Pinamonti, C. Clementi, *J. Chem. Theory Comput.* 10.1021/acs.jctc.8b00187 (2018).
26. M. J. Abraham *et al.*, *SoftwareX* **1–2**, 19–25 (2015).
27. J. C. Phillips *et al.*, *J. Comput. Chem.* **26**, 1781–1802 (2005).
28. K. J. Bowers *et al.*, in *Proceedings of the 2006 Association for Computing Machinery (ACM)/Institute of Electrical and Electronics Engineers (IEEE) Conference on Supercomputing* (ACM, New York, 2006); http://doi.acm.org/10.1145/1188455.1188544.
29. M. J. Harvey, G. Giupponi, G. D. Fabritiis, *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
30. D. A. Case *et al.*, *J. Comput. Chem.* **26**, 1668–1688 (2005).
31. I. Ohmura, G. Morimoto, Y. Ohno, A. Hasegawa, M. Taiji, *Phil. Trans. A Math. Phys. Eng. Sci.* **372**, 20130387 (2014).
32. D. E. Shaw *et al.*, *Science* **330**, 341–346 (2010).
33. D. E. Shaw *et al.*, in *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis* (ACM, New York, 2014), pp. 41–53.
34. M. Shirts, V. S. Pande, *Science* **290**, 1903–1904 (2000).
35. B. E. Husic, V. S. Pande, *J. Am. Chem. Soc.* **140**, 2386–2396 (2018).
36. J. D. Chodera, F. Noé, *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).
37. K. A. Beauchamp *et al.*, *J. Chem. Theory Comput.* **7**, 3412–3419 (2011).
38. M. K. Scherer *et al.*, *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).
39. J. Juraszek, J. Vreede, P. G. Bolhuis, *Chem. Phys.* **396**, 30–44 (2012).
40. R. Elber, A. West, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5001–5005 (2010).

41. Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **314**, 141–151 (1999).
42. D. Hamelberg, J. Mongan, J. A. McCammon, *J. Chem. Phys.* **120**, 11919–11929 (2004).
43. G. M. Torrie, J. P. Valleau, *J. Comput. Phys.* **23**, 187–199 (1977).
44. A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562–12566 (2002).
45. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, *Comput. Phys. Commun.* **185**, 604–613 (2014).
46. L. C. T. Pierce, R. Salomon-Ferrer, C. Augusto F. de Oliveira, J. A. McCammon, R. C. Walker, *J. Chem. Theory Comput.* **8**, 2997–3002 (2012).
47. Y. Wang, O. Valsson, P. Tiwary, M. Parrinello, K. Lindorff-Larsen, *J. Chem. Phys.* **149**, 072309 (2018).
48. J. M. L. Ribeiro, P. Bravo, Y. Wang, P. Tiwary, *J. Chem. Phys.* **149**, 072301 (2018).
49. C. Wehmeyer, F. Noé, *J. Chem. Phys.* **148**, 241703 (2018).
50. C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, V. S. Pande, *Phys. Rev. E* **97**, 062412 (2017).
51. H. van den Bedem, J. S. Fraser, *Nat. Methods* **12**, 307–318 (2015).
52. W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen, *PLOS Comput. Biol.* **10**, e1003406 (2014).
53. W. Boomsma *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 13852–13857 (2014).
54. A. Perez, J. A. Morrone, E. Brini, J. L. MacCallum, K. A. Dill, *Sci. Adv.* **2**, e1601274 (2016).
55. D. Russel *et al.*, *PLOS Biol.* **10**, e1001244 (2012).
56. E. Karaca, J. P. G. L. M. Rodrigues, A. Graziadei, A. M. J. J. Bonvin, T. Carlomagno, *Nat. Methods* **14**, 897–902 (2017).
57. T. Löhr, A. Jussupow, C. Camilloni, *J. Chem. Phys.* **146**, 165102 (2017).
58. W. Rieping, M. Habeck, M. Nilges, *Science* **309**, 303–306 (2005).
59. G. Hummer, J. Köfinger, *J. Chem. Phys.* **143**, 243150 (2015).
60. M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, *Sci. Adv.* **2**, e1501177 (2016).
61. T. A. Leonard, B. Różycki, L. F. Saidi, G. Hummer, J. H. Hurley, *Cell* **144**, 55–66 (2011).
62. S. Olsson, H. Wu, F. Paul, C. Clementi, F. Noé, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 8265–8270 (2017).
63. S. Bottaro, G. Bussi, S. D. Kennedy, D. H. Turner, K. Lindorff-Larsen, *Sci. Adv.* **4**, r8521 (2018).
64. N. Salvi, A. Abyzov, M. Blackledge, *J. Phys. Chem. Lett.* **7**, 2483–2489 (2016).
65. P. D. Dixit, K. A. Dill, *J. Chem. Theory Comput.* **14**, 1111–1119 (2018).
66. R. Capelli, G. Tiana, C. Camilloni, *J. Chem. Phys.* **148**, 184114 (2018).
67. Y. Shi *et al.*, *J. Chem. Theory Comput.* **9**, 4046–4063 (2013).
68. K. T. Debiec *et al.*, *J. Chem. Theory Comput.* **12**, 3926–3947 (2016).
69. C. Bergonzo, N. M. Henriksen, D. R. Roe, T. E. Cheatham 3rd, *RNA* **21**, 1578–1590 (2015).
70. A. Grossfield, D. M. Zuckerman, *Annu. Rep. Comput. Chem.* **5**, 23–48 (2009).
71. S. Hanot *et al.*, bioRxiv 113951 [Preprint]. 25 January 2018. https://doi.org/10.1101/113951
72. T. Nakane, D. Kimanius, E. Lindahl, S. H. Scheres, *eLife* **7**, e36861 (2018).
73. N. R. Latorraca *et al.*, *Cell* **169**, 96–107.e12 (2017).
74. H. S. Chung, S. Piana-Agostinetti, D. E. Shaw, W. A. Eaton, *Science* **349**, 1504–1510 (2015).
75. S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, *PLOS Comput. Biol.* **13**, e1005324 (2017).
76. T. A. Hopf *et al.*, *Nat. Biotechnol.* **35**, 128–135 (2017).
77. J. Shendure, J. M. Akey, *Science* **349**, 1478–1483 (2015).
78. S. V. Nielsen *et al.*, *PLOS Genet.* **13**, e1006739 (2017).
79. K. A. Matreyek *et al.*, *Nat. Genet.* **50**, 874–882 (2018).
80. V. Gapsys, S. Michielssens, D. Seeliger, B. L. de Groot, *Angew. Chem. Int. Ed. Engl.* **55**, 7364–7368 (2016).

ACKNOWLEDGMENTS

We thank Y. Wang for providing part of Fig. 1. Parts of Fig. 1A were designed by Freepik. **Funding:** We acknowledge funding from the Velux Foundations, the Lundbeck Foundation BRAINSTRUC initiative, and a Hallas-Møller stipend from the Novo Nordisk Foundation. **Competing interests:** None declared.

10.1126/science.aat4010

REVIEW

Inverse molecular design using machine learning: Generative models for matter engineering

Benjamin Sanchez-Lengeling[1] and Alán Aspuru-Guzik[2,3,4]*

The discovery of new materials can bring enormous societal and technological progress. In this context, exploring completely the large space of potential materials is computationally intractable. Here, we review methods for achieving inverse design, which aims to discover tailored materials from the starting point of a particular desired functionality. Recent advances from the rapidly growing field of artificial intelligence, mostly from the subfield of machine learning, have resulted in a fertile exchange of ideas, where approaches to inverse molecular design are being proposed and employed at a rapid pace. Among these, deep generative models have been applied to numerous classes of materials: rational design of prospective drugs, synthetic routes to organic compounds, and optimization of photovoltaics and redox flow batteries, as well as a variety of other solid-state materials.

Many of the challenges of the 21st century (*1*), from personalized health care to energy production and storage, share a common theme: materials are part of the solution (*2*). In some cases, the solutions to these challenges are fundamentally limited by the physics and chemistry of a material, such as the relationship of a materials bandgap to the thermodynamic limits for the generation of solar energy (*3*).

Several important materials discoveries arose by chance or through a process of trial and error. For example, vulcanized rubber was prepared in the 19th century from random mixtures of compounds, based on the observation that heating with additives such as sulfur improved the rubber's durability. At the molecular level, individual polymer chains cross-linked, forming bridges that enhanced the macroscopic mechanical properties (*4*). Other notable examples in this vein include Teflon, anesthesia, Vaseline, Perkin's mauve, and penicillin. Furthermore, these materials come from common chemical compounds found in nature. Potential drugs either were prepared by synthesis in a chemical laboratory or were isolated from plants, soil bacteria, or fungus. For example, up until 2014, 49% of small-molecule cancer drugs were natural products or their derivatives (*5*).

In the future, disruptive advances in the discovery of matter could instead come from unexplored regions of the set of all possible molecular and solid-state compounds, known as chemical space (*6*, *7*). One of the largest collections of molecules, the chemical space project (*8*), has mapped 166.4 billion molecules that contain at most 17 heavy atoms. For pharmacologically relevant small molecules, the number of structures is estimated to be on the order of 10^{60} (*9*). Adding consideration of the hierarchy of scale from subnanometer to microscopic and mesoscopic further complicates exploration of chemical space in its entirety (*10*). Therefore, any global strategy for covering this space might seem impossible.

Simulation offers one way of probing this space without experimentation. The physics and chemistry of these molecules are governed by quantum mechanics, which can be solved via the Schrödinger equation to arrive at their exact properties. In practice, approximations are used to lower computational time at the cost of accuracy.

Although theory enjoys enormous progress, now routinely modeling molecules, clusters, and perfect as well as defect-laden periodic solids, the size of chemical space is still overwhelming, and smart navigation is required. For this purpose, machine learning (ML), deep learning (DL), and artificial intelligence (AI) have a potential role to play because their computational strategies automatically improve through experience (*11*). In the context of materials, ML techniques are often used for property prediction, seeking to learn a function that maps a molecular material to the property of choice. Deep generative models are a special class of DL methods that seek to model the underlying probability distribution of both structure and property and relate them in a nonlinear way. By exploiting patterns in massive datasets, these models can distill average and salient features that characterize molecules (*12*, *13*).

Inverse design is a component of a more complex materials discovery process. The time scale for deployment of new technologies, from discovery in a laboratory to a commercial product, historically, is 15 to 20 years (*14*). The process (Fig. 1) conventionally involves the following steps: (i) generate a new or improved material concept and simulate its potential suitability; (ii) synthesize the material; (iii) incorporate the material into a device or system; and (iv) characterize and measure the desired properties. This cycle generates feedback to repeat, improve, and refine future cycles of discovery. Each step can take up to several years.

In the era of matter engineering, scientists seek to accelerate these cycles, reducing the

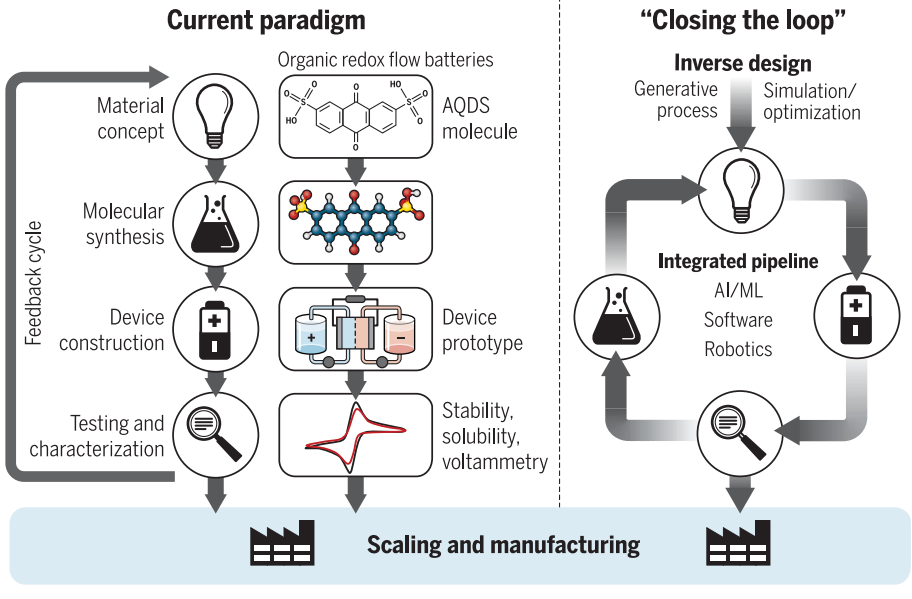


Fig. 1. Schematic comparison of material discovery paradigms. The current paradigm is outlined at left and exemplified in the center with organic redox flow batteries (*92*). A closed-loop paradigm is outlined at right. Closing the loop requires incorporating inverse design, smart software (*93*), AI/ML, embedded systems, and robotics (*87*) into an integrated ecosystem.

[1]Department of Chemistry and Chemical Biology, Harvard University 12 Oxford Street, Cambridge, MA 02138, USA. [2]Department of Chemistry and Department of Computer Science, University of Toronto, Toronto Ontario, M5S 3H6, Canada. [3]Vector Institute for Artificial Intelligence, Toronto, Ontario M5S 1M1, Canada. [4]Canadian Institute for Advanced Research (CIFAR) Senior Fellow Toronto, Ontario M5S 1M1, Canada.
***Corresponding author. Email: aspuru@utoronto.ca**

time between steps. The ultimate aim is to concurrently propose, create, and characterize new materials, with each component transmitting and receiving data simultaneously. This process is called "closing the loop," and inverse design is a critical facet (*12*, *15*).

Inverse design

Quantum chemical methods reveal properties of a molecular system only after specifying the essential parameters of the constituent atomic nuclei and their three-dimensional (3D) coordinate positions (*16*). Inverse design, as its name suggests, inverts this paradigm by starting with the desired functionality and searching for an ideal molecular structure. Here the input is the functionality and the output is the structure. Functionality need not necessarily map to one unique structure but to a distribution of probable structures. Inverse design (Fig. 2) uses optimization, sampling, and search methods to navigate the manifold of functionality of chemical space (*17*, *18*).

One of the earliest efforts in inverse design was the methodology of high-throughput virtual screening (HTVS). HTVS has its roots in the pharmaceutical industry for drug discovery, where simulation is an exploratory tool for screening a large number of molecules (*19*, *20*). HTVS starts with an initial library of molecules built on the basis of researchers' intuition, which narrows down the pool of possible candidate molecules to a tractable range of a thousand to a million. Initial candidates are filtered on the basis of focused targets such as ease of synthesis, solubility, toxicity, stability, activity, and selectivity. Molecules are also filtered by expert opinion, eventually considered as potential lead compounds for organic synthesis. Successful motifs and substructures are further incorporated in future cycles to further optimize functionality.

Although HTVS might seem like an ensemble version of the direct approach for material design, it differs in its underlying philosophy (*4*). HTVS is focused on data-driven discovery, which incorporates automation, time-critical performance, and computational funnels; promising candidates are further processed by more expensive methodologies. A crucial component is feedback between theory and experiment.

The HTVS methodology has been quite successful at generating new and high-performing materials in other domains. In organic photovoltaics, molecules have been screened on the basis of frontier orbital energies and photovoltaic conversion efficiency and orbital energies (*21*–*24*). In organic redox flow batteries, redox potential, solubility, and ease of synthesis (*25*, *26*) are prioritized. For organic light-emitting diodes, molecules have been screened for their singlet-triplet gap and photoluminescent emission (*27*). Massive explorations of reactions for catalysis (*28*) or redox potentials in biochemistry have been undertaken (*28*). For inorganic materials, the Materials Project (*29*) spawns many applications such as dielectric and optical materials (*30*), photoanode materials for generation of chemical fuels from sunlight (*31*), and battery electrolytes (*32*).

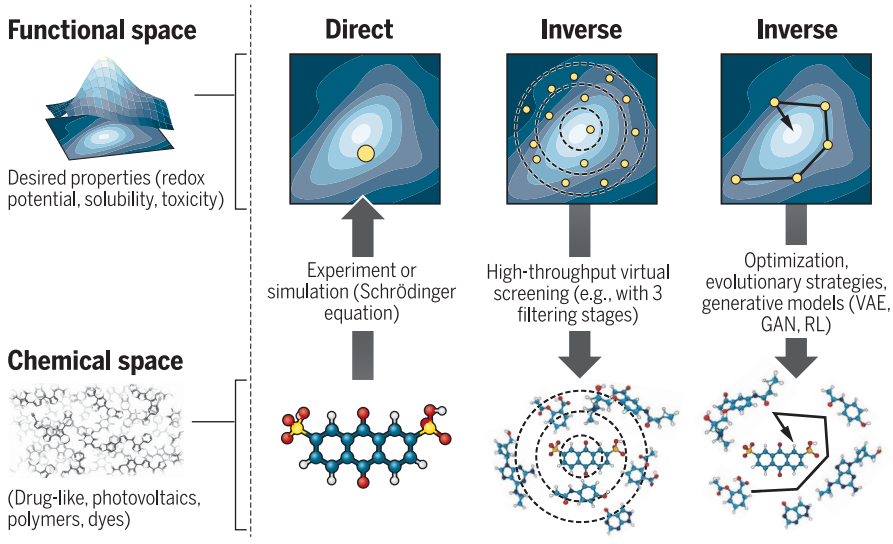


Fig. 2. Schematic of the different approaches toward molecular design. Inverse design starts from desired properties and ends in chemical space, unlike the direct approach that leads from chemical space to the properties.

Arguably, an optimization approach is preferable to HTVS because it generally visits a smaller number of configurations when exploring the manifold of functionality. An optimization incorporates and learns geometric information of the functionality manifold, guided by general trends, directions, and curvature (*17*).

Within discrete optimization methods, Evolution Strategies (ES) is a popular choice for global optimization (*33*–*35*) and has been used to map chemical space (*36*). ES involves a structured search that incorporates heuristics and procedures inspired by natural evolution (*37*). At each iteration, parameter vectors ("genotypes") in a population are perturbed ("mutated") and their objective function value ("fitness") evaluated. ES has been likened to hill-climbing in high-dimensional space, following the numerical finite difference across parameters that are more successful at optimizing the fitness. With appropriately designed genotypes and mutation operations, ES can be quite successful at hard optimization problems, even overcoming state-of-the-art machine learning approaches (*38*).

In other cases, inverse design is realized by incorporating expert knowledge into the optimization procedure, via improved Bayesian sampling with sequential Monte Carlo (*39*), invertible system Hamiltonians (*18*), deriving analytical gradients of properties with respect to a molecular system (*40*), optimizing potential energy surfaces of chemical systems (*41*), or discovering design patterns via data-mining techniques (*42*, *43*).

Finally, another approach involves generative models stemming from the field of machine learning. Before delving into the details, it is appropriate to highlight the differences between generative and discriminative models. A discriminative model tries to determine conditional probabilities ($p(y|x)$): that is, the probability of observing properties y (such as the bandgap energy or solvation energy), given x (a molecular representation). By contrast, a generative model attempts to determine a joint probability distribution $p(x, y)$: the probability of observing both the molecular representation and the physical property. By conditioning the probability on a molecule (x) or a property (y), we retrieve the notion of direct ($p(y|x)$) and inverse design ($p(x|y)$).

As expected, deep generative models are more challenging to create than direct ML approaches, but DL algorithms and computational strategies have advanced substantially in the last few years, producing astonishing results for generating natural-looking images (*44*), constructing high-quality audio waveforms containing speech (*45*), generating coherent and structured text (*46*), and most recently, designing molecules (*47*). There are several ways of building generative models, but for the purposes of this Review, we will focus on three main approaches: variational autoencoders (VAEs) (*48*), reinforcement learning (RL) (*49*), and generative adversarial networks (GANs) (*44*).

Before describing how each approach differs, we consider representations of molecules, which in turn determine the types of tools available and the types of information that can be exploited in the models.

Representation of molecules

To model molecular systems accurately, we must solve the Schrödinger equation (SE) for the molecular electronic Hamiltonian, from which we obtain properties relating to the energy, geometry, and curvature of the potential energy surface of our system. In the SE, the molecule is represented as a set of nuclear charges and the corresponding Cartesian coordinates of the atomic positions in 3D space. Meanwhile,

IMAGE: ADAPTED BY K. HOLOSKI

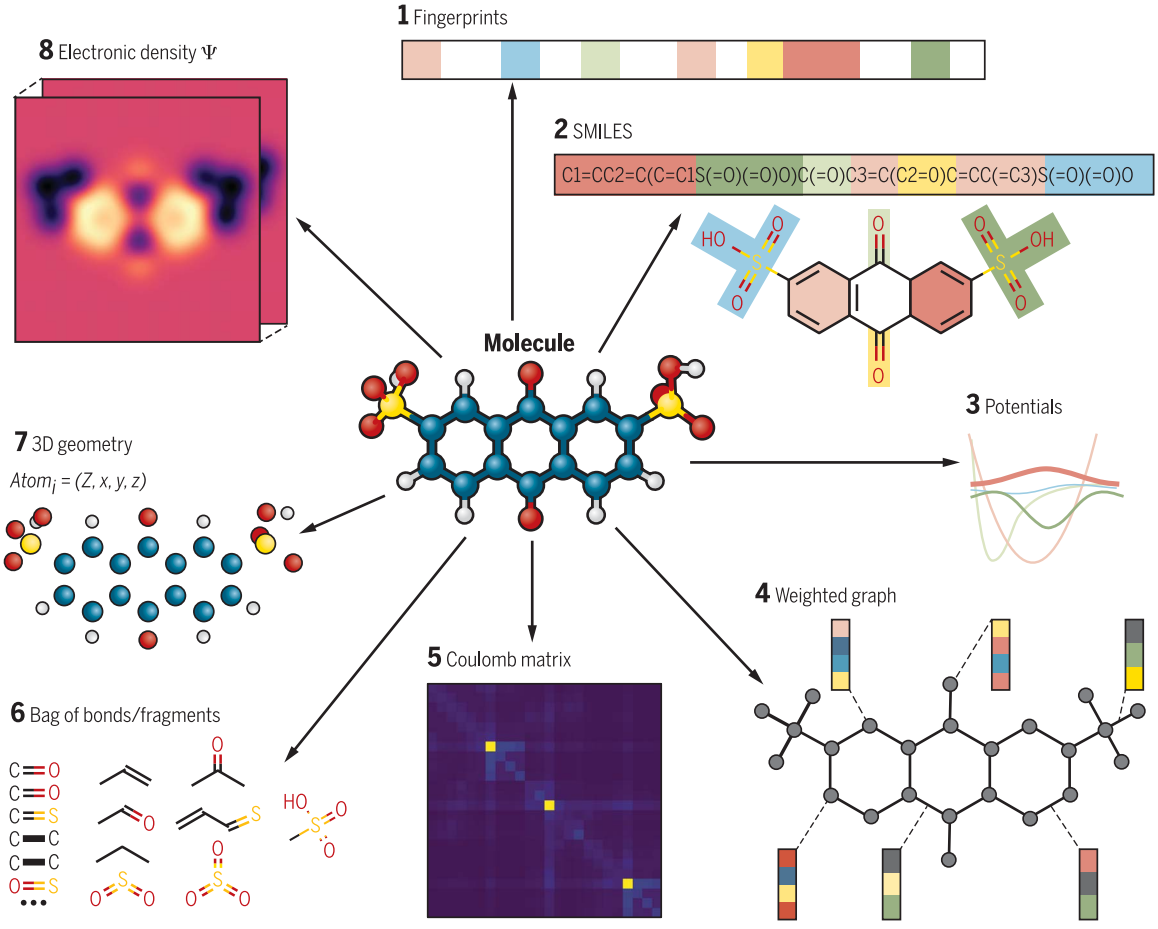


Fig. 3. Different types of molecular representations applied to one molecule, AQDS, which is used in the construction of organic redox flow batteries. Clockwise from top: (1) A fingerprint vector that quantifies presence or absence of molecular environments; (2) SMILES strings that use simplified text encodings to describe the structure of a chemical species; (3) potential energy functions that could model interactions or symmetries; (4) a graph with atom and bond weights; (5) Coulomb matrix; (6) bag of bonds and bag of fragments; (7) 3D geometry with associated atomic charges; and (8) the electronic density.

ML algorithms benefit from having representations that expose more easily constraints and properties of the physics of interest, so a 3D representation might not be the most efficient. Having a more direct representation allows the model to spend fewer computational resources learning patterns from first principles. A representation that can span all of chemical space should ideally capture all the symmetries of the SE: permutational, rotational, reflectional, and translational invariance for particles of the same type (*50*). Convolutions, Fourier transforms, and determinants are some of the mathematical structures that can preserve these symmetries and are often incorporated into the representation or model (*51*, *52*). Molecular representation is a current open research problem; there are many representations, and no one representation seems to work for all properties (*53*).

Current molecular representations fall into three broad categories: discrete (e.g., text), continuous (e.g., vectors and tensors), and weighted graphs. Although graphs can be represented as sparse matrices, they differ fundamentally in how they are processed within a model. Typically a representation will have a fixed length size via padding or the addition of dummy atoms. For inverse design, a desired property is invertibility—the capability to map back to a molecule structure that can then potentially be synthesized and characterized. Alternatively, if not invertible, it would be sufficient to have an ideal target representation and then either scan or evolve a molecule to match in a fast manner. Among invertible representations, we find molecular graphs and Hamiltonians.

Graphs are a natural representation of molecules. Following empirical principles of bonding, a molecule is interpreted as an undirected graph where each atom is a node and the bonds are the edges. To reduce complexity, hydrogen atoms are treated implicitly because they are deduced from standard chemistry valence rules. One standard for molecular graphs is SMILES strings (*54*), 1D text encodings that follow a particular grammar syntax. More advanced representations forgo the text encoding and use a weighted graph representation, with a variety of vectorized features on edges and nodes such as bonding type, aromaticity, charge, and distance. (*55*–*57*). Graphs are normally not uniquely represented, which can be advantageous for data augmentation (*58*) or disadvantageous when this representation degeneracy introduces noise to a model (*53*).

Whereas Hamiltonians rely only on the known physics and atomic constants of a molecule, the Coulomb matrix representation is based on Coulombic forces between charges of each atom (*59*). When combined via concatenation, summation, or differences, these base representations represent reactions (*60*), molecular ensembles, or conformers.

Other representations are better suited to prediction and could be rendered invertible via lookup tables: bag of bonds (*61*), amons (*62*), fingerprints (*63*, *64*), electronic density (*51*), symmetry functions (*65*), and chemical environments (*50*). Figure 3 shows several of these representations.

Generative models for exploring chemical space

Molecular representations are often inputs for deep neural network (DNN) models. The original data are transformed across several stages (layers),

IMAGE: ADAPTED BY K. HOLOSKI

usually by a linear transformation followed by a nonlinear function. For a given task and associated loss function, parameters for each layer (weights) are optimized via the backpropagation algorithm. When optimized, each intermediate (hidden) representation will tend to capture high- or low-level transformed features of the original data. In this sense, DL is a form of representation learning (*66*) because the DNN architecture is optimized to transform the original data into another representation that is more efficient for a given task such as regression, classification, or generation.

By attaching additional structure to the hidden representation, either in the form of statistical priors or probability distributions, we arrive at the idea of latent variable models. Each observed datum (molecule) has a corresponding latent representation, often a vector, within a latent variable space that encodes the relevant semantic features of the data.

The goal of a generative model is to model a data distribution, by training a model on large amounts of data and attempting to generate data like it. The loss function encodes the notion of likeness, measuring the differences between two distributions, the empirically observed and the generated.

We center most of our discussion on deep generative models using SMILES as a representation. Nonetheless, many of these approaches are quite general and are applicable to other representations. We expect future work to extend these architectures toward other molecular representations.

For the generation of sequences, recurrent neural networks (RNNs) (*46*, *67*) serve as a common starting point, creating sequences incrementally one step at a time and predicting what comes next. RNNs can be augmented to take into account complex time-dependent patterns with long short-term memory cells (*67*, *68*) (LSTMs), and attention and memory mechanisms (*69*). Figure 4 displays several architectures for generative models.

Variational autoencoders, reinforcement learning, and adversarial training

Besides generation tasks, for inverse design the generative process must be controlled or biased toward desirable qualities. With VAEs, the optimization of properties is performed explicitly over a continuous representation. By comparison, with GANs and RNNs, the optimization of properties can be achieved by biasing the generation process, typically with RL by rewarding or penalizing generative behaviors.

VAEs (*48*) give control over the data generation via latent variables. An autoencoder (AE) model includes an encoding and a decoding network. The encoder maps the molecule to a vector in a lower-dimensional space known as the latent space, and the decoder maps the latent vector back to the original representation. The encoder acts as a compression and the decoder as a decompression operation. The AE is trained to process and reproduce the original datum. In the act of distilling and expanding information, the AE is expected to learn some of the essential features of the data. The AE is sufficient to reproduce the training data, but it can easily learn to memorize the data. To be able to extrapolate and sample new molecules, we must fill the uncovered spaces of the latent space. The VAE achieves better generalizability by constraining the encoding network to generate latent vectors following a probability distribution on the latent space; often the distribution is Gaussian, owing to its accessible numerical and theoretical properties. Therefore, a molecule is represented not as a fixed point but as a probability distribution over latent space. In practice, this is done as a sampling procedure; when training, noise is added to the latent vector, so the VAE must reconstruct the same molecule from a noisy vector.

Arguably, the most interesting part of a VAE is the latent space. Molecules are represented as continuous and differentiable vectors residing on a probabilistic manifold. Latent space encodes a geometry; for a given molecule, we can sample nearby to decode similar molecules, and with increasing distance, we decode increasingly dissimilar molecules.

Given two molecules, we can trace a path between their corresponding latent coordinates, interpolate among the path, and decode interpolated molecules. Initially, VAEs were proposed for encoding characters of SMILES and then extended to take into account grammar and syntax features, which improve the generation of syntactically valid structures (*70*, *71*).

Latent space allows for direct gradient-based optimization of properties, as latent space is a continuous vector space. Nevertheless, the manifold of molecules has many local minima. One approach has been to explore a smoothed version of the manifold via Bayesian optimization (*71*) or constrained optimization with Gaussian processes (*47*).

By jointly training the VAE to reproduce molecules and properties, in a semi-supervised fashion, the latent molecular space reorganizes itself so that molecules with similar properties are close to each other. For a given property, there will exist preferred dimensions and directions. By changing the quality of their Gaussian processes, Gómez-Bombarelli *et al.* (*47*) demonstrated the capability of local or global optimization across the generated distribution.

Another way of building a generative model is with adversarial training under the GAN framework. Here, the generator competes against a discriminative model; specifically, the generator tries to generate synthetic data from sampling a noise space, whereas the discriminator tries to distinguish data as synthetic or real. Both models train in alternation, with the goal of the generator learning to structure noise toward producing data that the discriminator cannot classify better than chance. Convergence for GANs is not straightforward and can suffer from several issues, including mode collapse and overwhelming of the generator by the discriminator during training. Improving training for GANs is a current research topic (*72*), and dealing with discrete data, which suffers from nondifferentiability, has some workarounds (*73*, *74*).

To bias the generation process with GANs and RNNs, a gradient is needed to guide the optimization of a network toward desired properties. These properties could be modeled via neural networks and backpropagated to the generator, as is the case with GANs where the optimization metric is that the output looks like real data. To incorporate properties from chemoinformatic tools, simulations, or experimental measures, we need to create a gradient estimator that can backpropagate the generator.

The field of RL provides several approaches to this problem; among the most prominent are Q-learning (*49*) and policy gradients (*75*). RL considers the generator as an agent that must learn how to take actions (add characters) within an environment or task (SMILES generation) to maximize some notion of reward (properties). With SMILES, assigning rewards can only be done once the sequence is completed. To overcome this problem, Monte Carlo Tree Search (MCTS) is often used as it constructs a tree of probabilities and weights, simulating several possible completions for sequences, evaluating their reward, and weighting paths through the tree based on their success or failure at the given task. The completion behavior (policy) is learned as a neural network.

Because of these features, several molecular applications have adopted RL and MCTS for generation of drug-like molecules (*76*–*78*) and reaction synthesis planning (*79*, *80*).

The aforementioned approaches are not exclusive; they can be mixed to yield advantages from each. For instance, druGAN (*81*) adopts an adversarial autoencoder network, and ORGANIC (*82*, *83*) adopts both adversarial and RL approaches.

It should be noted that most results of generative models have been used in a pharmaceutical context, optimizing properties relevant to potential drugs such as solubility in water, melting temperature, synthesizability, and presence or absence of certain substructures. For example, Popova *et al.* (*78*) optimized molecules for putative inhibitors of Janus protein kinase 2 (JAK2), and Olivecrona *et al.* (*77*) optimized molecules active against the target dopamine receptor type 2.

Part of the focus on the SMILES representation has been driven by the adoption of natural language processing deep learning tools. It should be noted that SMILES represents only a subset of possible molecules; for example, a syntactically invalid SMILES string might still be a valid molecular structure, but its physics is not encoded by basic valence rules as used in SMILES. The introduction of more molecular representations and easy-to-use molecular property predictors will expand the use of generative models in other molecular contexts.

Looking ahead, new theoretical developments in ML are opening the door for generative models dealing with graphs. The VAE framework has been extended to molecular graphs (*84*), and message passing networks are used to incrementally build graphs (*57*). Even so, there many challenges remain; it is not yet clear how one

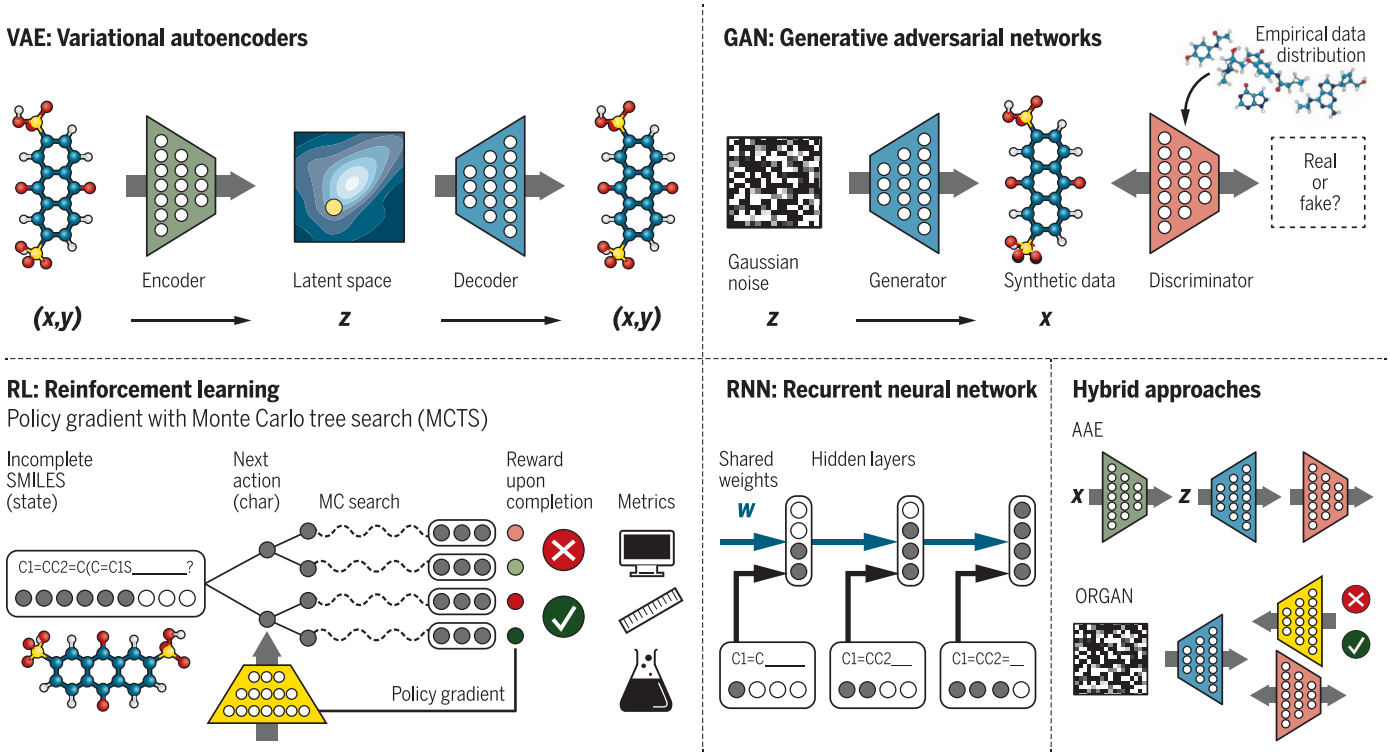


Fig. 4. Schematic representation of several architectures found in generative models. RNNs are used for sequence generation. The VAE shows the semi-supervised variant, jointly trained by molecules (x) and properties (y). Latent space is denoted with Z, and latent vectors with z. In the GAN setting, the noise eventually acquires structure via the adversarial training. Reinforcement learning (RL) shows a policy gradient with MTCS in the task of SMILES completion with arbitrary rewards. Shown in the lower right are hybrid architectures such as AAE (adversarial autoencoders) and ORGAN, which represents GAN and RL.

can deal practically with approximation methods for the graph isomorphism problem.

Additionally, improved sequence generation models are possible with the ability to read and write to memory (*69*). These approaches demonstrate better ability for learning long- and short-term patterns. More work is needed on Riemannian optimization methods that exploit the geometry of latent space. Structured architectures such as multilevel VAE (*85*) offer new ways of organizing latent space and are promising research directions. New approaches also lie in inverse RL, geared toward learning a reward or loss function (*86*). Research in this direction will allow for the discovery of reward functions associated with different materials discovery tasks.

Outlook

Inverse design is an important component of the complex framework required to design matter at an accelerated pace. The tools for inverse design, especially those stemming from the field of machine learning, have shown rapid progress in the last several years and have allowed chemical space to be framed into probabilistic data-driven models. Generative models produce large numbers of candidate molecules, and the physical realizations of these candidates will require automated high-throughput efforts to validate the generative approach. The community has yet has to show more than a few examples of successful closed-loop approaches for the design of matter (*87*). The blurring of the barriers between theory and experiment will lead to AI-enabled automated laboratories (*88*, *89*).

The combination of inverse design tools with active learning approaches such as Bayesian optimization (*90*, *91*) can enable a model that adapts as it explores chemical space, which allows for expanding a model in regions of high uncertainty and enabling the discovery of regions of molecular space with desirable properties as a function of composition. Active learning in the space of objective functions could lead to a better understanding of the best rewards to seek while carrying out machine learning.

As seen, central to machine learning methodologies is the representation of molecules; representations that encode the relevant physics will tend to generalize better. Despite considerable progress, much work remains. Graph and hierarchical representations of molecules are an area requiring further study.

The integration of machine learning as a new pillar of knowledge in the curricula of chemical, biochemical, medicinal, and materials sciences will allow for a more rapid adoption of the methodologies summarized in this work.

REFERENCES AND NOTES

1. Royal Geographical Society, *21st Century Challenges* (2015); https://21stcenturychallenges.org/challenges/.
2. D. Segal, *Materials for the 21st Century* (Oxford Univ. Press, 2017).
3. M. C. Scharber *et al.*, *Adv. Mater.* **18**, 789–794 (2006).
4. E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).
5. D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **79**, 629–661 (2016).
6. P. Kirkpatrick, C. Ellis, *Nature* **432**, 823–823 (2004).
7. A. Mullard, *Nature* **549**, 445–447 (2017).
8. J.-L. Reymond, *Acc. Chem. Res.* **48**, 722–730 (2015).
9. A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang, D. N. Beratan, *J. Am. Chem. Soc.* **135**, 7296–7303 (2013).
10. C. Qian, T. Siler, G. A. Ozin, *Small* **11**, 64–69 (2015).
11. M. I. Jordan, T. M. Mitchell, *Science* **349**, 255–260 (2015).
12. A. Aspuru-Guzik, R. Lindh, M. Reiher, *ACS Cent. Sci.* **4**, 144–152 (2018).
13. P. B. Jørgensen, M. N. Schmidt, O. Winther, *Mol. Inform.* **37**, 1700133 (2018).
14. E. Maine, E. Garnsey, *Res. Policy* **35**, 375–393 (2006).
15. A. Aspuru-Guzik, K. Persson, *Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence*. Mission Innovation (2018): Innovation Challenge 6.
16. T. Weymuth, M. Reiher, *Int. J. Quantum Chem.* **114**, 823–837 (2014).
17. A. Zunger, *Nat. Rev. Chem.* **2**, 0121 (2018).
18. C. Kuhn, D. Beratan, *J. Phys. Chem.* **100**, 10595–10599 (1996).
19. J. R. Broach, J. Thorner, *Nature* **384** (suppl.), 14–16 (1996).
20. S. Hoelder, P. A. Clarke, P. Workman, *Mol. Oncol.* **6**, 155–176 (2012).
21. D. Xiao, L. A. Martini, R. C. Snoeberger 3rd, R. H. Crabtree, V. S. Batista, *J. Am. Chem. Soc.* **133**, 9014–9022 (2011).
22. S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, A. Aspuru-Guzik, *Joule* **1**, 857–870 (2017).
23. I. Y. Kanal, S. G. Owens, J. S. Bechtel, G. R. Hutchison, *J. Phys. Chem. Lett.* **4**, 1613–1623 (2013).
24. J. Hachmann *et al.*, *Energy Environ. Sci.* **7**, 698–704 (2014).

IMAGE: ADAPTED BY K. HOLOSKI

25. L. Cheng *et al*., *J. Phys. Chem. Lett.* **6**, 283–291 (2015).
26. K. Lin *et al*., *Nat. Energy* **1**, 16102 (2016).
27. R. Gómez-Bombarelli *et al*., *Nat. Mater.* **15**, 1120–1127 (2016).
28. D. W. Robbins, J. F. Hartwig, *Science* **333**, 1423–1427 (2011).
29. A. Jain *et al*., *Comput. Mater. Sci.* **50**, 2295–2310 (2011).
30. I. Petousis *et al*., *Sci. Data* **4**, 160134 (2017).
31. Q. Yan *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3040–3043 (2017).
32. X. Qu *et al*., *Comput. Mater. Sci.* **103**, 56–67 (2015).
33. A. Supady, V. Blum, C. Baldauf, *J. Chem. Inf. Model.* **55**, 2338–2348 (2015).
34. I. Y. Kanal, G. R. Hutchison, Rapid computational optimization of molecular properties using genetic algorithms: Searching across millions of compounds for organic photovoltaic materials. arXiv:1707.02949 [physics.app-ph] (2017).
35. N. Yoshikawa, K. Terayama, T. Honma, K. Oono, K. Tsuda, Population-based de novo molecule generation, using grammatical evolution. arXiv:1804.02134 [physics.chem-ph] (2018).
36. C. Rupakheti, A. Virshup, W. Yang, D. N. Beratan, *J. Chem. Inf. Model.* **55**, 529–537 (2015).
37. G. Schneider, U. Fechner, *Nat. Rev. Drug Discov.* **4**, 649–663 (2005).
38. T. Salimans, J. Ho, X. Chen, S. Sidor, I. Sutskever, Evolution strategies as a scalable alternative to reinforcement learning. arXiv:1703.03864 [stat.ML] (2017).
39. H. Ikebata, K. Hongo, T. Isomura, R. Maezono, R. Yoshida, *J. Comput. Aided Mol. Des.* **31**, 379–391 (2017).
40. T. Weymuth, M. Reiher, Gradient-driven molecule construction: An inverse approach applied to the design of small-molecule fixating catalysts. arXiv:1401.1491 [physics.chem-ph] (2014).
41. M. Wang, X. Hu, D. N. Beratan, W. Yang, *J. Am. Chem. Soc.* **128**, 3228–3232 (2006).
42. P. Raccuglia *et al*., *Nature* **533**, 73–76 (2016).
43. B. Meredig, C. Wolverton, *Chem. Mater.* **26**, 1985–1991 (2014).
44. I. J. Goodfellow *et al*., Generative adversarial networks. arXiv:1406.2661 [stat.ML] (2014).
45. A. van den Oord *et al*., WaveNet: A generative model for raw audio. arXiv:1609.03499 [cs.SD] (2016).
46. S. R. Bowman *et al*., Generating sentences from a continuous space. arXiv:1511.06349 [cs.LG] (2015).
47. R. Gómez-Bombarelli *et al*., *ACS Cent. Sci.* **4**, 268–276 (2018).
48. D. P. Kingma, M. Welling, Auto-encoding variational Bayes. arXiv:1312.6114v10 [stat.ML] (2013).
49. V. Mnih *et al*., Playing Atari with deep reinforcement learning. arXiv:1312.5602 [cs.LG] (2013).
50. A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B Condens. Matter* **87**, 184115 (2013).
51. M. Hirn, S. Mallat, N. Poilvert, *Multiscale Model. Simul.* **15**, 827–863 (2017).
52. K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **8**, 13890 (2017).
53. O. A. von Lilienfeld, *Int. J. Quantum Chem.* **113**, 1676–1689 (2013).
54. D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).
55. S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).
56. Z. Wu *et al*., MoleculeNet: A benchmark for molecular machine learning. arXiv:1703.00564 [cs.LG] (2017).
57. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, Neural message passing for quantum chemistry. arXiv:1704.01212 [cs.LG] (2017).
58. E. J. Bjerrum, SMILES enumeration as data augmentation for neural network modeling of molecules. arXiv:1703.07076 [cs.LG] (2017).
59. M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301 (2012).
60. J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2**, 725–732 (2016).
61. K. Hansen *et al*., *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
62. B. Huang, O. A. von Lilienfeld, The "DNA" of chemistry: Scalable quantum machine learning with "amons." arXiv:1707.04146 [physics.chem-ph] (2017).
63. D. Duvenaud *et al*., Convolutional networks on graphs for learning molecular fingerprints. arXiv:1509.09292 [cs.LG] (2015).
64. D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **50**, 742–754 (2010).
65. J. Behler, *J. Chem. Phys.* **134**, 074106 (2011).
66. D. J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models. arXiv:1401.4082 [stat.ML] (2014).
67. A. Graves, Generating sequences with recurrent neural networks. arXiv:1308.0850 [cs.NE] (2013).
68. S. Hochreiter, J. Schmidhuber, *Neural Comput.* **9**, 1735–1780 (1997).
69. C. Olah, S. Carter, *Distill* **1**, (2016).
70. M. J. Kusner, B. Paige, J. M. Hernández-Lobato, Grammar variational autoencoder. arXiv:1703.01925 [stat.ML] (2017).
71. H. Dai, Y. Tian, B. Dai, S. Skiena, L. Song, Syntax-directed variational autoencoder for structured data. arXiv:1802.08786 [cs.LG] (2018).
72. M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN. arXiv:1701.07875 [stat.ML] (2017).
73. L. Yu, W. Zhang, J. Wang, Y. Yu, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *AAAI* (2017); www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14344/14489.
74. R. D. Hjelm *et al*., Boundary-seeking generative adversarial networks. arXiv:1702.08431 [stat.ML] (2017).
75. D. Silver *et al*., *Nature* **529**, 484–489 (2016).
76. X. Yang *et al*., *Sci. Technol. Adv. Mater.* **18**, 972–976 (2017).
77. M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *J. Cheminform.* **9**, 48 (2017).
78. M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for de-novo drug design. arXiv:1711.10907 [cs.AI] (2017).
79. M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **555**, 604–610 (2018).
80. Z. Zhou, X. Li, R. N. Zare, *ACS Cent. Sci.* **3**, 1337–1344 (2017).
81. A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, *Mol. Pharm.* **14**, 3098–3104 (2017).
82. G. L. Guimaraes, B. Sanchez-Lengeling, P. L. C. Farias, A. Aspuru-Guzik, Objective-Reinforced Generative Adversarial Networks (ORGAN) for sequence generation models. arXiv:705.10843 [stat.ML] (2017).
83. B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, A. Aspuru-Guzik, Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). *ChemRxiv* [Preprint] (2017).
84. W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation. arXiv:1802.04364 [cs.LG] (2018).
85. D. Bouchacourt, R. Tomioka, S. Nowozin, Multi-level variational autoencoder: learning disentangled representations from grouped observations. arXiv:1705.08841 [cs.LG] (2017).
86. C. Finn, P. Christiano, P. Abbeel, S. Levine, A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. arXiv:1611.03852 [cs.LG] (2016).
87. P. Nikolaev *et al*., *Npj Comput. Mater.* **2**, 16031 (2016).
88. S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, R. M. Myers, *Angew. Chem. Int. Ed. Engl.* **54**, 3449–3464 (2015).
89. P. J. Kitson *et al*., *Science* **359**, 314–319 (2018).
90. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, *Proc. IEEE* **104**, 148–175 (2016).
91. F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, PHOENICS: A universal deep Bayesian optimizer. arXiv:1801.01469 [stat.ML] (2018).
92. B. Huskinson *et al*., *Nature* **505**, 195–198 (2014).
93. L. M. Roch *et al*. *Sci. Robot*. **3**, eaat5559 (2018).

ACKNOWLEDGMENTS

We thank A. Fröseth for support of this work. A.A.-G. is a cofounder of Kebotix, a startup company that works in automated materials discovery.

10.1126/science.aat2663

REVIEW

Fundamentals of numerical relativity for gravitational wave sources

Bernd Brügmann*

Einstein's theory of general relativity affords an enormously successful description of gravity. The theory encodes the gravitational interaction in the metric, a tensor field on spacetime that satisfies partial differential equations known as the Einstein equations. This review introduces some of the fundamental concepts of numerical relativity—solving the Einstein equations on the computer—in simple terms. As a primary example, we consider the solution of the general relativistic two-body problem, which features prominently in the new field of gravitational wave astronomy.

The basic equations of general relativity are the Einstein equations, first published in 1915 (*1*). However, even today there are large gaps in our understanding of the physics implied by the Einstein equations. Stated in general terms, a major goal of research in general relativity is to solve the Einstein equations for the physical situations of interest. Fundamental analytic solutions of the Einstein equations include the flat Minkowski spacetime known from special relativity, the Schwarzschild and Kerr spacetimes describing single black holes, and the simple Big Bang cosmologies. Also predicted by general relativity are gravitational waves, which for weak fields can be obtained as analytic solutions of the linearized Einstein equations. However, the few known analytic solutions describe only very special situations, and approximation methods fail in the regime where the nonlinear, strong-field effects of relativity play a crucial role. If we are interested in the truly relativistic regime, we must turn to computer simulations to obtain numerical solutions to the full Einstein equations.

Solving the full Einstein equations on the computer is the subject of numerical relativity, which could also be called computational general relativity. Computers also play a role in algebraic computations and in approximation schemes, and such calculations are important topics in numerical relativity. But the distinguishing feature of numerical relativity is that, in principle, the Einstein equations in full generality can and must be solved numerically.

Numerical relativity spans a large range of different topics including mathematical general relativity, astrophysics, numerical methods for partial differential equations, computer programming, and simulation science. Current research in numerical relativity is in a transition from a self-contained topic in theoretical physics to a physical theory with numerous connections to observational astronomy (*2*, *3*). Gravitational wave astronomy holds much promise for the future, as recognized by the 2017 Nobel Prize in Physics, and numerical relativity is providing key theoretical predictions and analysis tools for the ongoing gravitational wave observations (*4*).

Theoretical Physics Institute, University of Jena, 07743 Jena, Germany.
*Corresponding author. Email: bernd.bruegmann@uni-jena.de

The general relativistic two-body problem

As a primary application of numerical relativity, we consider the gravitational two-body problem. The two-body problem in Newtonian gravitational physics can be formulated for two point masses moving in their mutual gravitational field. A particular solution of the Newtonian two-body problem is a Keplerian elliptical orbit. However, in Einsteinian gravity, such orbital motion generates gravitational waves that carry away energy and momentum. Binary orbits therefore decay, and the motion of the two bodies follows an inward spiral that eventually terminates with the collision and merger of the two objects. In most astrophysical situations, the energy loss due to the emission of gravitational waves is so small that a binary orbit decays only on time scales of millions or billions of years. However, for compact objects such as neutron stars or black holes in very tight binaries, general relativistic effects such as gravitational wave emission play a major role (*5*).

Research in this field seeks to provide a theoretical framework for the physics of binary black holes, neutron stars, and gravitational waves. Such an endeavor must rely on numerical simulations in general relativity and general relativistic hydrodynamics. But a reasonably complete framework still requires substantial progress in numerical relativity and related fields. Currently there are serious limitations in our ability to model the entire range of relevant physics, from the nuclear physics of neutron star matter to the large-scale, strong-gravity effects encountered in binary neutron star mergers (*6*). The different dynamical phases of the binary evolution—known as the inspiral, the merger, and the evolution of the remnant—are accompanied by characteristic gravitational wave signatures (Fig. 1). For binaries involving at least one neutron star, depending on the specifics of the system, key features include the disruption of the star(s) before merger, the formation of a hypermassive neutron star, the prompt or delayed collapse to a black hole, the dynamics of the accretion torus plus the central merged object, and the creation of unbound material, the ejecta. Before discussing simulations of these systems, we introduce the mathematical foundation of numerical relativity.

Mathematical foundation

Combining space and time into spacetime can be considered a triumph of human thought, allowing us to perceive the true nature of relativistic and gravitational physics (*7*). However, this does not mean that we cannot or need not consider space and time separately. Somewhat ironically, after working hard to unify space and time, the mathematical setup of numerical relativity starts by splitting spacetime again into space and time and by making gauge (coordinate) choices (*8*) in order to reformulate the Einstein equations as a well-posed mathematical problem.

General relativity is the theory of a metric tensor on a four-dimensional manifold, plus matter described by additional tensor fields. A manifold $\mathcal{M}$ endowed with a metric g_{ab} is called a spacetime $(\mathcal{M}, g_{ab})$. The metric measures lengths, here in four dimensions. The infinitesimal line element

$$ds^2 = g_{ab}dx^a dx^b \equiv \sum_{a,b=0}^{3} g_{ab}dx^a dx^b \qquad (1)$$

provides a generalization of the Pythagorean theorem. Repeated indices are summed over, following the Einstein summation convention. The metric is symmetric ($g_{ab} = g_{ba}$), has a Lorentz signature of − + + +, and there exists an inverse metric g^{ab} defined by $g^{ab}g_{bc} = \delta^a_c$, where δ^a_c is the identity matrix. A special example is the Minkowski line element $ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2$, where c is the speed of light, t is the time coordinate, and x, y, and z are spatial coordinates. The components of the Minkowski metric are constants, but in general g_{ab} is a field with nonconstant components.

The field equations of general relativity are the Einstein equations,

$$G_{ab}(g, \partial g, \partial^2 g) = 8\pi T_{ab}(g, \Phi) \qquad (2)$$

where G_{ab} is the Einstein tensor, which depends on the metric and its first and second derivatives, and T_{ab} is the stress-energy tensor constructed from the matter fields Φ and in general also from the metric. For numerical implementations, the first step is to write the Einstein equations as a well-posed system of partial differential equations (PDEs) for the metric. Equation 2 represents 10 coupled, nonlinear PDEs for the 10 independent components of the metric, but without further adjustments these equations are in no known sense hyperbolic (i.e., well-posed as an initial value problem).

The differential operator acting on the metric in the Einstein equations is given by the Ricci tensor,

$$R_{ab} = -\frac{1}{2}g^{cd}(\partial_c\partial_d g_{ab} + \partial_a\partial_b g_{cd} - \partial_c\partial_a g_{bd} - \partial_c\partial_b g_{ad}) + R^1_{ab}(g, \partial g) \quad (3)$$

The first term by itself, $g^{cd}\partial_c\partial_d g_{ab}$, which is often denoted ${}^{g}\Box g_{ab}$ (where is the d'Alembert operator), has the form of the principal part of a simple hyperbolic wave equation, but note that the metric appears in two places: as the wave field g_{ab} and as the inverse metric g^{cd} in the wave operator. The other second-derivative terms are not standard wave operators. The best we can say about the complete principal part in Eq. 3 is that it is quasi-linear in the metric; that is, it is linear in the highest-order derivatives but with coefficients that depend (nonlinearly) on the variable itself. The lower-order terms $R^1_{ab}(g, \partial g)$ are quite involved as well, with typical terms of the form $g^{-1}g^{-1}\partial g\partial g$. Approaching the problem in this way makes it difficult to recognize that these equations are describing the simple geometric concept of curvature and that there is a time evolution being defined. Further, their well-posedness properties are quite unclear.

The so-called 3+1 decomposition—for example, in the form of Arnowitt, Deser, and Misner (ADM) (*8*)—assumes that the manifold (at least locally) allows a split into time and space, $\mathcal{M} = R \times \Sigma$. Physics is then describable by time-dependent tensors on three-dimensional hypersurfaces Σ, which correspond to t = constant slices of $\mathcal{M}$, resulting in a "foliation" of spacetime in terms of three-dimensional spaces. Geometrically, we obtain a normal vector n^a to Σ that allows the decomposition of tensors in directions normal and tangential to the hypersurfaces. These are the time-like and space-like directions, respectively. For concreteness, we can assume coordinates $x^a = (t, x^i) = (t, x, y, z)$ with a time coordinate $x^0 = t$ and spatial coordinates x^i, where $i = 1, 2, 3$ and $a = 0, 1, 2, 3$.

Decomposing the Einstein equations is accomplished by projecting G_{ab} and R_{ab} in the directions parallel and orthogonal to n^a. We discover that the differential operator Eq. 3 leads to two types of equations: (i) evolution equations containing time derivatives, and (ii) four constraint equations that are essentially elliptic equations, highlighting the indeterminate type of Eq. 3. The constraints are the Hamiltonian constraint and the momentum constraints. The latter are reminiscent of the Gauss law constraint of electrodynamics, where the divergence of the electric field gives the charge density.

Given the evolution and constraint equations as PDEs, we still must choose a spatial and temporal domain with boundary conditions. For problems in astrophysics such as the two-body problem of black holes and neutron stars, we consider isolated systems where at large distances gravitational fields become weak and spacetime becomes asymptotically flat (in contrast to typical cosmological models). Because gravity is universally attractive and long-range, it is not natural to restrict a system to a finite box, especially given that the goal is to compute waves traveling to infinity. Nonetheless, a typical configuration for numerical simulations is a finite-size spatial domain (e.g., a sphere) with boundary conditions at some finite radius that implement the proper fall-off of the fields and an outgoing-wave boundary (*9*).

Features unique to numerical relativity are various aspects of black hole spacetimes, in particular the causal structure associated with black hole event horizons and the possibility of spacetime singularities. This latter aspect can be viewed as the problem of specifying additional boundaries that represent black holes within the simulation domain.

Building blocks of numerical relativity

To define a particular strategy to solve the Einstein equations, we consider the following building blocks that define the anatomy of a numerical relativity simulation, with a focus on the compact binary problem. The following items are certainly relevant to any evolution problem in computational physics: initial data, evolution, analysis, and numerics. We must specify the initial conditions, integrate the equations of motion to obtain the evolved data, and perform an analysis of the evolved data to extract physical information. The numerical treatment of each item may require the implementation of specific numerical techniques.

Evolution

Formulation: Choose one of many inequivalent formulations (i.e., choose variables and rewrite the Einstein equations to obtain a well-posed initial value problem). Choose the order of time and space derivatives; make structural choices about the gauge and the constraints.

Reformulating the Einstein equations as a well-posed initial value problem has been the subject of much research (*10*, *11*). To give an example, the result of the generalized harmonic gauge (GHG) formulation (*12*) can be cast in a standard first-order PDE form as

$$\partial_t u^\mu + A^{i\mu}{}_\nu(u)\partial_i u^\nu = S^\mu(u) \quad (4)$$

Here, the state vector u^μ collects all 10 components g_{ab}, the 40 first derivatives $\partial_c g_{ab}$, and a few additional fields depending on the formulation. In addition, there may be variables for the matter fields. Equation 4 for GHG is strongly and even symmetric hyperbolic (*12*) and is well suited for numerical implementation. Another standard way to proceed is the classic ADM formulation that makes the geometry of the 3+1 decomposition in time and space more explicit. Basic variables are the 3-metric g_{ij} and the extrinsic curvature K_{ij}, which is essentially the first time derivative of the metric (*8*). The ADM equations are only weakly hyperbolic and are not suitable for numerics. However, closely related systems, the so-called BSSN and Z4c formulations (*10*, *11*), are strongly hyperbolic. Most current simulations in numerical relativity rely on either the GHG or BSSN/Z4c families of formulations.

Constraint propagation: Maintain the constraints during evolution. Perform free evolutions and monitor the convergence of the constraints; use, for example, constraint damping to maintain the constraints explicitly.

Analytically, the constraints propagate; that is, if they are satisfied initially, they remain satisfied during a well-posed evolution. Numerically, even small rounding errors can trigger divergence from the constraint-satisfying solution, which can lead to a catastrophic failure of the simulation. How the constraints are controlled is a distinguishing feature of each formulation. A key ingredient in stable binary black hole evolutions (*13*) is the constraint-damping scheme (*14*). The Z4c formulation improves BSSN in the way the Hamiltonian constraint is treated, which leads to improved conservation of mass for neutron star simulations (*15*). Apart from instabilities, constraint violations in 3+1 relativity signify a problem with four-dimensional covariance. The 3+1 decomposition breaks

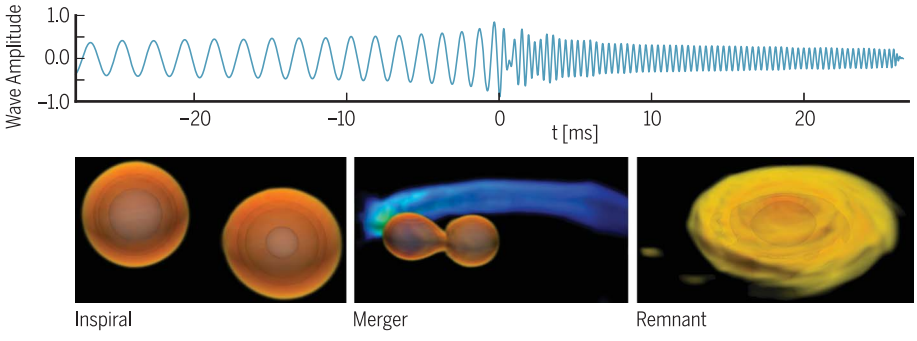


IMAGE: COURTESY OF T. DIETRICH, BASED ON (*78*)

Fig. 1. Binary neutron star mergers emit gravitational waves. The waves reveal unique information about extreme gravity and extreme matter—information that can be unraveled with the help of numerical relativity. Shown is a waveform and snapshots of the neutron star matter for the inspiral, merger, and remnant. The amplitude of the gravitational wave is plotted versus time. The merger occurs at $t = 0$.

covariance of the full theory by choosing a foliation, but the constraints ensure that four-dimensional covariance is maintained.

Gauge: Choose a coordinate condition—for example, in terms of lapse and shift or in terms of gauge source functions. Construct coordinates that avoid physical and coordinate singularities and are suitable for the black hole problem.

The main point about the gauge choice is that not only do we have the freedom to choose coordinates, but it is necessary to choose nontrivial coordinates. For example, even for the simplest black hole spacetimes, a foliation can fail by running into the physical singularity, and the hypersurface (or slice) may become badly distorted by slice stretching when points start falling into the black hole. The topic of how to dynamically construct good coordinates that lead to stable evolutions, cover spacetime with a regular foliation, avoid coordinate singularities, and avoid physical singularities inside black holes has become its own area of interest. In that context, the 3+1 decomposition is about "spacetime engineering" because we not only evolve the metric variables, but also build up the spacetime slice by slice in coordinates that are constructed dynamically during the evolution. The GHG formulation relies on the harmonic gauge to obtain hyperbolicity (*12*). For BSSN, the moving puncture gauge is essential to obtain long-term black hole evolutions, preventing slice stretching (*16*) and allowing the black hole punctures to move freely (*17*, *18*).

Boundary conditions: Specify outer boundary conditions appropriate for outgoing waves and asymptotic flatness. Specify inner boundaries for black holes; choose between black hole excision and black hole punctures. Handle coordinate patch boundaries.

Although approximate boundary conditions are possible, for a clean treatment, strong or symmetric hyperbolicity is required for well-posedness (*11*). We can then specify boundary conditions in terms of the ingoing and outgoing characteristic fields. The outer boundary conditions in numerical relativity tend to be substantially more complicated than the Einstein equations themselves, because outgoing-wave boundaries are typically constructed by taking additional derivatives of the right sides of the equations (*10*). The Einstein equations share with other nonlinear wave equations the feature that there is backscattering by waves off themselves (and, for binary systems, also due to the gradient in the gravitational well). This is a fundamental problem for boundaries at finite radius, because in principle we must account for all future backscattering from outside the domain. Consistent boundaries at finite radius have only been addressed quite recently, considering the long history of the Einstein equations (*19*, *20*).

Initial data

Formulation: Rewrite the constraints as elliptic equations, identifying suitable free and dependent variables.

To give an indication of what the formulation of the constraints entails (*8*), consider a conformal rescaling of the metric, $g_{ab} = \psi^4 \bar{g}_{ab}$, which conveniently transforms the Hamiltonian constraint into a scalar elliptic equation for ψ. We have the freedom to specify a conformal metric, $\bar{g}_{ab}$, that is not physical because it does not solve the constraints, but by solving the elliptic equation for ψ we find a physical solution g_{ab} that solves the constraints. The conformal transverse-traceless decomposition (*8*) is widely used for the full set of constraints; for neutron star initial data in particular, the conformal thin-sandwich construction (*21*, *22*) is used, where typically an additional elliptic equation is added to initialize the gauge condition.

Physical content: Solve the constraints for data that contain multiple black holes or neutron stars with arbitrary mass, spin, and momentum.

Because the constraints are nonlinear, we cannot simply "add up" the metric tensors of, for example, two Schwarzschild black holes to obtain binary data, although that can be a useful approximate initial guess. As a result, some aspects of the initial data construction are indirect. For example, we can start with two single black hole solutions for particular masses, which will be combined to form a binary. But solving the constraints for the binary leads to a change in the individual masses of the black holes because of the conformal rescaling. In some cases we have to perform evolutions to determine whether the initial data were constructed appropriately for a particular dynamical situation.

There is a growing variety of initial data constructions for binaries that correspond to the variety of physical configurations. For black holes, there are excision-type data, where the interior of black holes is removed (*23*, *24*). Alternatively, black hole puncture data handle the black hole interior with a coordinate singularity at a point (*25*), which sometimes is called automatic excision. The thin-sandwich formulation is well suited for quasi-equilibrium data of black holes and/or neutron stars, which, for example, can approximate the state of a binary system during a quasi-circular inspiral (*26*). Only quite recently have methods been developed for neutron stars that generalize the quasi-equilibrium, quasi-circular construction to eccentric orbits (*27*) and to neutron stars with spin (*28*) (Fig. 2). Solving the constraints for electromagnetic field configurations is another recent topic of investigation (*29*).

Analysis

Black holes and neutron stars: Determine all physical parameters during evolution. Find horizons of black holes. Analyze the rich phenomenology of neutron star mergers with the remnant, torus, jets, and ejecta. Connect to multi-messenger astronomy.

In any binary simulation, a wide range of detailed information is of interest, especially when matter is involved. The "relativity" in general relativity means, however, that many quantities have no direct physical meaning. In general, any tensor component (such as g_{tt} or g_{xy}) is not meaningful by itself; we have to construct proper gauge-invariant quantities. For example, mass and spin must be carefully defined because their local meaning at a point is problematic. For black holes, special methods are required to find the event horizon, which is a global concept in spacetime and therefore expensive to compute. See Fig. 3 for examples. Instead, black hole excision relies on the apparent horizon [e.g., (*13*)].

Gravitational waves: Compute gravitational wave emission; control numerical and systematic errors. Produce gravitational wave templates in a form that is ready to use for gravitational wave detectors. Treat both waveform prediction and waveform analysis.

Gravitational waves are propagating variations in the metric tensor, and the challenge is to separate the physical waves from various coordinate effects. In the weak-field limit, we can define gravitational waves as small perturbations around a background metric, and a first-order gauge-invariant formalism can be used to eliminate leading-order gauge effects (*30*). Such methods are applicable because we assume that

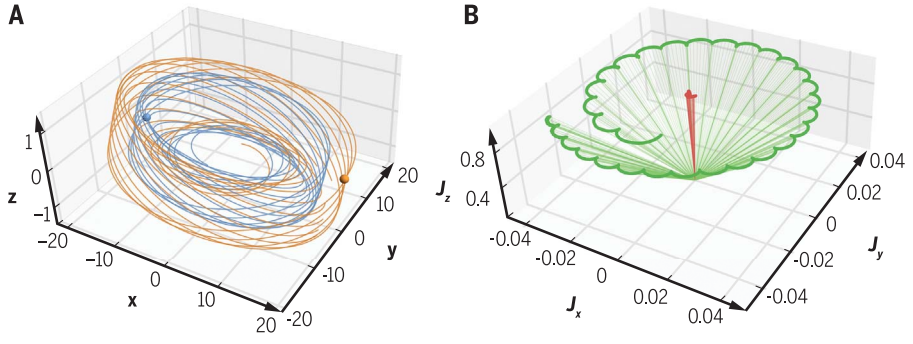


IMAGE: ADAPTED BY C. BICKEL FROM FIGURE 16 OF (*79*)

Fig. 2. Binary neutron star evolution with spin and precession. As a result of the general relativistic frame-dragging effect, a binary of neutron stars with (unaligned) spin will not move within a fixed orbital plane. (**A**) The orbital motion, indicated by two different colors for the two stars. (**B**) The angular momentum. Both can show precession and nutation effects, which will also be visible in the gravitational wave signal. The axes indicate spatial coordinates (A) and vector components of the spin (B).

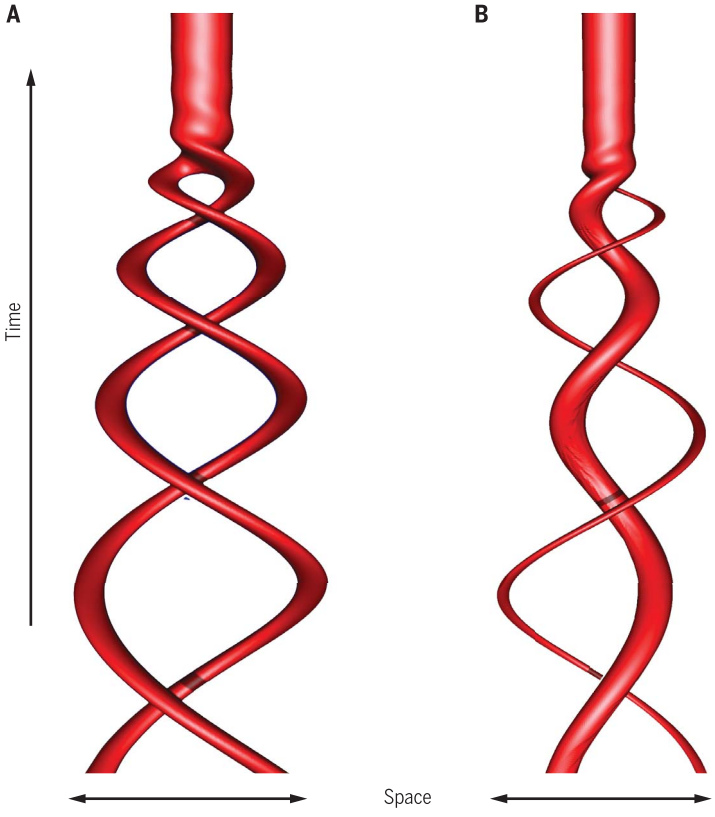


Fig. 3. The twisted pair of pants. (**A** and **B**) Spacetime plot of the event horizon of two inspiraling black holes that merge and ring down: equal (A) and unequal (B) masses, time *t* running up, horizontal *x*-*y* slices of the event horizon (*80*). (**C**) Pair of pants computed numerically in the 1990s for axisymmetric, head-on collisions, time *t* running up, horizontal slices in *z*-ρ coordinates (*81*).

IMAGE: ADAPTED BY C. BICKEL

the detectors are located far from the source where an asymptotically flat background is available. In simulations, the numerical grids often include extra patches for the far zone [e.g., (*9*)], possibly at lower resolution (see below).

A major effort in numerical relativity is directed toward obtaining accurate waveforms with controlled error bars for long time intervals. For the signal-to-noise ratio of current observations, a sufficiently accurate waveform model may begin with a post-Newtonian approximation (assuming nonrelativistic speeds) for the initial inspiral, matched to 10 to 20 orbits up to and including the merger from numerical simulations of the full Einstein equations. Initially, the goal was to filter the signal out of the noise by matching against theoretical waveforms. However, as the quality of the signals is improving, the main goal of gravitational wave astronomy is to estimate unknown source parameters. For example, we need detailed waveform models to distinguish black hole mergers from neutron star mergers, determine masses and spins, etc. The first detection of gravitational waves by Advanced LIGO (*2*) was accompanied by a theory paper describing how the properties of GW150914 were deduced from the observational data (*4*). Only by combining data with theory was it possible to arrive at the interpretation of GW150914 as the signature of a binary black hole merger, with specific parameters and credibility intervals. Two families of models were used, the EOBNR and Phenom families of waveforms (*2*) (Fig. 4). To analyze the data stream from the detectors, various parametrized waveform models are being developed for high-speed template matching (e.g., reduced-order surrogate models) (*31*).

Numerics

Discretization: Choose a discretization in space and time. Introduce adaptive mesh refinement (AMR) in space and time to efficiently represent different physical length scales. Choose coordinate patches and transformations to adapt coordinates to the underlying physics.

Once a suitably hyperbolic form of the PDEs of general relativity has been derived, we have access to several standard discretizations from applied mathematics. The recent trend has been toward high-order discretizations, with different choices for the geometry and the matter fields. In vacuum or where the matter is smooth, the geometry is smooth as well. For smooth metrics, fourth- to eighth-order finite differencing in space is applied routinely, as well as pseudospectral methods for exponential convergence. Neutron star matter is represented by general relativistic fluids, and handling relativistic shocks becomes important. Several high-resolution shock-capturing (HRSC) fifth-order methods are available (*6*), as is work on smoothed particle hydrodynamics (*32*, *33*).

The physics of a binary involves several physical scales. The wavelength of gravitational waves near merger is about 100 times the size of the black holes, and the simulation domain is typically chosen to be at least 1000 times the size of the black holes to accommodate several wave cycles. Simulations in three spatial dimensions therefore become several orders of magnitude more efficient with AMR, often of the Berger-Oliger type with refinements not just in space, but also in time. Many codes use several coordinate patches to transition from two (or more) central objects to spherical shells near the outer boundary.

Scientific computing: Implement parallel algorithms for high-performance computing. Invest in professional software engineering for a collaborative computational infrastructure.

Numerical relativity has been very successful with the hybrid MPI (message passing interface) plus OpenMP (open multiprocessing) or a similar parallelization strategy. Still, a typical numerical relativity simulation for a binary coalescence, representing just a single data point in a template catalog, may take roughly 1 month on 1000 to 10,000 cores of a supercomputer. The numerical relativity community is working on improving the efficiency of these methods, including spectral methods and improved AMR schemes, which tend to be a bottleneck for massive parallelism. Most efforts in numerical relativity are group efforts with a long-term investment in an evolving code base. These efforts include SpEC (*34*), SACRA (*35*), Whisky/THC (*36*), Pretorius (*37*), HAD (*38*), BAM (*39*), and the community code Einstein Toolkit (*40*). Some codes approximate general relativity but provide more advanced neutron star physics (*32*, *33*). Although similar in some regards—after all, the same or similar physics is studied—the different projects vary greatly in the range and the specifics of the physics modules, the flexibility and extensibility of the codes, the level of software optimization, and the collaboration and code-sharing models.

The main challenge common to all these projects is that they are implementing a "moving target," as formulations and basic equations are still changing and more physics is added to the simulations. Simultaneously, they must handle the trend in technology toward massively

parallel computers and heterogeneous hardware, which is challenging given the complex algorithms required for numerical relativity.

Short history of binary simulations

The first simulations of black holes in vacuum were attempted in 1964 (*41*). By the 1970s, many concepts of the 3+1 ADM formulation had been brought into numerical relativity (*42*), which led to the seminal numerical work on head-on (axisymmetric, 2+1-dimensional) black hole collisions and gravitational waves (*43*, *44*). It took until the early 1990s (*45*, *46*) to revisit the head-on collision with improved numerics, which confirmed the early results on gravitational waves (*46*). Numerical relativity in 3+1 dimensions began in 1995 with the evolution of a Schwarzschild black hole on a Cartesian grid (*47*) and the evolution of gravitational waves (*48*), followed by the first fully 3+1-dimensional simulation of a black hole binary (*49*, *50*). All the early black hole simulations mentioned so far were numerically unstable, with barely enough evolution time to start with two separate black holes that promptly merged. The first full orbit was achieved in 2004 (*51*). In 2005–2006, the last missing ingredients for long-term stable black hole evolutions were found in two different approaches, one based on a harmonic gauge formulation and excision (*13*) and the other based on the BSSN formulation and black hole punctures (*17*, *18*, *51*). By 2010, the robustness and flexibility of these methods had been established. Improvements in the formulations, the boundary conditions, etc., are still ongoing today (*11*, *12*, *15*).

Neutron star simulations were pursued in parallel with the black hole simulations. The Valencia formalism of general relativistic hydrodynamics (GRHD), now the primary approach, was developed in the 1990s (*52*). The first fully general relativistic binary neutron star simulations were published in 2000 (*53*), with enormous progress in many groups since then. As far as the geometry of general relativity matters in these simulations, it turns out that the methods established for stable black hole simulations carry over to neutron star simulations (gauge, boundaries, initial data formulation, etc.). However, GRHD introduces its own challenge of relativistic shocks, and the range of different physics phenomena makes this a much more complex problem than black holes in vacuum.

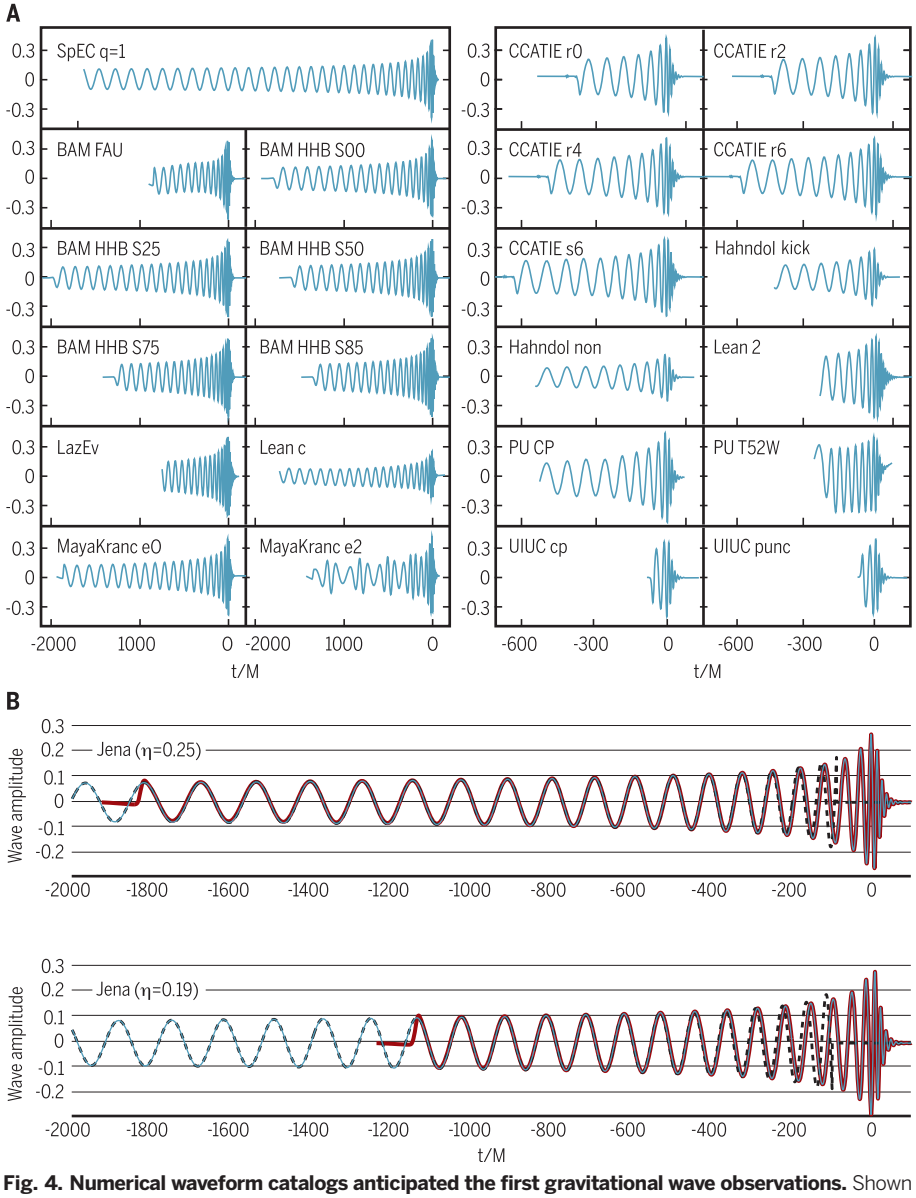


Fig. 4. Numerical waveform catalogs anticipated the first gravitational wave observations. Shown are examples for template construction for gravitational waves from binary black hole mergers. (**A**) Various numerical waveforms computed by different research groups forming an international collaboration. (**B**) Combining post-Newtonian models for the inspiral with numerical relativity. In (A) and (B), the amplitude of the gravitational wave is plotted versus time. The merger occurs at $t = 0$. In (B), the numerical waveform is preceded by a post-Newtonian waveform to cover more orbits of the inspiral. Such waveforms, which were purely theoretical, became real with the first observation of gravitational waves in 2015 [compare to figure 1 of (*2*)], making it possible to interpret the first signals as the merger events of two black holes.

IMAGES: (A) ADAPTED BY C. BICKEL FROM FIGURE 1 OF (*82*); (B) ADAPTED BY C. BICKEL FROM FIGURE 2 OF (*83*)

Outlook

Numerical relativity is developing rapidly in several directions, and we highlight a few representative examples.

High-order methods

High-order methods to address the ever-increasing demand for even more accurate and detailed simulations are a major topic of current research. Among the different high-order methods to solve partial differential equations, the discontinuous Galerkin (DG) method has emerged in recent years as a particularly successful general-purpose paradigm (*54*). It can be argued that the DG spectral-element method subsumes several of the key advantages of traditional finite-element and finite-volume methods. In particular, the DG method works with element-local stencils, which is a great advantage for parallelization and the construction of complicated grids. Furthermore, DG methods offer easy access to *hp*-adaptivity, where both the size of the computational elements (or cells) and the order of the polynomial approximation within each element can be adapted to the problem.

There are three major efforts to use DG methods for general relativity and/or GRHD (*55*–*57*). The first simulations of a single neutron star were achieved recently (*55*, *58*), and simple binaries are a work in progress. With regard to high-order approximations, there is no doubt that if exponentially convergent spectral methods such as DG (or pseudospectral methods) are applicable, they will constitute a big improvement over finite-difference approximations, which give only polynomial convergence. High-order methods can provide breakthroughs by reaching accuracies that make new physics possible (e.g., for magnetic field amplification due to small-scale turbulence) or by reducing numerical errors to

make gravitational wave analysis possible. Viewed differently, we can reach a given error criterion with much lower computational resources, making simulations feasible that are otherwise too computationally expensive.

Multi-physics

The spectacular first observation of both gravitational waves (*3*) and electromagnetic radiation (*59*, *60*) from a neutron star merger represents the beginning of multi-messenger astronomy including gravitational waves. To model such systems, we need to perform "multi-physics" simulations.

Modeling electromagnetic fields in GRHD can be accomplished by coupling the Maxwell equations to the GRHD equations, for which the prevalent approach has been ideal magnetohydrodynamics (IMHD). The assumption of IMHD is that the fluid has zero resistivity, but for the merger—and in particular for the fields surrounding the remnant with torus and ejecta—the quality of that approximation is unclear. Resistive magnetohydrodynamics (RMHD) is expected to be important for realistic models of plasma instabilities and magnetic reconnection. Apart from unknown physics, the mathematical character of the RMHD equations may be problematic (*61*, *62*). There are only a few general relativistic simulations with RMHD [e.g., (*61*, *63*, *64*)]. Developing a proper theory of resistive relativistic plasmas is a large project in itself (*65*).

The microphysical equation of state of neutron stars remains unknown and is also a target for numerical models and for observations. Investigations may include 20 or more different equations of state in an attempt to cover all sensible proposals. Even determining just one parameter—the existence of neutron stars with 2.0 solar masses (*66*, *67*)—provided a strong constraint. In principle, gravitational wave observations can do much better, gleaning information from the inspiral and the merger. Although inspiral signals will show rather systematic long-time effects (*68*–*70*), one of the grand challenges will be to disentangle the much more messy merger signal (*71*).

Standard merger models predict strong heating of the neutron star matter, which is expected to lead to an enormous amount of neutrino emission with luminosity on the order of 10^{54} erg s^{-1}. This burst of energy plays a role in models of short gamma ray bursts (*72*) and also for the ejecta, which in turn affects heavy-element production and macro- or kilonovae (*73*). However, currently the high dimensionality of such radiative transport problems (3+1 spacetime plus 3 for the radiative transport) is prohibitive, leading to a wide array of approximations with variable applicability (*74*, *75*). A coherent picture for neutrino physics in binary mergers is still lacking but should be a part of multi-messenger astrophysics.

Beyond current astrophysics

Numerical relativity has a large number of applications outside the area of compact binaries and gravitational waves (*76*, *77*). Topics include gravitational collapse with surprising critical phenomena, boson stars and other exotic matter, and cosmological simulations. Going beyond classical general relativity, the field of numerical relativity for alternative gravity theories and gravity in higher dimensions is wide open.

Conclusion

The next decade is sure to see numerical relativity grow in terms of computational power and applicability to different physical scenarios. The detailed theoretical models for black hole and neutron star binaries that are the target of research in numerical relativity are closely linked to the observation of gravitational waves. Numerical relativity, in combination with the highly anticipated future observations of gravitational waves, is expected to provide entirely new insights into extreme gravity and extreme matter.

REFERENCES AND NOTES

1. A. Einstein, *Preuss. Akad. Wiss. Berlin Sitzungsber.* **47**, 844–847 (1915).
2. B. P. Abbott *et al*., LIGO Scientific Collaboration and Virgo Collaboration, *Phys. Rev. Lett.* **116**, 061102 (2016).
3. B. P. Abbott *et al*., LIGO Scientific Collaboration and Virgo Collaboration, *Phys. Rev. Lett.* **119**, 161101 (2017).
4. B. P. Abbott *et al*., LIGO Scientific Collaboration and Virgo Collaboration, *Phys. Rev. Lett.* **116**, 241102 (2016).
5. B. S. Sathyaprakash, B. F. Schutz, *Living Rev. Relativ.* **12**, 2 (2009).
6. L. Baiotti, L. Rezzolla, *Rep. Prog. Phys.* **80**, 096901 (2017).
7. H. Minkowski, in *The Principle of Relativity*, H. A. Lorentz, A. Einstein, H. Minkowski, H. Weyl, Eds. (Dover, 1952), pp. 75–91.
8. J. W. York Jr., in *Sources of Gravitational Radiation*, L. Smarr, Ed. (Cambridge Univ. Press, 1979), pp. 83–126.
9. D. Pollney, C. Reisswig, E. Schnetter, N. Dorband, P. Diener, *Phys. Rev. D* **83**, 044045 (2011).
10. O. Sarbach, M. Tiglio, *Living Rev. Relativ.* **15**, 9 (2012).
11. D. Hilditch, *Int. J. Mod. Phys. A* **28**, 1340015 (2013).
12. L. Lindblom, M. A. Scheel, L. E. Kidder, R. Owen, O. Rinne, *Class. Quantum Gravity* **23**, S447–S462 (2006).
13. F. Pretorius, *Phys. Rev. Lett.* **95**, 121101 (2005).
14. C. Gundlach, G. Calabrese, I. Hinder, J. M. Martín-García, *Class. Quantum Gravity* **22**, 3767–3773 (2005).
15. D. Hilditch *et al*., *Phys. Rev. D* **88**, 084057 (2013).
16. M. Alcubierre *et al*., *Phys. Rev. D* **67**, 084023 (2003).
17. M. Campanelli, C. O. Lousto, P. Marronetti, Y. Zlochower, *Phys. Rev. Lett.* **96**, 111101 (2006).
18. J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, J. van Meter, *Phys. Rev. Lett.* **96**, 111102 (2006).
19. H. Friedrich, G. Nagy, *Commun. Math. Phys.* **201**, 619–655 (1999).
20. H.-O. Kreiss, J. Winicour, *Class. Quantum Gravity* **23**, S405–S420 (2006).
21. H. P. Pfeiffer, J. W. York, *Phys. Rev. D* **67**, 044022 (2003).
22. W. Tichy, *Rep. Prog. Phys.* **80**, 026901 (2017).
23. J. Thornburg, *Class. Quantum Gravity* **4**, 1119–1131 (1987).
24. S. Ossokine, F. Foucart, H. P. Pfeiffer, M. Boyle, B. Szilágyi, *Class. Quantum Gravity* **32**, 245010 (2015).
25. S. Brandt, B. Brügmann, *Phys. Rev. Lett.* **78**, 3606–3609 (1997).
26. E. Gourgoulhon, P. Grandclément, K. Taniguchi, J. Marck, S. Bonazzola, *Phys. Rev. D* **63**, 064029 (2001).
27. N. Moldenhauer, C. M. Markakis, N. K. Johnson-McDaniel, W. Tichy, B. Brügmann, *Phys. Rev. D* **90**, 084043 (2014).
28. W. Tichy, *Phys. Rev. D* **84**, 024041 (2011).
29. K. Uryū *et al*., *Phys. Rev. D* **90**, 101501 (2014).
30. E. T. Newman, R. Penrose, *J. Math. Phys.* **3**, 566–578 (1962).
31. J. Blackman *et al*., *Phys. Rev. Lett.* **115**, 121102 (2015).
32. A. Bauswein, H. T. Janka, R. Oechslin, *Phys. Rev. D* **82**, 084043 (2010).
33. S. Rosswog, *Living Rev. Comput. Astrophys.* **1**, 1 (2015).
34. M. D. Duez *et al*., *Phys. Rev. D* **78**, 104015 (2008).
35. T. Yamamoto, M. Shibata, K. Taniguchi, *Phys. Rev. D* **78**, 064054 (2008).
36. D. Radice, L. Rezzolla, *Astron. Astrophys.* **547**, A26 (2012).
37. W. E. East, F. Pretorius, B. C. Stephens, *Phys. Rev. D* **85**, 124010 (2012).
38. C. Palenzuela *et al*., *Phys. Rev. D* **92**, 044045 (2015).
39. M. Thierfelder, S. Bernuzzi, B. Brügmann, *Phys. Rev. D* **84**, 044012 (2011).
40. F. Löffler *et al*., *Class. Quantum Gravity* **29**, 115001 (2012).
41. S. G. Hahn, R. W. Lindquist, *Ann. Phys.* **29**, 304–331 (1964).
42. J. W. York, in *Sources of Gravitational Radiation*, L. L. Smarr, Ed. (Cambridge Univ. Press, 1979), pp. 83–126.
43. L. L. Smarr, thesis, University of Texas at Austin (1975).
44. K. R. Eppley, thesis, Princeton University (1975).
45. S. L. Shapiro, S. A. Teukolsky, *Phys. Rev. D* **45**, 2739–2750 (1992).
46. P. Anninos, D. Hobill, E. Seidel, L. Smarr, W.-M. Suen, *Phys. Rev. Lett.* **71**, 2851–2854 (1993).
47. P. Anninos, J. Massó, E. Seidel, W.-M. Suen, J. Towns, *Phys. Rev. D* **52**, 2059–2082 (1995).
48. M. Shibata, T. Nakamura, *Phys. Rev. D* **52**, 5428–5444 (1995).
49. B. Brügmann, *Int. J. Mod. Phys. D* **8**, 85–100 (1999).
50. M. Alcubierre *et al*., *Phys. Rev. Lett.* **87**, 271103 (2001).
51. B. Brügmann, W. Tichy, N. Jansen, *Phys. Rev. Lett.* **92**, 211101 (2004).
52. F. Banyuls, J. A. Font, J. M. Ibáñez, J. M. Martí, J. A. Miralles, *Astrophys. J.* **476**, 221–231 (1997).
53. M. Shibata, K. Uryu, *Phys. Rev. D* **61**, 064001 (2000).
54. J. S. Hesthaven, T. Warburton, *Nodal Discontinuous Galerkin Methods* (Springer, 2008).
55. M. Bugner, T. Dietrich, S. Bernuzzi, A. Weyhausen, B. Brügmann, *Phys. Rev. D* **94**, 084004 (2016).
56. L. E. Kidder *et al*., *J. Comput. Phys.* **335**, 84–114 (2017).
57. M. Dumbser, F. Guercilena, S. Köppel, L. Rezzolla, O. Zanotti, *Phys. Rev. D* **97**, 084053 (2018).
58. F. Hébert, L. E. Kidder, S. A. Teukolsky, General-relativistic neutron star evolutions with the discontinuous Galerkin method. arXiv:1804.02003 (5 April 2018).
59. B. P. Abbott *et al*., *Astrophys. J.* **848**, L12 (2017).
60. B. P. Abbott *et al*., *Astrophys. J.* **848**, L13 (2017).
61. C. Palenzuela, L. Lehner, O. Reula, L. Rezzolla, *Mon. Not. R. Astron. Soc.* **394**, 1727–1740 (2009).
62. A. Schoepe, D. Hilditch, M. Bugner, *Phys. Rev. D* **97**, 123009 (2018).
63. M. Ponce, C. Palenzuela, L. Lehner, S. L. Liebling, *Phys. Rev. D* **90**, 044007 (2014).
64. K. Dionysopoulou, D. Alic, L. Rezzolla, *Phys. Rev. D* **92**, 084064 (2015).
65. N. Andersson, G. L. Comer, I. Hawke, *Class. Quantum Gravity* **34**, 125001 (2017).
66. P. B. Demorest, T. Pennucci, S. M. Ransom, M. S. E. Roberts, J. W. T. Hessels, *Nature* **467**, 1081–1083 (2010).
67. J. Antoniadis *et al*., *Science* **340**, 1233232 (2013).
68. J. S. Read *et al*., *Phys. Rev. D* **88**, 044042 (2013).
69. K. Hotokezaka, K. Kyutoku, Y. Sekiguchi, M. Shibata, *Phys. Rev. D* **93**, 064082 (2016).
70. T. Dietrich, S. Khan, R. Dudi, S. J. Kapadia, P. Kumar, A. Nagar, F. Ohme, F. Pannarale, A. Samajdar, S. Bernuzzi, G. Carullo, W. Del Pozzo, M. Haney, C. Markakis, M. Pürrer, G. Riemenschneider, Y. E. Setyawati, K. W. Tsang, C. Van Den Broeck, Matter imprints in waveform models for neutron star binaries: Tidal and self-spin effects. arXiv:1804.02235 (6 April 2018).
71. A. Bauswein, H. T. Janka, *Phys. Rev. Lett.* **108**, 011101 (2012).
72. T. Piran, *Rev. Mod. Phys.* **76**, 1143–1210 (2005).
73. B. D. Metzger, Welcome to the Multi-Messenger Era! Lessons from a Neutron Star Merger and the Landscape Ahead. arXiv:1710.05931 (16 October 2017).
74. F. Foucart *et al*., *Phys. Rev. D* **94**, 123016 (2016).
75. K. Kyutoku, K. Kiuchi, Y. Sekiguchi, M. Shibata, K. Taniguchi, *Phys. Rev. D* **97**, 023009 (2018).
76. V. Cardoso, L. Gualtieri, C. Herdeiro, U. Sperhake, *Living Rev. Relativ.* **18**, 1 (2015).
77. M. W. Choptuik, L. Lehner, F. Pretorius, in *General Relativity and Gravitation: A Centennial Perspective*, A. Ashtekar, B. K. Berger, J. Isenberg, M. MacCallum, Eds. (Cambridge Univ. Press, 2015), pp. 361–411.
78. T. Dietrich, S. Bernuzzi, M. Ujevic, B. Brügmann, *Phys. Rev. D* **91**, 124041 (2015).
79. T. Dietrich *et al*., *Phys. Rev. D* **92**, 124007 (2015).
80. M. Thierfelder, thesis, University of Jena (2008).
81. R. A. Matzner *et al*., *Science* **270**, 941–947 (1995).
82. B. Aylott *et al*., *Class. Quantum Gravity* **26**, 165008 (2009).
83. P. Ajith *et al*., *Phys. Rev. D* **77**, 104017 (2008).

ACKNOWLEDGMENTS

I gratefully acknowledge the joint work evident from the list of references. Without my collaborators, this review would not have been possible. **Funding:** Supported in part by DFG/NSF grant BR 2176/5-1. **Author contributions:** B.B. is responsible for the entire manuscript. **Competing interests:** None. **Data and materials availability:** There are no new data in this review.

10.1126/science.aat3363

AAAS NEWS & NOTES

Public engagement helps scientists tackle global challenges

Leshner Leadership Institute fellows use public interactions for mutual learning

By **Anne Q. Hoy**

Over the coming year, scientist Merritt Turetsky will be working with the Royal Society of Canada to coordinate a series of public dialogues on challenges facing residents of Canada's Arctic region, home of the world's fastest-changing climate.

Climbing at twice the world's average rate, air temperatures throughout the Arctic region are altering ecosystems. In Canada's Northwest Territories, the shifts are impacting plants, animals, and residents and putting pressure on food and water security, Turetsky noted during a training program to make scientists effective listeners and communicators with the public. The fellowship program is offered under the auspices of the American Association for the Advancement of Science's Alan I. Leshner Leadership Institute for Public Engagement with Science.

Residents of Ottawa, Toronto, Montreal, and other Canadian cities will be invited to explore and discuss the transformations unfolding in Canada's northern reaches. Turetsky's strategic dialogues will follow the Science Café or World Café format designed to bring together stakeholders—in this case scientists, indigenous community members, and other segments of the public—to "provide a forum for sharing ideas on complex issues and gaining mutual understanding of the challenges ahead," said Turetsky. "We will use these events as an opportunity to bring stories from the Arctic to southern audiences."

One of 15 mid-career research scientists selected as 2018–19 AAAS Leshner Leadership Institute fellows, Turetsky is moving forward on one of the program's key objectives in spearheading public dialogues about Canada's northwest. The Arctic region was selected as a focus in recognition of Canada's 2018 presidency of the Group of Seven, the organization of industrialized democracies of the United States, Canada, France, the United Kingdom, Germany, Italy, and Japan.

The G7's scientific organizations committed to devoting their scientific expertise to the sustainability of Arctic communities, particularly those with changing ocean systems. The societies agreed to share research, open data-sharing platforms that build science capacity, and continue satellite monitoring of the Arctic region.

The program's roots

Turetsky, along with others, participated in the week-long, intensive public engagement training sessions beginning on 11 June, a program Leshner envisioned as a pathway to perfecting "two-way communication between the scientific enterprise and the community it is intended to serve."

The Leshner Leadership Institute fellowship was established in 2015 in honor of Leshner's long-standing dedication to enhancing discourse between scientists and the public following his retirement as AAAS's CEO and executive publisher of the *Science* family of journals and his earlier career as a neuroscientist, psychologist practitioner, and science agency administrator.

Leshner Leadership Institute fellows are called upon to leverage their scientific research and interact with the public through a series of activities focused on generating bidirectional dialogue about the promise and limits of science and technology in addressing significant societal issues. This year's group was selected to apply its expertise to food and water security.

The fellows make up the program's third contingent since its establishment. Like their predecessors, participants also share their training by mentoring other scientists, helping scientists become strategic in

PHOTO: MARY CATHERINE LONGSHORE/AAAS

Leshner Leadership Institute fellows recently gathered for an intensive week of public engagement training on food and water security.

their public outreach, and working within their academic institutions to establish public engagement programs.

"There is a tremendous need to bring the scientific community into closer contact with the rest of society in the form of a respectful, interacting dialogue," said Leshner. "That's a learned skill. That is not an innate skill and therefore there is a tremendous need to train people in how to do it. This approach is to train a small cadre and then have that cadre, in turn, leverage their experience by sharing what they've learned with the rest of their community."

With that statement, Leshner articulated what has become one of the program's central pursuits as it expands its network of scientists and engineers who recognize that researchers have an obligation to interact meaningfully with the public, particularly on issues of societal significance and concern. Significantly, Leshner's appreciation of a missing need came during a webinar presentation prerecorded to attract scientists to apply for the inaugural class in the fall of 2015. Knowing what the program needed to accomplish before it had gotten off the ground revealed how Leshner's vision has become an applied recognition that scientists have an obligation to interact attentively with the public.

The program's goals also represent a response to Leshner's view of the ineffectiveness of a previously held proposition that scientists need only to educate the public about emerging scientific issues to instill public understanding and build support. That approach "is not working," he said of what is known as the "deficit model" in a 2003 editorial published in *Science*.

The Leshner Leadership Institute fellows program was developed using relevant findings from social science research and the experiences of scientists and other professionals working in public engagement and science communication. It uses a logic model to guide practitioners in "intentional interaction with opportunities for mutual, two-way learning." Scientific researchers in the inaugural 2016–17 class focused on climate change, a topic of continuing public controversy. The second, in 2017–18, pursued the equally contentious area of infectious disease at a time of public misconceptions of the value of vaccines and risks of global pandemics.

"After this year's group completes its work on food and water security, the 2019–20 fellows are slated to explore public engagement around new technologies applied to humans, including artificial intelligence, nanotechnology, genetic engineering, and cognitive science," said Emily Cloyd, project director for the AAAS Center for Public Engagement with Science and Technology.

The fellows at work

Benjamin Preston, director of the RAND Corporation's Infrastructure Resilience and Environmental Policy Program and among the first group of Leshner Leadership Institute fellows, said the training helped him shift his focus to the audience, its views, and the impact of climate on its communities and positions when discussing the science of climate change.

The recitation of "overwhelming evidence of climate change and the significant consequences that change is having" often erodes an audience's view of a scientist's objectivity, Preston said. A more effective approach is one that focuses on the trade-offs.

"There are many voices talking about the consequences of failing to act on climate change, but few that acknowledge that both action and inaction have consequences that need to be managed," said Preston. "Rather than trying to defend a scientific finding or policy position, I choose to explore the trade-offs associated with different courses of action for things people care about."

For Maria Elena Bottazzi, associate dean of the National School of Tropical Medicine at Baylor College of Medicine and a member of the second group examining infectious disease, the fellowship came at an auspicious moment. Shortly after the program, Bottazzi was bestowed the 2018 Slim Health Award, a lifetime research achievement honor in recognition of her two decades of advancing health in Latin America and the Caribbean.

The distinction swept her into a swirl of public communications obligations. She was called upon to deliver speeches, conduct dozens of press interviews, and work with young Latin American and Caribbean scientists to expand their research capacity, among other commitments. "My ability to stand out and highlight efforts devoted to the dissemination of knowledge and the training of human resources, strengthening research capacities in the region, and, consequently, consolidating health systems for the benefit of the population has been directly related to the skills, tools, and practice I received during my year as a AAAS Leshner fellow," Bottazzi said.

Bottazzi and Luis Martinez, an infectious microbes research scientist at the University of Texas, El Paso, also a fellow, organized a science communications workshop in Houston to train his university's faculty members and students, an activity that contributed to expanding engagement capacity across research institutions.

> ***"We need to engage the public in a more open and honest bidirectional dialogue about science and technology."***
>
> **Alan I. Leshner,** AAAS CEO Emeritus

Engagement in the Arctic region

As part of the fellows' concentration on food and water security, Turetsky, an ecosystem ecologist at the University of Guelph in Ontario, has devoted her research to understanding the systems of the Arctic region, which, after oceans, represent the world's largest ecosystem.

Her work examines permafrost deterioration and the rising prevalence of wildfires in the Canadian northwest, events that threaten to increase the release of atmospheric gases including carbon dioxide and methane, emissions linked to climate change. More recently, she has focused on soil and water quality and their connection to the region's food and water security.

An example of rising concerns about water security has emerged from the once-considered-secure resting place for some 237,000 tons of highly toxic arsenic trioxide dust produced by the now-abandoned Giant Mine, a gold-mining facility that operated in Yellowknife, capital of the Northwest Territories, from 1948 until 2004. It had no pollution controls in its first 3 years of operation. Some 16,500 pounds of toxic dust a day spewed out of the smelters until 1953, settling on lakes and atop the area's terrain, according to a research report funded by the Social Sciences and Humanities Research Council of Canada.

Eventually, the toxic debris was captured and buried in storage chambers and underground rocky spaces from which gold was once extracted. The debris stockpiles have become unstable and threaten Yellowknife's groundwater supply. A remediation plan to freeze the stored toxic waste is being pursued after years of scientific examination.

Turetsky's work in both the research and public discussion of these issues demonstrates the range of challenges being addressed by AAAS Leshner Leadership Institute fellows, current and past, and the increasing need for effective engagement between scientists and the public.

As Leshner wrote in his 2003 *Science* editorial, "We need to engage the public in a more open and honest bidirectional dialogue about science and technology and their products, including not only their benefits but also their limits, perils, and pitfalls. We need to respect the public's perspective and concerns even when we do not fully share them, and we need to develop a partnership that can respond to them."

AAAS NEWS & NOTES

AAAS S&T Policy Forum explores U.S. competitiveness, now and in the future

The R&D ecosystem needs change to reap benefits from emerging innovation

By **Anne Q. Hoy**

The U.S. scientific research and development enterprise stands at a critical juncture that requires the convergence of basic research and commercial applications to more quickly deliver benefits of next-generation technological innovation, said Arati Prabhakar in a lecture at the American Association for the Advancement of Science's Forum on Science & Technology Policy.

Prabhakar, a fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University and former director of both the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology, also called on the R&D community to leverage advances in the social sciences to meet challenges of the coming decades.

Change was a theme that ran through the 43rd annual S&T Policy Forum on 21–22 June at AAAS headquarters in Washington, D.C., focused on U.S. science, innovation, and competitiveness. The forum featured speakers whose careers span public service, academia, and the private sector, including Susan Hockfield, president emerita of the Massachusetts Institute of Technology and chair of the AAAS board, and Peter Lee, corporate vice president of artificial intelligence and research at Microsoft.

"We need new kinds of advances and new ways of working if we are to contribute to our nation for the next handful of decades as we have for the decades just passed," said Prabhakar, whose career also has traversed Silicon Valley technological start-ups and venture capital firms. "That's what it will take if we are to renew and fulfill R&D's promise of a better future for our country."

Arati Prabhakar of Stanford delivers lecture at AAAS Science & Technology Policy Forum.

Closing the gap between high-quality, multidisciplinary basic research and the private sector's skillful talent for commercial success calls for novel approaches. Start-up incubators such as California's Cyclotron Road and The Engine at MIT give innovators time and facilities to turn basic research into commercial products, she said.

The convergence of social and behavioral sciences with mathematics, digital technology, and cryptographic methods also is producing technologies. Prabhakar pointed, for instance, to a research collaboration between University of Washington scientists and Enlearn, an educational nonprofit, that is advancing adaptive learning for individual students by linking their responses with visual content. Likewise, a Stanford medical team identified end-of-life care practices able to curb health care costs.

Accelerating R&D development was at the heart of Prabhakar's William D. Carey Lecture, an annual address begun in 1989 to honor former AAAS executive officer Carey by presenting speakers who exemplify his leadership on public policy issues.

On that point, Prabhakar said ethical and societal implications of new technologies are not to be overlooked. "We need to play our part in helping our society make wise choices about the fruits of our labor," she said.

Microsoft's Lee looked to R&D's future by tracing his journey from disappointing his father, a physics professor, and his mother, a chemistry professor, for pursuing mathematics at the University of Michigan and being drawn "to the crystalline beauty of some mathematical foundations in computer science."

"It was the stereotypical hard-core art, science Asian upbringing that you might read about in the papers," Lee deadpanned. He went on to establish a technology office at DARPA and head Carnegie Mellon University's computer science department.

The technology sector's growth, Lee said, is so exponential that it may call for a "fourth paradigm," to expand science beyond the experimental, theoretical, and computational sciences to data-driven science. One reason: the growth of data centers able to serve, store, and connect to an expanding range of digital tools through networks, a move away from giant computers.

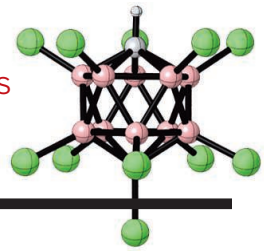
Microsoft's network more than doubles each year, Lee said, noting that it has added more than 250 data centers, including the June launch of an experimental, submerged data center on the seafloor off the Orkney Islands at Scotland's northern point. It is powered through tidal forces, is constructed of recycled materials, and connects to a modest 850 servers.

"We see these sorts of inflection points very infrequently," said Lee, discussing the growth of hyperscale data computing, the emergence of large-scale machine learning, and the possible inclusion of a new data-driven science paradigm. "We may in fact be in a similar type of inflection point today in the same way that the invention of mass-produced movable type, the Gutenberg press, was an industrial foundation for human knowledge."

RESEARCH

How to make vinyl cations

Popov et al., *p. 381*



IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

EUSOCIALITY

The benefits of being well fed

In eusocial insects, the vast majority of individuals sacrifice their reproductive potential to support the reproductive queen. Although this system has evolved repeatedly, there is still much debate surrounding its origin. Working with seven different species of ants, Chandra *et al*. used a transcriptomic approach to show that a single gene is consistently up-regulated in queens. This gene seems to confer reproductive status through integration with increased nutrition. In a clonal ant, larval signals disrupt this gene up-regulation, destabilizing the division of reproductive labor. Increasing levels of the associated peptide override these larval signals and establish eusociality. —SNV

Science, this issue p. 398

A transcriptomics approach reveals the underpinnings of queen versus worker ants.

DEVICE TECHNOLOGY

Cooler electrons for transistors

The operating power of field-effect transistors is constrained in part by the minimum change in voltage needed to change the current output. This subthreshold swing (SS) limit is caused by hotter electrons from a thermal electron source leaking over the potential of the gate electrode. Qiu *et al*. show that graphene can act as a Dirac source that creates a narrower distribution of electron energies. When coupled to a carbon nanotube channel, the decrease in SS would allow the supply voltage to be decreased from 0.7 to 0.5 volts. —PDS

Science, this issue p. 387

MICROBIOTA

Sticky bacteria tolerated as future food

Dictyostelium discoideum amoebae consume bacteria until the supply is exhausted. Then the amoeba cells clump together into a "slug" and initiate a complex multicellular reproductive phase. Specialized cells within aggregates rid the slug of any extracellular bacteria. However, some strains of amoeba tolerate live, intracellular bacteria. Dinh *et al*. discovered that these carrier strains bear surface lectins that bind *Klebsiella* bacteria, promote cell entry, and prevent the bacteria from being immediately digested. These bacteria then provide a future food source. Moreover, the internalized bacteria transfer DNA into the amoeba nucleus, resulting in transient genetic transformation. —CA

Science, this issue p. 402

NEUROSCIENCE

Support for the memory index theory

The link between contextual memory representations and locations or routes represented by hippocampal place cells during exploration remains unknown. Tanaka *et al*. examined spatial firing properties of neurons in hippocampal area CA1 on the basis of whether they had recently expressed the immediate-early activity-induced gene *c-Fos* in response to a novel context. The c-Fos–positive neurons displayed a more on-off firing pattern than the c-Fos–negative cells during context discrimination. In a contextual recognition paradigm, these results support the index theory of hippocampal function over a cognitive mapping theory. —PRS

Science, this issue p. 392

CREDITS: (TOP TO BOTTOM) S. POPOV *ET AL.*; DANIEL KRONAUER

GUT INFLAMMATION

Inflammation promotes gut polyposis

Peutz–Jeghers Syndrome (PJS) causes benign polyps in the gut and a higher risk of several cancers caused by mutations in the tumor suppressor gene *STK11*, which encodes liver kinase B1 (LKB1). LKB1's role in this disease is thought to be related to its tumor suppressor function. Now, Poffenberger *et al*. show that the T cell–specific heterozygous deletion of *Stk11* is sufficient to reproduce PJS symptoms in mice (see the Perspective by Hollstein and Shaw). Polyps in mice and humans are characterized by immune cell infiltration, enhanced STAT3 signaling, and increased levels of inflammatory cytokines such as interleukin-6 (IL-6). Targeting STAT3 signaling, IL-6, or T cells ameliorated the polyps, suggesting potential therapies for this disease. —STS

Science, this issue p. 406; see also p. 332

CLIMATOLOGY

Colder Siberian winters in a warming world

Winters have been more severe in Siberia over the past decade, even while loss of sea ice indicates dramatic warming of the Arctic. One hypothesis proposed to explain this apparent paradox is that Arctic sea ice loss forces a change in atmospheric circulation patterns that leads to colder Siberian winters. Using comprehensive atmospheric circulation models, Zhang *et al.* demonstrate that late autumn sea ice loss in the Barents-Kara Seas should result in a colder climate in Siberia the following winter. Their experiments also suggest that changes in the stratosphere, particularly shifts in and weakening of the stratospheric polar vortex owing to a reduction in sea ice, contribute substantially to the severity of winters in Siberia. —KVH

Sci. Adv. 10.1126/sciadv.aat6025 (2018).

CANCER

Blocking activin actively treats cancer

Platinum-based chemotherapy is a mainstay of treatment for lung cancer. However, resistance to this therapy is common, as are dose-limiting side effects, particularly kidney toxicity. To search for mechanisms contributing to treatment resistance, Marini *et al.* performed a whole-genome RNA interference screen, which implicated the activin pathway. Inhibition of this pathway offered a dual benefit by potentiating the effects of platinum drugs in mouse models of cancer and protecting the animals from kidney damage. Thus, activin inhibitors could be a valuable addition to platinum chemotherapy, enhancing the efficacy of treatment while also allowing higher doses or longer periods of drug use. —YN

Sci. Transl. Med. **10**, eaat3504 (2018).

Siberian winters are becoming colder, despite global warming.

IN OTHER JOURNALS

Edited by **Sacha Vignieri** *and* **Jesse Smith**

MARINE POLLUTION

Use the beach

Coastal power stations regularly draw water from the ocean as a cooling source, and this water is later returned to the marine environment. The water and any marine organisms that happen to be drawn in with it are exposed to high temperatures and other mechanical processes that generally reduce survival and affect populations near outflows. Jebakumar *et al*. looked at the impacts of this process on an Indian creek system influenced by power station cooling draws. They found reductions in abundance, decreases in survival, and changes in community structure across marine taxa influenced by the draws. Further, regulations to reduce the temperature of the water before it is returned, though necessary, increased the impact owing to the need to draw larger amounts of water into the plant for cooler outflows. The authors suggest that a relatively simple solution would be to widely institute subsurface intake systems, or "beach wells," which naturally filter out marine organisms before intake and improve water quality, across tropical regions. —SNV

Mar. Pol. Bull. 10.1016/j.marpolbul.2018.05.053 (2018).

Beach-filtered intake systems could decrease the impact of power plant draws on aquatic ecosystems.

EDUCATION

Taking anxiety out of active learning

As STEM (science, technology, engineering, and mathematics) education becomes more centered on active learning practices, what happens to students' anxiety levels? Specifically, what aspects of evaluative active learning practices cause student anxiety to increase or decrease? After measuring students' baseline anxiety levels, Cooper *et al*. conducted semi-structured interviews to explore how students' anxiety levels were altered in an active learning classroom. Results show that the way that the active learning activity is implemented and the extent to which students perceive the activity to be beneficial influence its effect on their anxiety. The authors encourage instructors to consider student anxiety when implementing active learning. —MMc

Int. J. STEM Educ. 10.1186/s40594-018-0123-6 (2018).

MOLECULAR BIOLOGY

tRNA lost in translation

Transfer RNAs (tRNAs), the adaptor molecules between messenger RNAs (mRNAs) and ribosomes during translation, are subjected to various types of chemical modifications, one of which is N^7-methyguanosine (m^7G). Mutations in the human m^7G methyltransferase complex lead to developmental disorders

PHOTOS: (LEFT TO RIGHT) DMITRY FEOKTISTOV/CONTRIBUTOR/GETTY IMAGES; MICHELLE MCMAHON/GETTY IMAGES

such as microcephalic primordial dwarfism and Down syndrome. Lin *et al*. mapped the m^7G tRNA methylome at single-nucleotide resolution and demonstrated its essential role in mouse embryonic stem cells. Depletion of members of the m^7G methyltransferase complex resulted in increased ribosome pausing on, and inefficient translation of, mRNAs involved in the cell cycle and brain development, thereby disrupting differentiation to neural lineages. This study is an important step toward a fuller understanding of how defects in tRNA methylation cause neurodevelopmental disorders. —SYM

Mol. Cell 10.1016/j.molcel.2018.06.001 (2018).

BIOMATERIALS

Using biology to remove cells

Decellularization is used to obtain extracellular matrix (ECM) scaffolds for use in tissue engineering and regenerative medicine, but the methods for stripping out the cells can be chemically harsh and cause damage to the ECM components or their overall architecture. Cornelison *et al*. treated peripheral nerve tissue with camptothecin, which triggers apoptosis, causing cell death and degradation. The fragments and any remaining cytotoxic compounds were removed with only a mild wash using saline solution or deoxyribonuclease, leaving the nerve scaffold intact without substantive loss of collagen or glycosaminoglycan content. In vivo testing in rats found no immunogenic response and a lower histological score than in isograft negative controls after 4 weeks. —MSL

Acta Biomat. 10.1016/j.actbio.2018.07.009 (2018).

NEURODEGENERATION

Alzheimer's disease in a dish?

Alzheimer's disease is characterized by extracellular deposits of β-amyloid (Aβ) and intracellular aggregation of hyperphosphorylated tau in the brain. These lead to the hyperactivation of glial cells in the brain and neuronal loss. However, it has been difficult to serially link this series of pathological events. Park *et al*. developed a three-dimensional triculture system composed of human-derived neurons, astrocytes, and microglia. The model demonstrated hallmark features of Alzheimer's disease and Aβ aggregation, phospho-tau accumulation, and neuroinflammation. The ability to see ongoing microglia recruitment in the microfluidic platform and consequent neurotoxicity bodes well for the utility of this type of approach as a platform for dissecting pathomechanisms and preliminary testing of potential therapeutics. —SMH

Nat. Neurosci. **21**, 941 (2018).

CANCER IMMUNOLOGY

Neutrophils take a different tack

Antibody-dependent cellular cytotoxicity is an important process by which cancer cells can be targeted and killed. Immune cells, such as T and NK (natural killer) cells, express Fc receptors that recognize cancer cells coated with antibodies. They then release cytotoxic granules that result in apoptosis. Matlung *et al*. show that neutrophils induce antibody-dependent cellular cytotoxicity via a completely different mechanism. This process entails trogocytosis-induced lysis of tumor cells, which the authors term "trogoptosis." They find that neutrophils endocytose cytoplasmic fragments of target cells, which induces cancer cell necroptosis. In addition to direct killing, this mechanism may be important for the release of cancer neoantigens and damage-associated molecular patterns, which further activate and direct the immune response to tumors. —STS

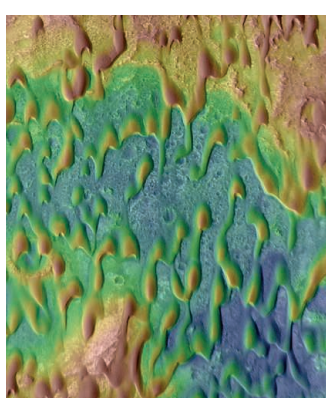
Cell Rep. **23**, 3946 (2018).

MARTIAN GEOLOGY

Ghost dunes spotted on Mars

Sand dunes are imposing physical structures, but few signs of them are preserved in the geological record because their grains do not stick together and are moved by wind. Preservation can occur if a dune field is partially buried by lava or mud, which then forms a layer of hard rock. Subsequent erosion of the sand leaves holes where the dunes once were. These features, termed ghost dunes, were only recently identified on Earth. Day and Catling used images from the Mars Reconnaissance Orbiter spacecraft to identify two locations on Mars where there are hundreds of ghost dunes. Revealing long-vanished dunes will enable studies of Mars' ancient winds. —KTS

J. Geophys. Res. Planets 10.1029/2018JE005613 (2018).



Voids in martian surface strata show where sand dunes once existed.

CREDIT: M. D. DAY AND D. C. CATLING, *J. GEOPHYS. RES. PLANETS* 10.1029/2018JE005613 (2018)

ALSO IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

MOLECULAR BIOLOGY

Phase separation and gene control

Many components of eukaryotic transcription machinery—such as transcription factors and cofactors including BRD4, subunits of the Mediator complex, and RNA polymerase II—contain intrinsically disordered low-complexity domains. Now a conceptual framework connecting the nature and behavior of their interactions to their functions in transcription regulation is emerging (see the Perspective by Plys and Kingston). Chong *et al*. found that low-complexity domains of transcription factors form concentrated hubs via functionally relevant dynamic, multivalent, and sequence-specific protein-protein interaction. These hubs have the potential to phase-separate at higher concentrations. Indeed, Sabari *et al.* showed that at super-enhancers, BRD4 and Mediator form liquid-like condensates that compartmentalize and concentrate the transcription apparatus to maintain expression of key cell-identity genes. Cho *et al.* further revealed the differential sensitivity of Mediator and RNA polymerase II condensates to selective transcription inhibitors and how their dynamic interactions might initiate transcription elongation. —SYM

Science, this issue p. 378, p. 379, p. 412; see also p. 329

BIOTECHNOLOGY

Transcriptome mapping in the 3D brain

RNA sequencing samples the entire transcriptome but lacks anatomical information. In situ hybridization, on the other hand, can only profile a small number of transcripts. In situ sequencing technologies address these shortcomings but face a challenge in dense, complex tissue environments. Wang *et al*. combined an efficient sequencing approach with hydrogel-tissue chemistry to develop a multidisciplinary technology for three-dimensional (3D) intact-tissue RNA sequencing (see the Perspective by Knöpfel). More than 1000 genes were simultaneously mapped in sections of mouse brain at single-cell resolution to define cell types and circuit states and to reveal cell organization principles. —SYM

Science, this issue p. 380; see also p. 328

CLIMATE

Improving climate model projections

How much will Earth's atmosphere warm if the carbon dioxide concentration doubles? Climate models give different answers to this question, making it difficult to determine how much carbon can be emitted while avoiding dangerous levels of climate change. In a Perspective, Soden *et al.* argue that the radiative forcing (i.e., the warming) from carbon dioxide is much less uncertain than model results suggest. The uncertainty is reduced substantially when a more sophisticated method is used to calculate the forcing, but models cannot adopt this method because it is too computationally expensive. Ensuring that model parameterizations are vetted against the more accurate calculations would help to reduce uncertainty in this crucial component of climate modeling. —JFU

Science, this issue p. 326

CANCER THERAPY

Modeling chemotherapy-drug combinations

Pharmacological inhibitors of DNA repair pathways can enhance the efficacy of chemotherapy. Alkan *et al*. modeled cancer cell responses to various chemotherapies and tested the efficacy of combining chemotherapy with DNA repair inhibition. In addition to the expected synergy between chemotherapy and an inhibitor of ATR (a DNA repair pathway kinase) in cultured cells and in vivo, there were some unexpected differences in outcome when applying low- versus high-dose chemotherapy regimens. Thus, modeling in cells can help inform therapeutic development. —LKF

Sci. Signal. **11**, eaat0229 (2018).

T CELL DIFFERENTIATION

Signal strength seals fate

T cell differentiation into effector and memory T cell subsets is influenced by T cell receptor (TCR) signals. Snook *et al*. examined how TCR signals influence $CD4^+$ T cell differentiation by using a panel of cloned TCRs that recognize the same MHC (major histocompatibility complex) class II–restricted epitope of lymphocytic choriomeningitis virus. Strong TCR signals were associated with T helper (T_H1) differentiation, whereas lower TCR signal strength corresponded with follicular helper T (Tfh) and memory T cell differentiation. Low CD25 expression by early effector T cells also predicted memory differentiation, although CD25 expression levels were not predictive of recall responses. Enhanced TCR signaling via knockdown of SHP-1 favored T_H1 over Tfh and memory T cell differentiation. Thus, stronger TCR signaling promotes terminal effector T_H1 differentiation. —CNF

Sci. Immunol. **3**, eaas9103 (2018).

ORGANIC CHEMISTRY

Silicon smooths the way to vinyl cations

Saturated carbon centers often undergo substitution reactions by initial cleavage of a carbon-halogen or carbon-oxygen bond, which leaves the carbon positively charged. Analogous cations have proven much harder to access from vinylic carbons that are engaged in double bonds. Popov *et al*. now show that silicon cations paired with noncoordinating anions can pull triflate groups off such vinylic carbons under ambient conditions (see the Perspective by Kennedy and Klumpp). The resultant vinyl cations react with simple alkanes through C–H insertion. Theoretical and mechanistic studies suggest that these reactions proceed through nonclassical pathways that bifurcate after the transition state. —JSY

Science, this issue p. 381; see also p. 331

RESEARCH ARTICLE SUMMARY

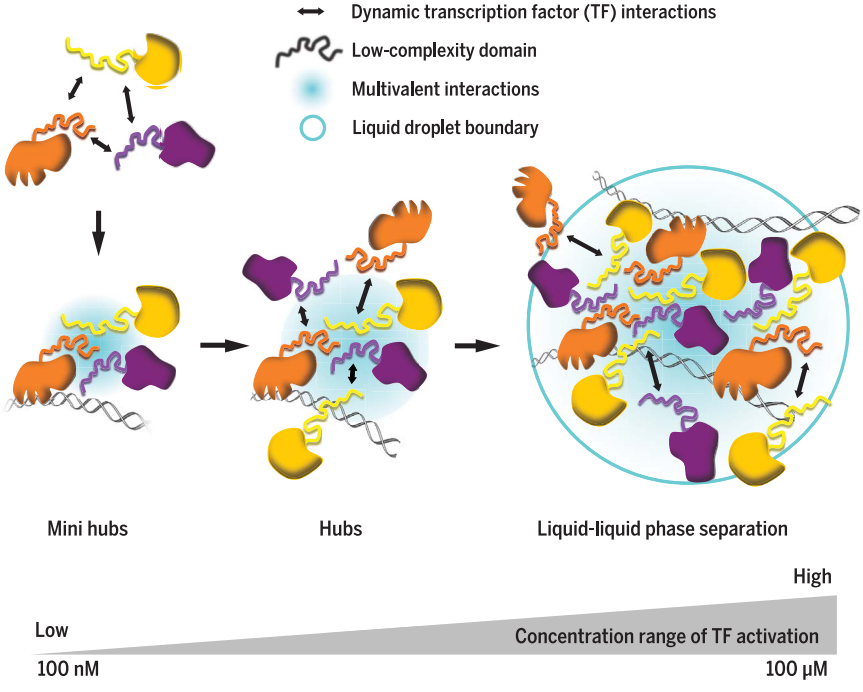
MOLECULAR BIOLOGY

Imaging dynamic and selective low-complexity domain interactions that control gene transcription

Shasha Chong, Claire Dugast-Darzacq, Zhe Liu, Peng Dong, Gina M. Dailey, Claudia Cattoglio, Alec Heckert, Sambashiva Banala, Luke Lavis, Xavier Darzacq, Robert Tjian*

INTRODUCTION: DNA binding transcription factors (TFs) are quintessential regulators of eukaryotic gene expression. Early studies of TFs revealed their well-structured DNA binding domains (DBDs) and identified functionally critical activation domains (ADs) required for transcription. It later became evident that many ADs contain intrinsically disordered low-complexity sequence domains (LCDs), but how LCDs activate transcription has remained unclear. Although it is known that transcriptional activation by LCDs requires selective interaction with binding partners, it has been challenging to directly measure selective LCD-LCD recognition in vivo and unravel its mechanism of action.

RATIONALE: Traditional biochemical reconstitution and genetics studies have identified most of the molecular players central to transcription regulation. However, the mechanism by which weak, dynamic protein-protein interactions drive gene activation in living cells has remained unknown. Advances in live-cell single-molecule imaging have opened a new frontier for studying transcription in vivo. In this study, we used synthetic LacO (Lac operator) arrays as well as endogenous GGAA microsatellite loci to study LCD-LCD interactions of TFs such as EWS/FLI1, TAF15, and Sp1 in live cells. To probe the dynamic behavior of TF LCDs at target genomic loci, we have combined CRISPR-Cas9 genome editing, mutagenesis, gene activation, cell transformation assays, and various high-resolution imaging approaches including fluorescence correlation spectroscopy, fluorescence recovery after photobleaching, lattice light-sheet microscopy, three-dimensional DNA fluorescence in situ hybridization, and live-cell single-particle tracking.



From hubs to phase separation: Activation occurs in a wide range of TF concentrations. In vivo LCD-dependent transactivation occurs in hubs formed over a broad range of TF concentrations (100 nM to 100 μM) and time scales (<1 s to minutes). At endogenous concentrations, TF LCDs form transactivation hubs at native genomic loci without undergoing evident phase separation. Upon TF LCD overexpression, phase separation is observed at synthetic TF binding site arrays.

RESULTS: Live-cell single-molecule imaging revealed that TF LCDs interact to form local high-concentration hubs at both synthetic DNA arrays and endogenous genomic loci. TF LCD hubs stabilize DNA binding, recruit RNA polymerase II (RNA Pol II), and activate transcription. LCD-LCD interactions within hubs are highly dynamic (seconds to minutes), selective for binding partners, and differentially sensitive to disruption by hexanediols. These findings suggest that under physiological conditions, rapid, reversible, and selective multivalent LCD-LCD interactions occur between TFs and the RNA Pol II machinery to activate transcription. We observed formation of functional TF LCD hubs at a wide range of intranuclear TF concentrations. Although we detected apparent liquid-liquid phase separation with gross overexpression of LCDs, transcriptionally competent TF LCD hubs were observed at physiological TF levels at endogenous chromosomal loci in the absence of detectable phase separation. In addition, mutagenesis, gene expression, and cell transformation assays in Ewing's sarcoma cells revealed a functional link between LCD-LCD interactions, transactivation capacity, and oncogenic potential.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aar2555

CONCLUSION: The use of various imaging methods in live cells powerfully complements in vitro studies and provides new insights into the nature of LCD interactions and their role in gene regulation. We propose that transactivation domains function by forming local high-concentration hubs of TFs via dynamic, multivalent, and specific LCD-LCD interactions. It also seems likely that weak, dynamic, and transient contacts between TFs play a role in disease-causing dysregulation of gene expression (i.e., EWS/FLI1 in Ewing's sarcoma), suggesting that LCD-LCD interactions may represent a new class of viable drug targets. Although we examined a small subset of TF LCDs, the principles uncovered regarding the dynamics and mechanisms driving LCD-LCD interactions may be applicable to other classes of proteins and biomolecular interactions occurring in many cell types. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jmlim@berkeley.edu
Cite this article as S. Chong *et al.*, *Science* **361**, eaar2555 (2018). DOI: 10.1126/science.aar2555

RESEARCH ARTICLE

MOLECULAR BIOLOGY

Imaging dynamic and selective low-complexity domain interactions that control gene transcription

Shasha Chong[1,2], Claire Dugast-Darzacq[1,3], Zhe Liu[4], Peng Dong[4], Gina M. Dailey[1], Claudia Cattoglio[1,2], Alec Heckert[1], Sambashiva Banala[4], Luke Lavis[4], Xavier Darzacq[1,3], Robert Tjian[1,2,3]*

Many eukaryotic transcription factors (TFs) contain intrinsically disordered low-complexity sequence domains (LCDs), but how these LCDs drive transactivation remains unclear. We used live-cell single-molecule imaging to reveal that TF LCDs form local high-concentration interaction hubs at synthetic and endogenous genomic loci. TF LCD hubs stabilize DNA binding, recruit RNA polymerase II (RNA Pol II), and activate transcription. LCD-LCD interactions within hubs are highly dynamic, display selectivity with binding partners, and are differentially sensitive to disruption by hexanediols. Under physiological conditions, rapid and reversible LCD-LCD interactions occur between TFs and the RNA Pol II machinery without detectable phase separation. Our findings reveal fundamental mechanisms underpinning transcriptional control and suggest a framework for developing single-molecule imaging screens for drugs targeting gene regulatory interactions implicated in disease.

Sequence-specific DNA binding transcription factors (TFs) are preeminent players in eukaryotic gene regulation. From the earliest studies of human TFs, it was recognized that regulatory proteins such as Sp1 contain well-structured DNA binding domains (DBDs) and functionally critical transactivation domains that participate in specific TF-TF interactions to direct gene transcription (*1*–*3*). Numerous atomic structures of DBDs have provided a concrete understanding of TF-DNA interactions. In contrast, many transactivation domains contain low-complexity sequence domains (LCDs) that persist in an intrinsically disordered conformation not amenable to conventional structural determination. Mutations in TF LCDs not only disrupt transcription but also have been implicated in cancer and neurodegenerative disorders (*4*, *5*). However, the mechanism by which TF LCDs execute specific transactivation functions has remained an enigma. Elucidation of how LCDs operate in vivo, given the dynamic nature of TF-TF interactions required for gene regulation, has been an equally challenging problem.

Several in vitro studies have suggested that purified LCDs from the FET protein family (FUS, EWS, and TAF15) can undergo reversible hydrogel formation or liquid-liquid phase separation at high concentrations and low temperatures (*6*–*8*). Moreover, the C-terminal domain of RNA polymerase II (RNA Pol II) is itself an LCD and can undergo phase separation (*9*) and be incorporated into FET LCD hydrogels in a phosphorylation-regulated manner (*10*). FET LCDs were also reported to undergo phase separation in live cells upon overexpression (*7*, *11*).

However, stark differences exist between in vivo physiological conditions and those used for in vitro or overexpression studies. Temperature, protein concentration, purity, and microenvironment may all substantially affect the behavior of LCDs. There is also a vigorous debate as to whether LCDs undergo cross-β polymerization or remain in a disordered conformation when interacting with partners (*6*–*8*, *11*–*17*). From the perspective of elucidating how TFs work in vivo, an equally pressing unresolved mechanistic question concerns the dynamics and time scales governing LCD-LCD interactions that would allow TFs to function in rapid cellular processes. Selectivity of cognate LCD-LCD interactions is another important yet poorly understood feature that is required for proper TF function in vivo. Thus far, selective LCD-LCD recognition has not been directly demonstrated in vivo, let alone understood at a mechanistic level. In this study, we combined a variety of high-resolution imaging strategies—including fluorescence correlation spectroscopy (FCS) (*18*, *19*), fluorescence recovery after photobleaching (FRAP) (*20*), lattice light-sheet microscopy (*21*), three-dimensional (3D) DNA fluorescence in situ hybridization (FISH) (*22*), and live-cell single-particle tracking (SPT) (*23*, *24*)—to probe the dynamic behavior of TF LCDs at target genomic loci under physiological conditions.

Synthetic LacO arrays mediate formation of LCD interaction hubs

We first established proof-of-concept experiments by using a synthetic Lac operator (LacO) array (~50,000 LacO repeats) integrated into the genome of human U2OS cells (*25*) that express various enhanced yellow fluorescent protein (EYFP)–tagged TF LCDs fused to the Lac repressor (LacI) (Fig. 1A). To probe potential sequence specific LCD-LCD interactions, we examined two distinct classes of LCDs: QGYS-rich LCDs from the FET family and a QGTS-rich LCD from Sp1 that is low in tyrosine (table S1) (Q, Gln; G, Gly; Y, Tyr; S, Ser; T, Thr).

As expected, the LacO array recruits a large number of EYFP-LCD-LacI molecules via targeted DNA binding, forming a concentrated local interaction hub in the nucleus. LacO-associated hubs formed by LCD-LacI but not LacI are visible by bright-field microscopy (Fig. 1B), suggesting that the refractive index and mass density of LCD-LacI hubs differ considerably from the surrounding nuclear environment.

We found that both TAF15-LacI and Sp1-LacI give rise to much brighter and larger LacO-associated LCD hubs than LacI alone. To quantify this effect, we used two orthogonal methods to measure the protein concentrations in live cells. Specifically, we performed fluorescence intensity and FCS measurements of intracellular protein and compared the results with standard concentration curves of a purified fluorescent tag (EYFP or mCherry) calibrated by the same method (fig. S1). Next, we estimated the absolute protein copy number in each hub by using the average in-hub protein concentration (fig. S2A) and hub dimensions measured in single-cell images (fig. S2B). The two independent concentration measurements consistently showed that the LCD-LacI copy number in the LacO-associated hub increases with the TF nuclear concentration much faster than with LacI alone and reaches levels up to orders of magnitude greater than the number of LacO repeats available for direct LacO-LacI binding (Fig. 1C and fig. S2, C and D), suggesting likely cooperative multivalent interactions at the LacO array that is contributed by extensive LCD self-interactions. Similarly, smaller LacO arrays containing substantially fewer (~15,000) LacO repeats similarly nucleate LCD self-interactions (Fig. 1D and fig. S2, E and F).

Moreover, LCD-LacI but not LacI alone can form hundreds of smaller puncta throughout the nucleus once its intranuclear concentration reaches a certain threshold (Figs. 1B and 2A, and fig. S2, G and H). LCDs can form intranuclear puncta in some cases even without being fused to a DBD such as LacI (Fig. 3A, bottom). These results suggest that LCD-LCD interactions can promote self-assembly of LCD hubs upon overexpression without assistance from DNA (*7*, *11*).

In addition, FRAP dynamics of LCD-LacI at the LacO array was also significantly different

[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA. [2]Howard Hughes Medical Institute, University of California, Berkeley, CA, USA. [3]CIRM Center of Excellence, University of California, Berkeley, CA, USA. [4]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA.
*Corresponding author. Email: jmlim@berkeley.edu

from that of LacI (Fig. 1E). Because diffusion contributes negligibly to the FRAP dynamics (fig. S3, A and B), such differences can be attributed to changes in dissociation rates. Specifically, when we fit the FRAP curves with a reaction-dominant model (*26*), we found that fusing TAF15 or FUS LCD to LacI leads to more than 60% reduction in the dissociation rate constant of LacI (fig. S3, C to F). This result suggests that at increased local TF concentrations, TF LCD hubs driven by LCD-LCD interactions stabilize TF binding to its cognate genomic site via multivalent contacts that could include both DBDs for tethering to chromatin and one or more LCDs within TF proteins that can form multiple transient interactions with different partner proteins.

LCD hubs interact with RNA Pol II

Having demonstrated homotypic LCD self-interactions, we next investigated the potential role of heterotypic LCD-LCD interactions in hub formation. First, we tested whether TF LCD hubs can interact with RNA Pol II in vivo by using a LacO-containing U2OS line in which we replaced the endogenous RPB1 (major and catalytic subunit of RNA Pol II) with an α-amanitin–resistant Halo-tagged RPB1 (*27*). We subsequently labeled the cells with a fluorescent HaloTag ligand and visualized RNA Pol II distribution in vivo. We found that mCherry-FET-LacI expression mediates significant enrichment of RNA Pol II in LacO-associated hubs compared with background levels recorded using LacI alone (Fig. 1, F and G, and fig. S4, A to C). Moreover, self-assembled LCD-LacI hubs that are unaffiliated with a LacO array also enrich RNA Pol II (fig. S4D), suggesting that LCD hubs interact with RNA Pol II potentially without assistance from DNA. While recapitulating the in vitro incorporation of RNA Pol II into LCD hydrogels (*10*), these experiments go one step further and suggest that LCD hub formation can facilitate the recruitment of the general transcription machinery in vivo—a key step toward transactivation.

LCD-LCD interactions are sequence specific and differentially sensitive to hexanediol disruption

Next, we probed the sequence specificity of interactions between various classes of LCD. To this

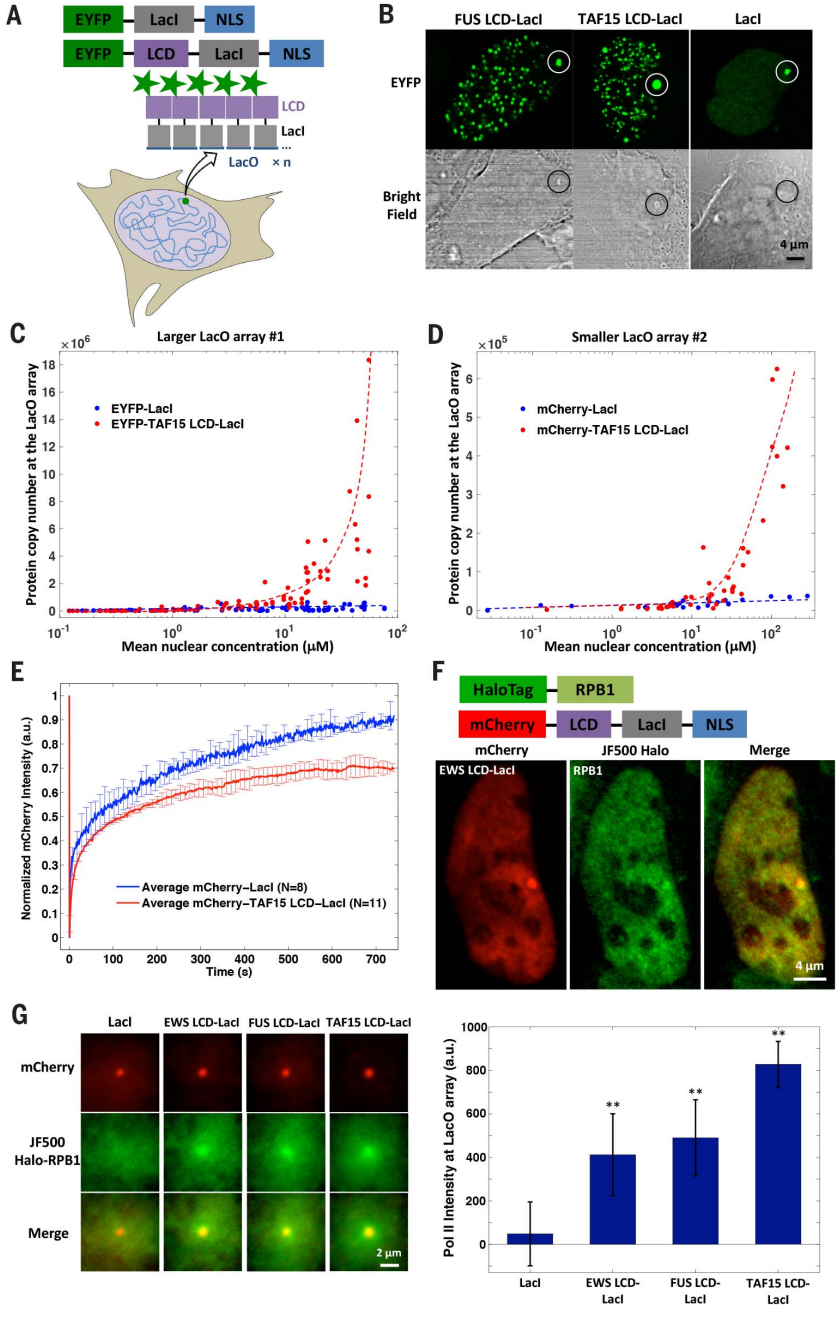


Fig. 1. A LacO array can mediate the formation of an LCD hub in live cells, which involves extensive LCD self-interaction and recruits RNA Pol II. (**A**) Schematic for a LacO array ($n \approx 50{,}000$ repeats for array 1, $n \approx 15{,}000$ repeats for array 2) in the U2OS genome nucleating an LCD hub when EYFP-LCD-LacI is transiently expressed. Alternatively, EYFP-LacI is expressed as a control. NLS, nuclear localization signal. (**B**) Confocal fluorescence and bright-field images of LacO-containing U2OS cells where LacO array 1 (highlighted by circles) is bound by EYFP-labeled LCD-LacI or LacI. LCD-LacI–bound, but not LacI-bound, LacO arrays are visible in bright-field images. (**C** and **D**) Copy number of EYFP-labeled (C) or mCherry-labeled (D) TAF15 LCD-LacI (red) or LacI (blue) molecules bound to LacO array 1 (C) or 2 (D) as a function of mean nuclear concentration of the TF. Concentrations were measured by fluorescence intensity comparison. Each dot represents one cell. (**E**) Averaged FRAP curves at LacO array 1 bound by mCherry-labeled TAF15 LCD-LacI (red) or LacI (blue). Error bars represent SD. a.u., arbitrary units; *N*, number of cells analyzed. (**F**) (Top) Schematic of the proteins expressed in the LacO-containing U2OS line. (Bottom) Confocal fluorescence images show that Halo-RPB1 (labeled with 200 nM Halo ligand JF500, green) is enriched at LacO array 2 bound by mCherry-EWS LCD-LacI (red). (**G**) (Left) Averaged Halo-RPB1 images at LacO array 2 bound by mCherry-labeled LacI, EWS LCD-LacI, FUS LCD-LacI, or TAF15 LCD-LacI (N = 55, 69, 81, or 143). (Right) Fluorescence intensity of Halo-RPB1 at the LacO array center in the average images after subtraction of nuclear Halo-RPB1 background (see supplementary methods). ** denotes a statistically significant increase compared with the LacI condition ($P < 0.01$, two-sample t test). Error bars represent bootstrapped SD (*45*).

end, we coexpressed both EYFP-LCD-LacI and mCherry-LCD in LacO-containing cells. mCherry-LCD lacking a DBD becomes enriched at the array only when it can interact with the coexpressed LCD that is fused to LacI (Fig. 2A and fig. S5A). Notably, the array can enrich mCherry-LCD over a wide range of expression levels. The EWS-LacI–bound LacO array also enriches endogenous EWSR1, as detected by immunofluorescence (fig. S5B). Therefore, mCherry-LCD enrichment at the array is most likely due to specific LCD-LCD interactions rather than potential nonspecific overexpression artifacts. Using this two-color imaging assay, we confirmed homotypic self-interactions of all tested LCDs (from the FET family and Sp1). Intriguingly, although all three FET LCDs interacted among themselves, none of them interacted with the Sp1 LCD (Fig. 2, A and B), suggesting that LCD interactions exhibit strong sequence specificity that is likely an essential feature underlying combinatorial TF regulation of gene expression.

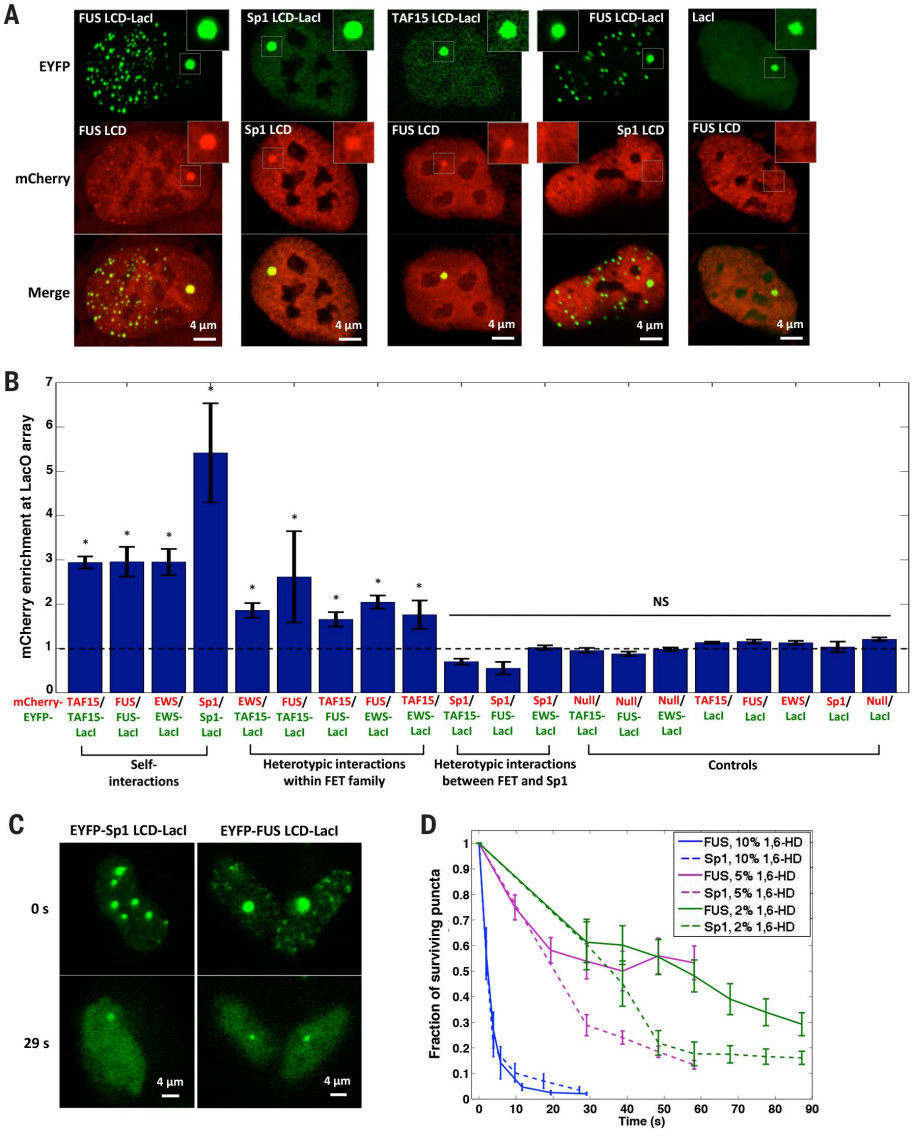


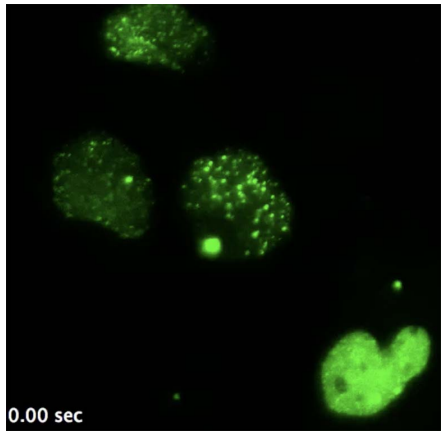
Fig. 2. LCD hub formation involves selective protein-protein interactions, which can be disrupted by 1,6-HD with sequence-dependent sensitivity. (**A**) Confocal fluorescence images of U2OS cells containing LacO array 1 that coexpress various combinations of mCherry-LCD and EYFP-LCD-LacI. The region surrounding the LacO array is zoomed in. (**B**) Quantification of the enrichment of mCherry-LCD (red) at the LacO array 1 bound by various EYFP-labeled LCD-LacI fusion proteins (green), calculated as the peak mCherry fluorescence intensity at the array divided by the average intensity immediately surrounding the array (fig. S5A). Null, mCherry not fused to any LCD. An mCherry enrichment at the array above 1 suggests LCD-LCD interactions. * denotes a statistically significant difference above 1 ($P < 0.05$, one-sample t test). NS, nonsignificant difference above 1. Error bars represent SE. (**C**) Fluorescence images of FUS and Sp1 LCD hubs before (0 s) and after (29 s) addition of 10% 1,6-HD. (**D**) Number of nuclear puncta formed by FUS or Sp1 LCD surviving over time upon addition of 1,6-HD at different concentrations. Error bars represent SE.

To better understand the nature of LCD-LCD interactions, we treated cells with 1,6-hexanediol (1,6-HD), an aliphatic alcohol known to dissolve various intracellular membraneless compartments and FUS hydrogels in vitro through disruption of hydrophobic interactions (*28*–*30*). We observed that both FUS and Sp1 LCD hubs rapidly disassemble within 30 s when exposed to 10% 1,6-HD. The LacO-associated LCD–induced hub shrank to a size comparable to that of the array bound by LacI alone, whereas all nuclear puncta not associated with LacO disappeared (Fig. 2C and Movie 1). We also found that 2,5-hexanediol (2,5-HD), a less hydrophobic derivative of 1,6-HD that barely melts FUS hydrogels in vitro (*28*), disrupts LCD hubs less efficiently in live cells (fig. S5, C and D). This correlation between hydrophobicity of hexanediols and LCD hub melting suggests that these aliphatic alcohols may directly influence LCD-LCD interactions by disrupting key hydrophobic contacts. These in vivo results also mirror in vitro hydrogel studies using these same disrupting agents (*28*).

Sp1 LCD hubs were disrupted significantly faster and more extensively than FUS LCD hubs with 2 or 5% 1,6-HD (Fig. 2D). Thus, although a combination of intermolecular forces may contribute to LCD hub formation, our results indicate that hydrophobic interactions might be more sensitive to disruption and play a more dominant role in Sp1 LCD self-interactions than FUS LCD, consistent with the Sp1 LCD containing hydrophobic residues sparsely interspersed among Q repeats (*31*). The differential sequence dependence of LCD-LCD interactions revealed by 1,6-HD treatment may be correlated to the selectivity of homo- and heterotypic LCD interactions observed above.

LCD-LCD interactions are highly dynamic

To study the dynamics of protein-protein interactions between LCD pairs, we coexpressed



Movie 1. LacO array–associated and self-aggregated non-array–affiliated FUS LCD hubs in live cells are quickly disrupted by 10% 1,6-hexanediol (related to Fig. 2C).

EYFP-LCD-LacI and Halo-LCD in the LacO-containing U2OS cells and performed SPT of Halo-LCD to measure residence times (RTs) of LCD-LCD interactions within the LacO-associated hub (Fig. 3A, top). For all LCDs tested, RTs resulting from self-interactions fell in the range of 11 to 33 s (Fig. 3B). When EYFP-LCD and Halo-LCD from the FET family were coexpressed at high levels, they spontaneously formed hubs that are unaffiliated with the array and resemble intranuclear puncta (Fig. 3A, bottom). These non-array hubs bind Halo-LCD via homo- or heterotypic interactions with even shorter RTs (7 to 10 s). As expected, the Sp1 LCD that failed to interact with the FUS LCD had an RT of <1 s at the non-array FUS LCD hubs (Fig. 3B). The fact that RTs of many LCDs in self-aggregated hubs unaffiliated with genomic DNA are substantially shorter than in hubs formed at the LacO array suggests that TF-DNA interactions that maintain a high local concentration of TF LCDs contribute to stabilizing LCD-LCD interactions and vice versa. Together, these findings reveal the rapid, reversible, and interdependent nature of LCD-LCD and TF-DNA interactions as well as their propensity to form local high-concentration hubs that likely stabilize multicomponent complexes—e.g., transcription preinitiation complex, a prerequisite for transactivation.

EWS/FLI1 forms hubs at endogenous GGAA microsatellites

Having unmasked the sequence specificity and dynamic nature of LCD-LCD interactions by using synthetic LacO arrays in living cells, we next tested LCD behavior at native GGAA microsatellites (>20 GGAA repeats) in the Ewing's sarcoma cell line A673 (*32*–*35*). These cancer-derived cells carry a chromosomal translocation t(11;22)(q24;q12) producing a fusion oncogene, *EWS/FLI1*, that encodes a potent TF consisting of a trans-activating LCD from EWSR1 and the DBD from FLI1 that targets GGAA sequences (Fig. 4A).

To visualize the behavior of endogenously expressed EWS/FLI1, we fused a HaloTag to its DBD using CRISPR-Cas9–mediated genome editing of A673 cells (Fig. 4, A and B, and fig. S6A) (*36*). This knock-in strategy allowed us to image fluorescently tagged endogenous EWS/FLI1 at its normal expression levels (Fig. 4B), which was essential because LCDs tend to self-aggregate and behave aberrantly upon overexpression. To ensure that Halo-tagging does not disrupt transactivation functions of EWS/FLI1, we confirmed that EWS/FLI1-Halo activates a luciferase reporter construct containing a GGAA microsatellite–driven promoter (*33*) as efficiently as wild-type (WT) EWS/FLI1 (fig. S6C). More importantly, using the gold standard neoplastic transformation assay (*37*), we confirmed that the EWS/FLI1-Halo knock-in A673 cells form colonies in soft agar much like the WT A673 cells—albeit, less efficiently (fig. S6, D and E).

We next performed high-resolution lattice light-sheet microscopy and found that EWS/FLI1 forms many small interaction hubs (>1000 per nucleus) in the nucleus (Fig. 4C and Movie 2).

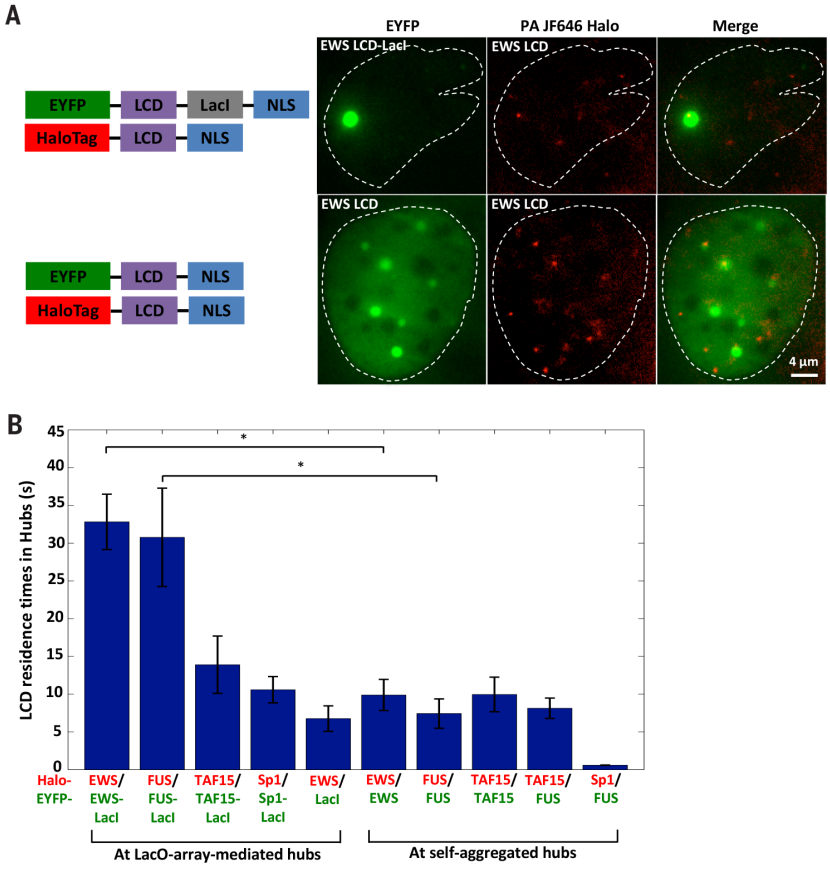


Fig. 3. LCD-LCD interactions involved in hub formation are highly dynamic. (**A**) Snapshots of a two-color SPT movie simultaneously imaging EYFP-labeled (green) EWS LCD-LacI (top, forming a LacO-associated LCD hub) or EWS LCD (bottom, forming self-aggregated LCD hubs not affiliated with the LacO array) and Halo-tagged EWS LCD (2 nM PA-JF646 labeled, red) in U2OS cells containing LacO array 1. A white dashed contour outlines the cell nucleus. We imaged the hubs in the EYFP channel (green) and tracked individual Halo-EWS LCD molecules with an acquisition time of 500 ms in the PA-JF646 channel (red). (**B**) Residence times of LCD (red) bound at the LacO-array–associated LCD hub or at self-aggregated LCD hubs not affiliated with the array (green). *P < 0.05, two-sample t test. Error bars represent SE.

The detected number of intranuclear hubs has the same order of magnitude as the total number of EWS/FLI1-bound GGAA microsatellites across the human genome (~6000) estimated by chromatin immunoprecipitation sequencing and bioinformatics analyses (*38*). To examine the spatial relationship between EWS/FLI1 hubs and GGAA microsatellites, we performed simultaneous confocal imaging of EWS/FLI1-Halo and 3D DNA FISH targeting genes adjacent to GGAA microsatellites that are regulated by EWS/FLI1 (Fig. 4D), including *CAV1*, *FCGRT*, *ABHD6*, *KDSR*, and *KIAA1797* (*33*, *39*). Although EWS/FLI1 enrichment is detected at many single loci of these genes, the crowded distribution of intranuclear EWS/FLI1 hubs makes it difficult to clearly visualize EWS/FLI1 enrichment at single target loci. By recording images of ~1000 loci for each gene, the signal-to-noise ratio is markedly improved to reveal specific EWS/FLI1 enrichment at GGAA repeats, whereas no enrichment was seen at non-GGAA gene loci (Fig. 4E). These results indicate that EWS/FLI1 forms hubs at endogenous GGAA microsatellite DNA elements.

Dynamic EWS LCD-LCD interactions mediate formation of EWS/FLI1 hubs

We previously showed that because of LCD-LCD interactions, the LCD-LacI copy number in the LacO-associated hub increases much faster with the TF concentration than that of LacI alone, reaching levels that are orders of magnitude higher than the number of available TF binding sites at the LacO array (Fig. 1, C and D, and fig. S2, C to F). Using the methods established earlier, we estimated the intranuclear concentration of endogenously Halo-tagged EWS/FLI1 in A673 cells to be ~200 nM and the median copy number of EWS/FLI1 per hub at GGAA microsatellites to be 24 (fig. S7, A and B). It was reported that most EWS/FLI1-bound microsatellites contain 11 to 19 GGAA motifs, with a median around 15 (*38*). Because the DBD of FLI1 occupies two consecutive GGAA repeats (*33*), the median

number of EWS/FLI1 molecules recruited to a microsatellite via direct DNA-protein interaction is estimated to be 8. The fact that there are a significantly greater number of EWS/FLI1 molecules in a GGAA-affiliated hub than what direct binding to GGAA repeats could accommodate suggests that, like the synthetic LacO arrays, these native genomic elements can also efficiently nucleate EWS LCD-LCD interactions to form local high-concentration TF hubs. Moreover, when transiently expressing EWS/FLI1-Halo or a Halo-tagged LCD deletion mutant (FLI1 DBD) in U2OS cells, we again observe that the total number of EWS/FLI1 molecules in the GGAA-affiliated hubs increases much faster than FLI1 DBD as a function of TF concentration (fig. S7, C and D), providing strong evidence that LCD-LCD interactions occur at GGAA-affiliated EWS/FLI1 hubs.

We previously observed that the formation of an LCD interaction hub slows down dissociation of LCD-LacI from the LacO array (Fig. 1E and fig. S3F). If LCD-LCD interactions are also involved in EWS/FLI1 hub formation at GGAA microsatellites, we expect the RT of EWS/FLI1 within GGAA-affiliated hubs to be longer than that of EWS/FLI1 outside hubs. We stained the EWS/FLI1-Halo knock-in A673 cells with two fluorescent ligands (*40*, *41*): High-concentration JF549 staining allows visualization of EWS/FLI1 hubs in the cell nucleus, whereas low-concentration PA-JF646 staining allows real-time tracking of individual EWS/FLI1 molecules (Fig. 5A and Movie 3). SPT revealed the average RTs of EWS/FLI1 in and outside the GGAA-affiliated hubs to be 90 and 16 s, respectively (Fig. 5B and fig. S8, A and B). The fact that EWS/FLI1 binds to GGAA repeats for a significantly longer time again suggests that EWS LCD-LCD interactions are involved in the formation of the GGAA-affiliated hubs. It is very likely that both LCD-LCD interactions and DBD binding to GGAA repeats work together to stabilize hub formation, much as we observed for LCD-LacI at the LacO array.

To confirm that hub formation in this native setting is dependent on the EWS LCD, we determined how mutations in the LCD might affect RTs of EWS/FLI1. We started by replacing different numbers (m = 3, 7, 10, 17, or 29) of tyrosines

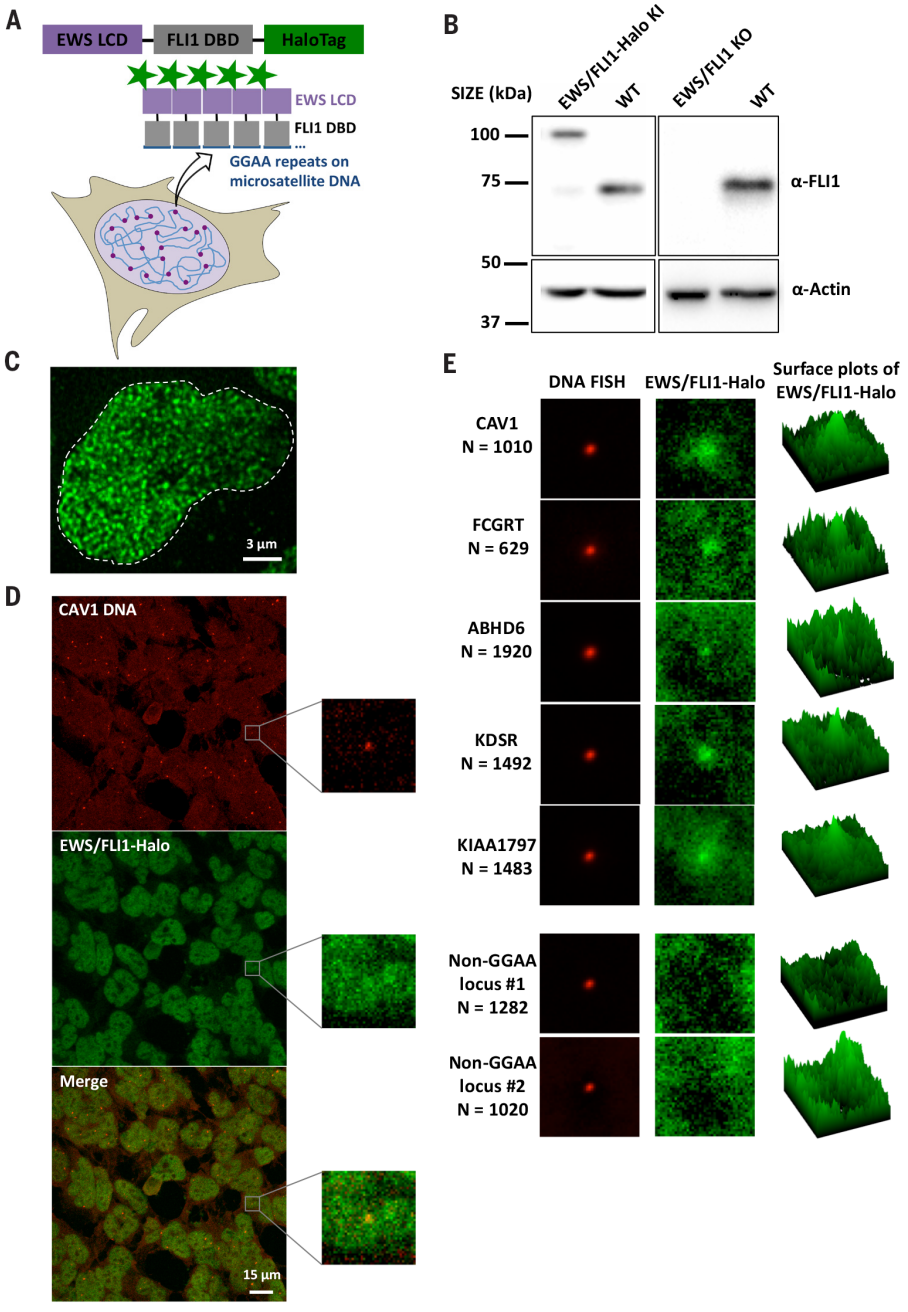


Fig. 4. Combined DNA FISH and EWS/FLI1-Halo imaging show that endogenous EWS/FLI1 forms hubs at GGAA microsatellites. (**A**) Schematic for GGAA microsatellites in the A673 genome nucleating hubs of endogenously Halo-tagged EWS/FLI1. (**B**) Western blot of EWS/FLI1 and β-actin (normalization control) from clonal EWS/FLI1-Halo knock-in (KI), WT and clonal EWS/FLI1 knockout (KO) A673 lines. (**C**) *z*-projected 3D image of endogenous EWS/FLI1-Halo in an A673 cell nucleus (stained with 200 nM Halo ligand JF549) taken on the lattice light-sheet microscope. (**D**) Confocal fluorescence images of 3D DNA FISH targeting GGAA microsatellite–adjacent *CAV1* gene (enhanced Cy5 labeled, red) and endogenous EWS/FLI1-Halo (JF549 labeled, green). The zoomed-in views depict the region surrounding one particular *CAV1* locus. EWS/FLI1-Halo enrichment at the locus is visible but buried in high background. (**E**) Averaged two-color images of five GGAA microsatellite–adjacent gene loci (*CAV1*, *FCGRT*, *ABHD6*, *KDSR*, and *KIAA1797*) and two gene loci not containing a GGAA microsatellite (Non-GGAA locus 1 targeting *ADGRA3* and locus 2 targeting *REEP5*). The right column shows average surface plots of EWS/FLI1-Halo.

(Y) in the EWS LCD (residues 47 to 266 of EWSR1) with serines (S) and then testing the self-interaction capability of mutant LCDs [EWS(YS_m)] using the LacO array assay established earlier. As previously shown, when we coexpressed EYFP-EWS-LacI and mCherry-EWS, the mCherry signal became enriched at the LacO array owing to EWS LCD self-interaction. Notably, when we replaced WT EWS in both fusion proteins with EWS(YS_m), mCherry enrichment at the array progressively decreased with an increasing number of Y-to-S mutations (fig. S9A) and vanished for EWS(YS_{29}), in which all tyrosines are replaced (Fig. 5C). Similarly, we found that EWS(YS_{29}) does not interact with

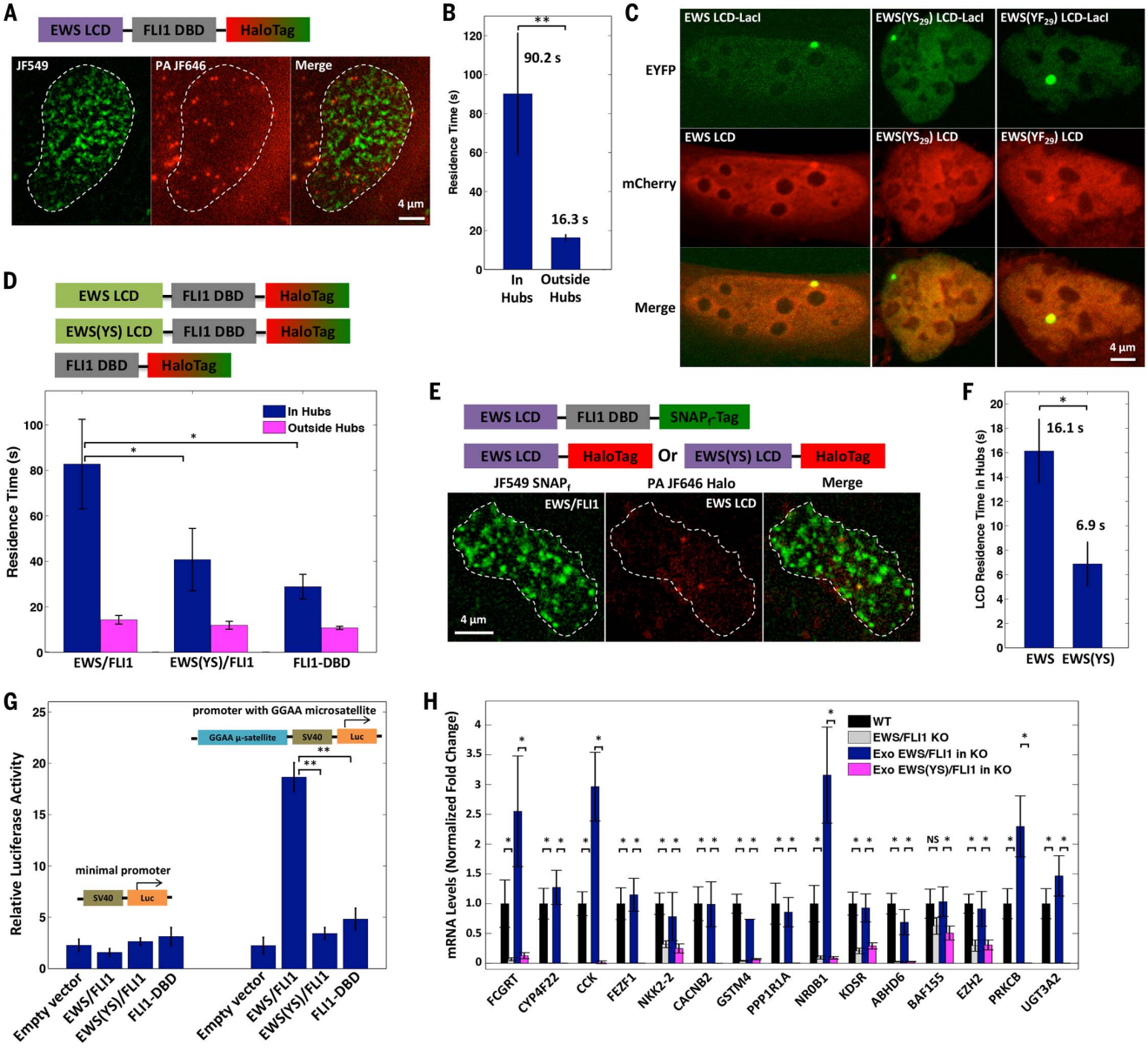
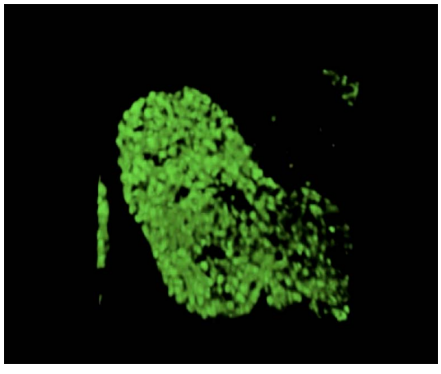


Fig. 5. Dynamic LCD-LCD interactions occur at GGAA microsatellites, which stabilize EWS/FLI1 binding and drive its transactivation function. (**A**) Snapshots of an SPT movie imaging endogenous EWS/FLI1-Halo labeled with two Halo ligands, JF549 (200 nM) and PA-JF646 (20 nM). We imaged the EWS/FLI1-Halo hubs in the JF549 channel (green) and tracked individual EWS/FLI1-Halo molecules in and outside the hubs in the PA-JF646 channel (red). (**B**) Residence times of EWS/FLI1 bound in hubs are longer than its residence times outside hubs, as determined by SPT (**$P < 0.01$, two-sample t test). Error bars represent SE. (**C**) EWS LCD is enriched at LacO array 1 bound by EWS LCD-LacI, but EWS(YS_{29}) LCD is not recruited to the array by EWS(YS_{29}) LCD-LacI. However, EWS(YF_{29}) LCD is recruited to the array by EWS(YF_{29}) LCD-LacI. (**D**) (Top) Schematic of proteins transiently expressed in EWS/FLI1 KO A673 cells: Halo-tagged EWS/FLI1, EWS(YS)/FLI1, or FLI1 DBD. (Bottom) Residence times of EWS/FLI1 and its variants binding in and outside their hubs, as determined by SPT. *$P < 0.05$, two-sample t test. Error bars represent SE. (**E**) Snapshots of an SPT movie simultaneously imaging $SNAP_f$-tagged EWS/FLI1 (200 nM JF549 labeled, green) and Halo-tagged EWS or EWS(YS) LCD (20 nM PA-JF646 labeled, red) in EWS/FLI1 KO A673 cells. Individual LCD-Halo molecules were tracked with the strategy described in (A). (**F**) Residence times of EWS bound at EWS/FLI1 hubs are longer than for EWS(YS) LCD, as determined by SPT (*$P < 0.05$, two-sample t test). Error bars represent SE. (**G**) Luciferase assay shows that EWS/FLI1 but not EWS(YS)/FLI1 or FLI1 DBD transactivates a GGAA microsatellite–driven reporter (**$P < 0.01$, two-sample t test). Error bars represent SE. (**H**) RT-qPCR shows down-regulation of GGAA microsatellite–associated EWS/FLI1 target genes in A673 cells upon EWS/FLI1 KO. Stable expression of exogeneous (Exo) EWS/FLI1, but not of the mutant EWS(YS)/FLI1, rescues the expression defect in EWS/FLI1 KO A673 cells. For each target gene, the mRNA level was normalized using five different invariant genes (fig. S10A) and graphed as a fold change relative to the mRNA level present in the WT A673 line (set to 1). *$P < 0.05$, two-sample t test. NS, not statistically significant. Error bars represent SD.

WT EWS (fig. S9, B and C). By contrast, a mutant replacing all 29 tyrosines with phenylalanine (F) [EWS(YF_{29})] retains hub formation activity with itself and with WT EWS (Fig. 5C and fig. S9D), suggesting that aromatic amino acids and hydrophobic contacts represent major drivers of EWS LCD-LCD interactions. The conformational disorder, proline residues, and acidic amino acids of the EWS LCD might play a role in keeping these key hydrophobic residues exposed to solvent and potential binding partners without being sequestered in a collapsed state (*42*).

Next, we probed the effects of mutations that disrupt LCD hub formation on RTs of EWS/FLI1. To examine behaviors of EWS/FLI1 variants in A673 cells without interference of endogenous EWS/FLI1, we generated an EWS/FLI1 knockout A673 line using CRISPR-Cas9–mediated genome editing (Fig. 4B and fig. S6B) and verified that transiently and moderately reexpressed EWS/FLI1-Halo in the knockout line exhibited binding dynamics comparable to that of endogenous EWS/FLI1-Halo (Fig. 5, B and D, and fig. S8C). We then transiently expressed similar levels of FLI1 DBD-Halo or a Halo-tagged 37-residue Y-to-S mutant [EWS(YS)/FLI1]. Both mutants still displayed some hubs in the nucleus, but the hubs are considerably diminished. SPT revealed that the in-hub RTs of the mutants become significantly reduced (by 51 to 65%) relative to those of WT EWS/FLI1, whereas their outside-hub RTs remain largely unchanged (Fig. 5D). Together, these results confirm that LCD-LCD interactions drive the formation of EWS hubs at GGAA microsatellites.

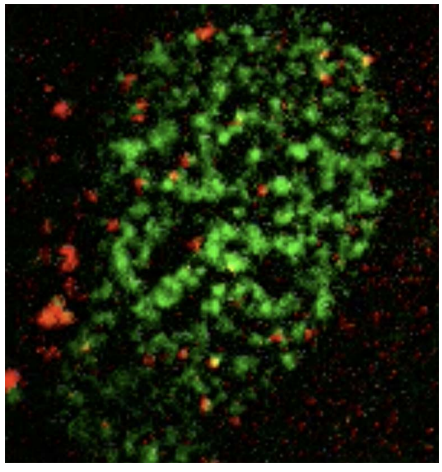
To measure the dynamics of just the protein-protein interactions occurring within the EWS LCD hubs, we transiently expressed $SNAP_f$-tagged EWS/FLI1 and Halo-tagged EWS in the EWS/FLI1 knockout line and labeled both fusion proteins using fluorescent ligands with distinct emission spectra (*40*, *41*). Whereas EWS/FLI1-$SNAP_f$ forms hubs at GGAA microsatellites via protein-DNA binding, EWS LCD-Halo, which does not interact with DNA, binds to EWS/FLI1 hubs only via protein-protein interactions. We visualized EWS/FLI1 hubs and simultaneously tracked individual EWS LCD molecules that bind to the hubs (Fig. 5E). SPT revealed the average RT of EWS LCD in EWS/FLI1 hubs to be 16 s, which suggests that LCD-LCD interactions are highly dynamic (Fig. 5F). As expected, the mutant EWS(YS) LCD has a significantly shorter RT (~7 s) at EWS/FLI1 hubs, consistent with its diminished interaction with the EWS LCD.



Movie 2. 360° view of a 3D image of endogenous EWS/FLI1-Halo in an A673 cell nucleus (stained with 200 nM Halo ligand JF549) taken on the lattice light-sheet microscope (related to Fig. 4C).

EWS LCD-LCD interactions are essential for transcription and transformation functions of EWS/FLI1

Finally, we tested whether LCD-LCD interactions influence EWS/FLI1 functions. We found that whereas EWS/FLI1 efficiently induces gene activation at a GGAA microsatellite in a luciferase assay, the mutant EWS(YS)/FLI1 and FLI1 DBD do not (Fig. 5G). We further engineered the EWS/FLI1 knockout A673 line to stably express EWS/FLI1 or EWS(YS)/FLI1 and performed reverse transcription quantitative polymerase chain reaction (RT-qPCR) to measure the expression levels of GGAA microsatellite–associated EWS/FLI1 target genes. As expected, we found that expression of EWS/FLI1, but not EWS(YS)/FLI1, specifically rescues the gene expression defect in the knockout line, indicating that EWS LCD-LCD interactions are required for transactivation (Fig. 5H and fig. S10A). Moreover, the knockout line stably expressing EWS/FLI1, but not EWS(YS)/FLI1, forms colonies in soft agar like WT A673 cells (fig. S10B). This demonstrates that EWS LCD-LCD interactions are required for oncogenic transformation. Taken together with previously published RT-qPCR and RNA sequencing data in mesenchymal stem cells showing the important role of EWS LCD in inducing expression of GGAA microsatellite–associated genes (*43*), our results suggest that the formation of EWS LCD-dependent hubs is essential for EWS/FLI1 to activate transcription of these target genes and drive oncogenic gene expression programs in Ewing's sarcoma.



Movie 3. Two-color movie imaging endogenous EWS/FLI1-Halo in a knock-in A673 cell nucleus stained with two HaloTag ligands, JF549 (200 nM) and PA-JF646 (20 nM). The image acquisition time was 500 ms. We took time-lapse images with a 10-s interval in the JF549 channel to visualize EWS/FLI1 hubs and imaged continuously in the PA-JF646 channel for SPT of individual EWS/FLI1-Halo molecules (related to Fig. 5A).

Discussion

DNA binding TFs are key regulators of eukaryotic gene expression. Early studies of TFs revealed their well-structured DBDs and identified functionally critical activation domains (ADs) required for transcription. It later became evident that many transactivation sequences contain intrinsically disordered LCDs, but how they mediate transactivation remained unclear. Despite the composition of LCDs and their generally unstructured nature, ADs must interact with specific binding partners to activate transcription. However, it has been challenging if not impossible to directly measure selective LCD-mediated target recognition by ADs in vivo. Two other key aspects of transcriptional control mechanisms have also remained largely uncharted: What is the nature of the protein-protein transactions that drive gene activation, and how stable or dynamic are these critical interactions in living cells?

We addressed these problems by employing single-molecule imaging to visualize LCD-LCD interactions in live cells. Our findings indicate that TF-TF interactions are extremely transient, with RTs of 5 to 20 s and rarely longer than 1 min. These studies lead us to propose that transactivation domains function by forming transient local regions of high TF concentrations or "hubs" (sometimes also referred to as clusters) via dynamic, multivalent, and sequence-specific LCD-LCD interactions. A key future endeavor will be to unlock the biochemical and structural basis for selective LCD-LCD interactions. Such selectivity is likely essential to implement the complex combinatorial logic of transcriptional regulation.

Recent advances in live-cell single-molecule imaging have not only opened a new frontier for studying transcription in vivo but have also provided an opportunity to gain greater insight into the nature and behavior of LCDs and their propensity to remain largely as intrinsically disordered peptide sequences. By deploying high-resolution single-molecule imaging of TF LCD interactions in live cells, our studies offer a powerful complement to pioneering in vitro studies that provided the first clues about LCD interactions (*10*). Importantly, to the extent that one can make comparisons between hydrogels and intracellular LCD hub formation, many aspects of FET-LCD function uncovered in vitro are borne out within live cells. For example, the behavior of LCDs upon mutations, disruption by hexanediols, and interaction with RNA Pol II observed in vivo are generally consistent with the role of LCDs proposed on the basis of in vitro assays. In addition, single-molecule live-cell imaging revealed several new aspects of LCD-driven interactions. Most notable are the fast dynamics and

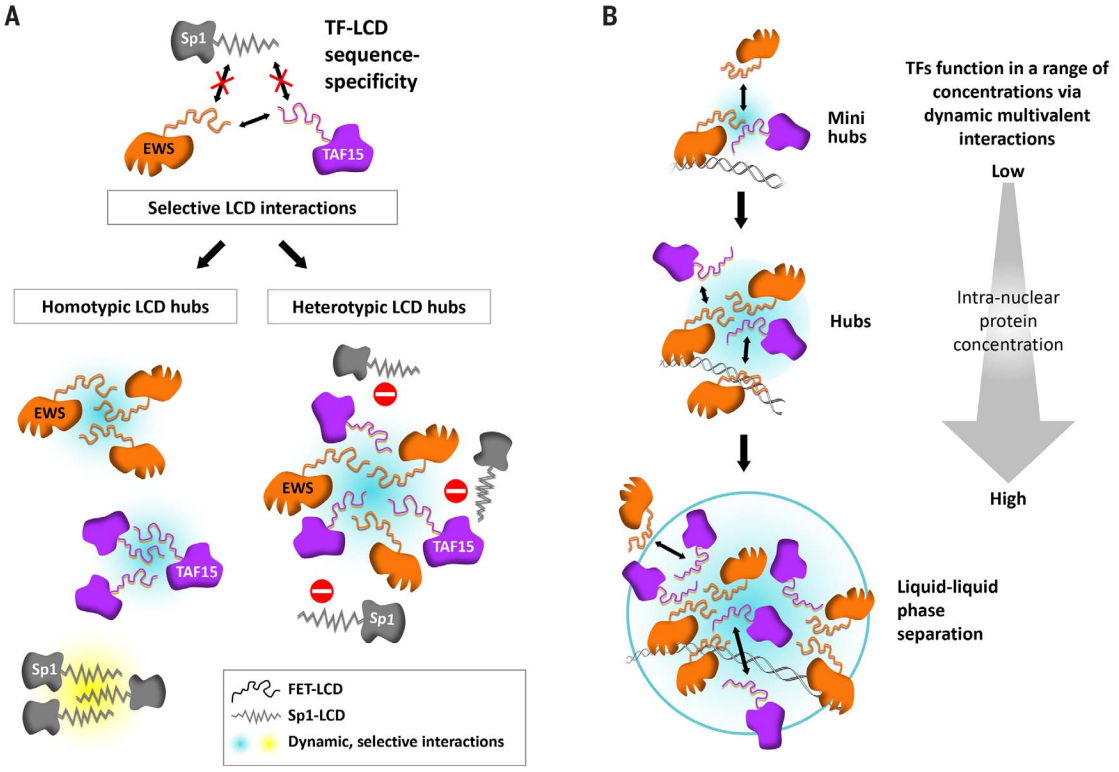


Fig. 6. A model for functional LCD-LCD interactions in vivo: From hubs to phase separation. (**A**) Dynamic and sequence-specific LCD-LCD interactions drive hub formation in live cells. (**B**) LCD-dependent transactivation occurs in hubs formed over a broad range of TF concentrations. At endogenous concentrations, TF LCDs form transactivation hubs at native genomic loci without undergoing evident phase separation. Upon TF LCD overexpression, phase separation is observed at synthetic TF binding site arrays.

sequence specificity of LCD interactions as they form transient local high-concentration hubs that drive transactivation (Fig. 6). We were also intrigued by the formation of LCD-dependent hubs throughout the nucleoplasm that are not associated with cognate genomic DNA. These LCD interaction-driven puncta, which display a capacity to interact with RNA Pol II, sensitivity to 1,6-HD, and fast dynamics (recovery times of 7 to 10 s), may provide an opportunity in future studies to further probe mechanisms governing LCD-LCD–mediated hub formation and transactivation.

To analyze the behavior of LCD-LCD interactions in live cells, we exploited the advantages of various imaging platforms and developed two distinct in vivo cell-based assays: one involving large synthetic TF binding site (LacO) arrays and the other targeting endogenous GGAA microsatellite regulatory elements in the human genome. When used in combination, the two assays afford powerful and complementary platforms for probing the properties of TF LCD interactions in the live-cell context. The highly reiterated LacO binding sites can serve as a useful cell-based assay system capable of nucleating local high-concentration hubs in vivo for any protein domain or LCD of interest through LacO-LacI mediated binding. These arrays can be readily detected by fluorescence imaging with a high signal-to-noise ratio and allow flexibility in probing various interaction properties of LCDs while offering a convenient alternative to various in vitro assays such as hydrogel polymerization and droplet formation for studying gelation and phase separation of intrinsically disordered proteins, two processes that are likely coupled with each other under certain physiological settings (*44*).

Our studies were not designed to address the structure and nature of LCD-driven phase separated compartments, but under certain overexpression conditions we detected what appears to be liquid-liquid phase separation (i.e., spherical shape and local changes in refractive index). Although we can detect apparent liquid-liquid phase separation with gross overexpression of LCDs, we did not obtain evidence for phase separation of TF hubs (i.e., EWS/FLI1) formed at endogenous expression levels. However, transactivation by TF LCD hubs is observed at endogenous TF levels at native chromosomal loci in the absence of detectable phase separation. Given the transience of LCD-LCD interactions and our direct measurements of TF concentrations in the nucleus and within hubs, we surmise that LCD-dependent transactivation can occur in hubs formed over a broad range of TF concentrations (100 nM to 100 μM) and time scales—from extremely brief specific and nonspecific LCD-LCD and TF-DNA binding events (0.1 to 1 s) to assembly of relatively stable hubs (minutes) driven by specific, multivalent interactions. Both the composition and diversity of LCDs in hubs and their interaction specificity could influence the range of their operational concentrations and their potential for phase separation and/or polymer formation. New insights regarding the rapid binding dynamics and functional importance of TF LCDs (i.e., LCD-dependent oncogenic potential of EWS/FLI1) suggest that understanding these mechanisms may also inform our ability to develop strategies to modulate gene expression in the context of disease. Given the transient interactions exhibited by TFs and the critical role of gene dysregulation in disease (i.e., EWS/FLI1 in Ewing's sarcoma), our findings offer the potential to develop single-molecule imaging platforms to screen drugs targeting gene regulatory pathways. In particular, we imagine that moderate- to high-throughput screens for gene expression inhibitors or activators based on high-resolution single-cell and single-molecule imaging could provide a strategy to probe large sectors of the proteome that have resisted traditional in vitro and cell-based screens for small molecules. With new classes of chemicals, natural products, or peptidomimetic libraries, it may even be possible to eventually target LCDs that are the key drivers of regulatory protein-protein interactions and hub formation involved in gene activation and potentially many other biomolecular interactions implicated in disease. Finally, although we examined only a small subset of TF LCDs, the fundamental principles that we have uncovered about the rapid dynamics and hub mechanisms driving LCD-LCD interactions may be applicable to other classes of regulatory proteins and biomolecular interactions occurring in a variety of cell types.

Materials and methods summary

The number of LacO repeats in the LacO arrays was determined by RT-qPCR. FCS and fluorescence intensity measurement were performed on fluorescently labeled TFs in live cells. By comparing the results with standard concentration curves of a purified fluorescent tag, the TF nuclear concentration and its copy number at hubs were

determined. FRAP and SPT were performed to measure interaction dynamics between various TFs and their target genomic loci and to examine how LCD-LCD interactions affect TF-DNA interaction dynamics. SPT was also used to determine LCD-LCD interaction dynamics. Two-color confocal fluorescence imaging was used to examine interactions between different classes of LCDs and between LCD hubs and RNA Pol II.

CRISPR-Cas9–mediated genome editing was performed to label the endogenous EWS/FLI1 in A673 cells with a HaloTag or to knock out the protein, allowing fluorescence imaging or functional studies of EWS/FLI1. Luciferase and soft agar colony formation assays were used to verify the functions of EWS/FLI1-Halo. Lattice light-sheet microscopy was used to visualize intranuclear hubs of the endogenous EWS/FLI1-Halo. Simultaneous confocal imaging of EWS/FLI1-Halo and 3D DNA FISH were performed to examine the spatial relationship between hubs of endogenous EWS/FLI1 and GGAA microsatellites. Luciferase, RT-qPCR, and soft agar colony formation assays were used to examine the effects of Y-to-S mutations on the transactivation and transformation functions of EWS/FLI1. Detailed descriptions for all materials and methods are provided in the supplementary materials.

REFERENCES AND NOTES

1. J. T. Kadonaga, A. J. Courey, J. Ladika, R. Tjian, Distinct regions of Sp1 modulate DNA binding and transcriptional activation. *Science* **242**, 1566–1570 (1988). doi: 10.1126/science.3059495; pmid: 3059495
2. A. J. Courey, R. Tjian, Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* **55**, 887–898 (1988). doi: 10.1016/0092-8674(88)90144-4; pmid: 3142690
3. W. Su, S. Jackson, R. Tjian, H. Echols, DNA looping between sites for transcriptional activation: Self-association of DNA-bound Sp1. *Genes Dev.* **5**, 820–826 (1991). doi: 10.1101/gad.5.5.820; pmid: 1851121
4. M. J. Friedman *et al*., Polyglutamine domain modulates the TBP-TFIIB interaction: Implications for its normal function and neurodegeneration. *Nat. Neurosci.* **10**, 1519–1528 (2007). doi: 10.1038/nn2011; pmid: 17994014
5. H. Kovar, Dr. Jekyll and Mr. Hyde: The Two Faces of the FUS/EWS/TAF15 Protein Family. *Sarcoma* **2011**, 837474 (2011). doi: 10.1155/2011/837474; pmid: 21197473
6. M. Kato *et al*., Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012). doi: 10.1016/j.cell.2012.04.017; pmid: 22579281
7. A. Patel *et al*., A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **162**, 1066–1077 (2015). doi: 10.1016/j.cell.2015.07.047; pmid: 26317470
8. K. A. Burke, A. M. Janke, C. L. Rhine, N. L. Fawzi, Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Mol. Cell* **60**, 231–241 (2015). doi: 10.1016/j.molcel.2015.09.006; pmid: 26455390
9. M. Boehning *et al*., RNA polymerase II clustering through CTD phase separation. bioRxiv 316372 [Preprint]. 7 May 2018. doi: 10.1101/316372
10. I. Kwon *et al*., Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell* **155**, 1049–1060 (2013). doi: 10.1016/j.cell.2013.10.033; pmid: 24267890
11. M. Altmeyer *et al*., Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nat. Commun.* **6**, 8088 (2015). doi: 10.1038/ncomms9088; pmid: 26286827
12. S. Xiang *et al*., The LC Domain of hnRNPA2 Adopts Similar Conformations in Hydrogel Polymers, Liquid-like Droplets, and Nuclei. *Cell* **163**, 829–839 (2015). doi: 10.1016/j.cell.2015.10.040; pmid: 26544936
13. D. T. Murray *et al*., Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell* **171**, 615–627 (2017). doi: 10.1016/j.cell.2017.08.048; pmid: 28942918
14. Y. Lin, D. S. Protter, M. K. Rosen, R. Parker, Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol. Cell* **60**, 208–219 (2015). doi: 10.1016/j.molcel.2015.08.018; pmid: 26412307
15. S. Elbaum-Garfinkle *et al*., The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7189–7194 (2015). doi: 10.1073/pnas.1504822112; pmid: 26015579
16. T. J. Nott *et al*., Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015). doi: 10.1016/j.molcel.2015.01.013; pmid: 25747659
17. A. Molliex *et al*., Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015). doi: 10.1016/j.cell.2015.09.015; pmid: 26406374
18. E. L. Elson, D. Magde, Fluorescence correlation spectroscopy. I. Conceptual basis and theory. *Biopolymers* **13**, 1–27 (1974). doi: 10.1002/bip.1974.360130102
19. R. T. Youker, H. Teng, Measuring protein dynamics in live cells: Protocols and practical considerations for fluorescence fluctuation microscopy. *J. Biomed. Opt.* **19**, 090801 (2014). doi: 10.1117/1.JBO.19.9.090801; pmid: 25260867
20. B. L. Sprague, J. G. McNally, FRAP analysis of binding: Proper and fitting. *Trends Cell Biol.* **15**, 84–91 (2005). doi: 10.1016/j.tcb.2004.12.001; pmid: 15695095
21. B. C. Chen *et al*., Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *Science* **346**, 1257998 (2014). doi: 10.1126/science.1257998; pmid: 25342811
22. I. Solovei, M. Cremer, 3D-FISH on cultured cells combined with immunostaining. *Methods Mol. Biol.* **659**, 117–126 (2010). doi: 10.1007/978-1-60761-789-1_8; pmid: 20809307
23. J. Chen *et al*., Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell* **156**, 1274–1285 (2014). doi: 10.1016/j.cell.2014.01.062; pmid: 24630727
24. A. S. Hansen, I. Pustova, C. Cattoglio, R. Tjian, X. Darzacq, CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife* **6**, e25776 (2017). doi: 10.7554/eLife.25776; pmid: 28467304
25. S. M. Janicki *et al*., From silencing to gene expression: Real-time analysis in single cells. *Cell* **116**, 683–698 (2004). doi: 10.1016/S0092-8674(04)00171-0; pmid: 15006351
26. B. L. Sprague, R. L. Pego, D. A. Stavreva, J. G. McNally, Analysis of binding reactions by fluorescence recovery after photobleaching. *Biophys. J.* **86**, 3473–3495 (2004). doi: 10.1529/biophysj.103.026765; pmid: 15189848
27. X. Darzacq *et al*., In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **14**, 796–806 (2007). doi: 10.1038/nsmb1280; pmid: 17676063
28. Y. Lin *et al*., Toxic PR Poly-Dipeptides Encoded by the C9orf72 Repeat Expansion Target LC Domain Polymers. *Cell* **167**, 789–802 (2016). doi: 10.1016/j.cell.2016.10.003; pmid: 27768897
29. A. R. Strom *et al*., Phase separation drives heterochromatin domain formation. *Nature* **547**, 241–245 (2017). doi: 10.1038/nature22989; pmid: 28636597
30. O. Rog, S. Köhler, A. F. Dernburg, The synaptonemal complex has liquid crystalline properties and spatially regulates meiotic recombination factors. *eLife* **6**, e21455 (2017). doi: 10.7554/eLife.21455; pmid: 28045371
31. G. Gill, E. Pascal, Z. H. Tseng, R. Tjian, A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 192–196 (1994). doi: 10.1073/pnas.91.1.192; pmid: 8278363
32. A. Martínez-Ramírez *et al*., Characterization of the A673 cell line (Ewing tumor) by molecular cytogenetic techniques. *Cancer Genet. Cytogenet.* **141**, 138–142 (2003). doi: 10.1016/S0165-4608(02)00670-2; pmid: 12606131
33. K. Gangwal *et al*., Microsatellites as EWS/FLI response elements in Ewing's sarcoma. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10149–10154 (2008). doi: 10.1073/pnas.0801073105; pmid: 18626011
34. S. L. Lessnick, M. Ladanyi, Molecular pathogenesis of Ewing sarcoma: New therapeutic and transcriptional targets. *Annu. Rev. Pathol.* **7**, 145–159 (2012). doi: 10.1146/annurev-pathol-011110-130237; pmid: 21942527
35. K. M. Johnson *et al*., Role for the EWS domain of EWS/FLI in binding GGAA-microsatellites required for Ewing sarcoma anchorage independent growth. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 9870–9875 (2017). doi: 10.1073/pnas.1701872114; pmid: 28847958
36. F. A. Ran *et al*., Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013). doi: 10.1038/nprot.2013.143; pmid: 24157548
37. S. L. Lessnick, C. S. Dacwag, T. R. Golub, The Ewing's sarcoma oncoprotein EWS/FLI induces a p53-dependent growth arrest in primary human fibroblasts. *Cancer Cell* **1**, 393–401 (2002). doi: 10.1016/S1535-6108(02)00056-9; pmid: 12086853
38. K. M. Johnson, C. Taslim, R. S. Saund, S. L. Lessnick, Identification of two types of GGAA-microsatellites and their roles in EWS/FLI binding and gene regulation in Ewing sarcoma. *PLOS ONE* **12**, e0186275 (2017). doi: 10.1371/journal.pone.0186275; pmid: 29091716
39. N. Guillon *et al*., The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *PLOS ONE* **4**, e4932 (2009). doi: 10.1371/journal.pone.0004932; pmid: 19305498
40. J. B. Grimm *et al*., A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nat. Methods* **12**, 244–250 (2015). doi: 10.1038/nmeth.3256; pmid: 25599551
41. J. B. Grimm *et al*., Bright photoactivatable fluorophores for single-molecule imaging. *Nat. Methods* **13**, 985–988 (2016). doi: 10.1038/nmeth.4034; pmid: 27776112
42. M. V. Staller *et al*., A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst.* **6**, 444–455 (2018). doi: 10.1016/j.cels.2018.01.015; pmid: 29525204
43. G. Boulay *et al*., Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell* **171**, 163–178 (2017). doi: 10.1016/j.cell.2017.07.036; pmid: 28844694
44. T. S. Harmon, A. S. Holehouse, M. K. Rosen, R. V. Pappu, Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife* **6**, e30294 (2017). doi: 10.7554/eLife.30294; pmid: 29091028
45. B. Efron, R. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, CRC Press, 1993).

ACKNOWLEDGMENTS

We thank S. Lessnick, S. McKnight, and M. Kato for discussions and providing reagents; Q. Gan and A. Hansen for providing codes to analyze imaging data; J. Bosco and P. Sharma for help with molecular cloning; K. Heydari and the CRL Flow Cytometry Facility for assistance with flow cytometry; the CRL Molecular Imaging Center for providing access to confocal fluorescence microscopes; and J. Goodrich, G. S. Martin, and members of Tjian and Darzacq labs for critical reading of the manuscript. **Funding:** This work was supported by California Institute of Regenerative Medicine grant LA1-08013 (to X.D.), NIH grants UO1-EB021236 and U54-DK107980 (to X.D.), and the Howard Hughes Medical Institute (to Z.L., L.L., and R.T.). **Author contributions:** R.T. conceived of and supervised the project. S.C., R.T., and X.D. designed experiments and interpreted the data. S.C. performed genome editing, soft colony formation assays, live-cell imaging in the presence of hexanediols, confocal imaging, DNA FISH, immunofluorescence, FRAP and SPT measurements, and all of the imaging data analyses. C.D.-D. established and characterized Halo-RPB1 cell lines, characterized LacO-containing cell lines, and performed all of the RT-qPCR assays and related data analyses. S.C. and Z.L. developed codes for SPT data analyses. P.D. performed lattice light-sheet imaging. G.M.D. and S.C. designed and generated constructs. S.C. and C.C. performed luciferase assays. A.H. performed FCS measurements. S.B. and L.L. synthesized and characterized fluorescent HaloTag and $SNAP_f$-tag ligands. S.C., R.T., C.D.-D., C.C., Z.L., X.D., P.D., and A.H. prepared the manuscript and supplementary materials. **Competing interests:** L.L. has filed a patent and patent applications (e.g., U.S. Patent 9,933,417) covering azetidine-containing fluorophores such as JF549. **Data and materials availability:** All codes for imaging data analyses have been submitted to GitHub and are publicly available (https://github.com/Shasha-Chong/CodeFor2018SciencePaper).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/eaar2555/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 to S5
References (*46*–*55*)

23 October 2017; resubmitted 12 March 2018
Accepted 13 June 2018
Published online 21 June 2018
10.1126/science.aar2555

RESEARCH ARTICLE SUMMARY

MOLECULAR BIOLOGY

Coactivator condensation at super-enhancers links phase separation and gene control

Benjamin R. Sabari*, Alessandra Dall'Agnese*, Ann Boija, Isaac A. Klein, Eliot L. Coffey, Krishna Shrinivas, Brian J. Abraham, Nancy M. Hannett, Alicia V. Zamudio, John C. Manteiga, Charles H. Li, Yang E. Guo, Daniel S. Day, Jurian Schuijers, Eliza Vasile, Sohail Malik, Denes Hnisz, Tong Ihn Lee, Ibrahim I. Cisse, Robert G. Roeder, Phillip A. Sharp, Arup K. Chakraborty, Richard A. Young†

INTRODUCTION: Mammalian genes that play prominent roles in healthy and diseased cellular states are often controlled by special DNA elements called super-enhancers (SEs). SEs are clusters of enhancers that are occupied by an unusually high density of interacting factors and drive higher levels of transcription than most typical enhancers. This high-density assembly at SEs has been shown to exhibit sharp transitions of formation and dissolution, forming in a single nucleation event and collapsing when chromatin factors or nucleation sites are deleted. These features led us to postulate that SEs are phase-separated multimolecular assemblies, also known as biomolecular condensates. Phase-separated condensates, such as the nucleolus and other membraneless cellular bodies, provide a means to compartmentalize and concentrate biochemical reactions within cells.

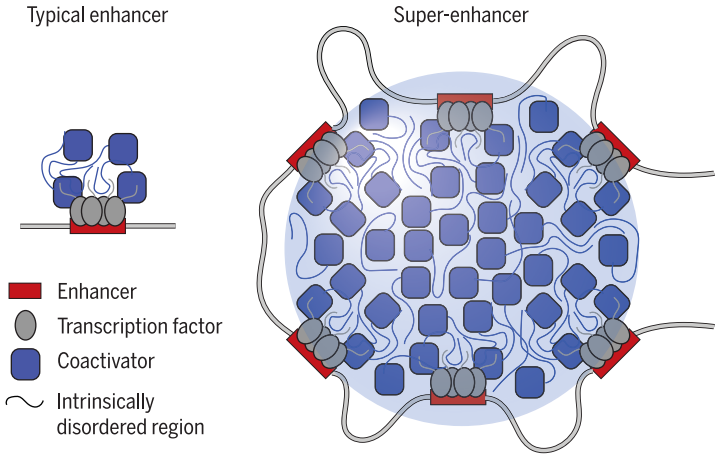
RATIONALE: SEs are formed by the binding of master transcription factors (TFs) at each component enhancer, and these recruit unusually high densities of coactivators, including Mediator and BRD4. Mediator is a large (~1.2 MDa) multisubunit complex that has multiple roles in transcription, including bridging interactions between TFs and RNA polymerase II (RNA Pol II). BRD4 facilitates the release of RNA Pol II molecules from the site of transcription initiation. The presence of MED1, a subunit of Mediator, and BRD4 can be used to define SEs. We reasoned that if transcriptional condensates are formed at SEs, then MED1 and BRD4 should be visualized as discrete bodies at SE elements in cell nuclei. These bodies should exhibit behaviors described for liquid-like condensates. We investigated these possibilities by using murine embryonic stem cells (mESCs), in which SEs were originally described. Because intrinsically disordered regions (IDRs) of proteins have been implicated in condensate formation, we postulated that the large IDRs present in MED1 and BRD4 might be involved.

RESULTS: We found that MED1 and BRD4 occupy discrete nuclear bodies that occur at SEs in mESCs. These bodies exhibit properties of other well-studied biomolecular condensates, including rapid recovery of fluorescence after photobleaching and sensitivity to 1,6-hexanediol, which disrupts liquid-like condensates. Disruption of MED1 and BRD4 bodies by 1,6-hexanediol was accompanied by a loss of chromatin-bound MED1 and BRD4 at SEs, as well as a loss of RNA Pol II at SEs and SE-driven genes. The IDRs of both MED1 and BRD4 formed phase-separated liquid droplets in vitro, and these droplets exhibited features characteristic of condensates formed by networks of weak protein-protein interactions. The MED1-IDR droplets were found to concentrate BRD4 and RNA Pol II from transcriptionally competent nuclear extracts, which may reflect their contribution to compartmentalizing and concentrating biochemical reactions associated with transcription at SEs in cells.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aar3958

CONCLUSION: Our results show that coactivators form phase-separated condensates at SEs and that SE condensates compartmentalize and concentrate the transcription apparatus at key cell-identity genes. These results have implications for the mechanisms involved in the control of genes in healthy and diseased cell states. We suggest that SE condensates facilitate the compartmentalization and concentration of transcriptional components at specific genes through the phase-separating properties of IDRs in TFs and cofactors. SE condensates may thus ensure robust transcription of genes essential to cell-identity maintenance. These properties may also explain why cancer cells acquire large SEs at driver oncogenes and why SEs that facilitate transcriptional dysregulation in disease can be especially sensitive to transcriptional inhibitors. ■



Phase separation of coactivators compartmentalizes and concentrates the transcription apparatus. Enhancers are gene regulatory elements bound by transcription factors that recruit coactivators and the transcription apparatus (not shown) to regulate gene expression. Super-enhancers are clusters of enhancers bound by master transcription factors that concentrate high densities of coactivators and the transcription apparatus to drive robust expression of genes that play prominent roles in cell identity. This is achieved by the phase separation of coactivators, which is driven in part by high-valency and low-affinity interactions of intrinsically disordered regions.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: young@wi.mit.edu
Cite this article as B. R. Sabari *et al.*, *Science* **361**, eaar3958 (2018). DOI: 10.1126/science.aar3958

RESEARCH ARTICLE

MOLECULAR BIOLOGY

Coactivator condensation at super-enhancers links phase separation and gene control

Benjamin R. Sabari[1]*, Alessandra Dall'Agnese[1]*, Ann Boija[1], Isaac A. Klein[1,2], Eliot L. Coffey[1,3], Krishna Shrinivas[4,5], Brian J. Abraham[1], Nancy M. Hannett[1], Alicia V. Zamudio[1,3], John C. Manteiga[1,3], Charles H. Li[1,3], Yang E. Guo[1], Daniel S. Day[1], Jurian Schuijers[1], Eliza Vasile[6], Sohail Malik[7], Denes Hnisz[1], Tong Ihn Lee[1], Ibrahim I. Cisse[8], Robert G. Roeder[7], Phillip A. Sharp[3,6], Arup K. Chakraborty[4,5,8,9,10,11], Richard A. Young[1,3]†

Super-enhancers (SEs) are clusters of enhancers that cooperatively assemble a high density of the transcriptional apparatus to drive robust expression of genes with prominent roles in cell identity. Here we demonstrate that the SE-enriched transcriptional coactivators BRD4 and MED1 form nuclear puncta at SEs that exhibit properties of liquid-like condensates and are disrupted by chemicals that perturb condensates. The intrinsically disordered regions (IDRs) of BRD4 and MED1 can form phase-separated droplets, and MED1-IDR droplets can compartmentalize and concentrate the transcription apparatus from nuclear extracts. These results support the idea that coactivators form phase-separated condensates at SEs that compartmentalize and concentrate the transcription apparatus, suggest a role for coactivator IDRs in this process, and offer insights into mechanisms involved in the control of key cell-identity genes.

Phase separation of fluids is a physicochemical process by which molecules separate into a dense phase and a dilute phase. Phase-separated biomolecular condensates, which include the nucleolus, nuclear speckles, stress granules, and others, provide a mechanism to compartmentalize and concentrate biochemical reactions within cells (*1*–*3*). Biomolecular condensates produced by liquid-liquid phase separation allow rapid movement of components into and within the dense phase and exhibit properties of liquid droplets such as fusion and fission (*4*). Dynamic and cooperative multivalent interactions among molecules, such as those produced by certain intrinsically disordered regions (IDRs) of proteins, have been implicated in liquid-liquid phase separation (*5*–*7*).

Enhancers are gene regulatory elements bound by transcription factors (TFs) and other components of the transcription apparatus that function to regulate expression of cell type–specific genes (*8*–*13*). Super-enhancers (SEs)—clusters of enhancers that are occupied by exceptionally high densities of transcriptional machinery—regulate genes with especially important roles in cell identity (*14*, *15*). DNA interaction data show that enhancer elements in the clusters are in close spatial proximity with each other and the promoters of the genes that they regulate (*16*–*18*), consistent with the notion of a dense assembly of transcriptional machinery at these sites. This high-density assembly at SEs has been shown to exhibit sharp transitions of formation and dissolution, forming as the consequence of a single nucleation event (*19*, *20*) and collapsing when concentrated factors are depleted from chromatin (*21*–*25*) or when nucleation sites are deleted (*26*–*29*). These properties of SEs led to the proposal that the high-density assembly of biomolecules at active SEs is due to phase separation of enriched factors at these genetic elements (*30*). Here we provide experimental evidence that the transcriptional coactivators BRD4 and MED1 (a subunit of the Mediator complex) form condensates at SEs. This establishes a new framework to account for the diverse properties described for these regulatory elements and expands the known biochemical processes regulated by phase separation to include the control of cell-identity genes.

[1]Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, MA 02142, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [5]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [7]Laboratory of Biochemistry and Molecular Biology, The Rockefeller University, New York, NY 10065, USA. [8]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [9]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [10]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [11]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard, Cambridge, MA 02139, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: young@wi.mit.edu

BRD4 and MED1 coactivators form nuclear puncta

The enhancer clusters that make up SEs are occupied by master TFs and unusually high densities of factors, including BRD4 and MED1, that are coactivators (*31*–*35*) whose presence can be used to define SEs (*14*, *15*, *21*). We reasoned that if BRD4 and MED1 are components of nuclear condensates, then they might be visualized as discrete puncta in the nuclei of cells, and the properties of these puncta could be investigated. Fixed cell immunofluorescence (IF) with antibodies against BRD4 and MED1 in murine embryonic stem cells (mESCs) revealed nuclear puncta for both factors (Fig. 1A). To determine whether such puncta occur in live cells, we engineered mESCs by using CRISPR-Cas9 to tag endogenous BRD4 and MED1 with monomeric enhanced green fluorescent protein (mEGFP) (fig. S1). Live-cell fluorescence microscopy of the engineered mESC lines also revealed discrete nuclear puncta (Fig. 1B). Analysis of these images showed that there were 1034 ± 130 BRD4 and 983 ± 102 MED1 puncta per nucleus (means ± SEM) (table S1). These results demonstrate that BRD4 and MED1 are components of puncta within the nuclei of mESCs.

SEs are associated with coactivator puncta

Several lines of evidence suggest that SEs are likely to be associated with some of the BRD4 and MED1 puncta in mESCs. ChIP-seq (chromatin immunoprecipitation followed by sequencing) data for BRD4 and MED1 show that SEs are especially enriched in these coactivators (*14*, *15*). DNA interaction data suggest that SE constituents occupied by BRD4 and MED1 are in close spatial proximity to one another (Fig. 1C and fig. S2A). Co-occupancy of the genome by BRD4 and MED1 is most evident at SEs (fig. S2B) (*14*, *15*). To determine whether SEs are associated with some of the BRD4 and MED1 puncta, we performed IF for BRD4 or MED1 together with DNA-FISH or nascent RNA-FISH for the genomic region containing the *Nanog* gene and its SEs (FISH, fluorescence in situ hybridization) (Fig. 1, D to G). We found that BRD4 and MED1 puncta consistently overlapped the DNA-FISH foci (Fig. 1D) or RNA-FISH foci (Fig. 1F). An average image analysis (details are given in the methods) of the BRD4 and MED1 IF signals centered at DNA-FISH foci (n = 137 for BRD4 and 125 for MED1) and RNA-FISH foci (n = 121 for BRD4 and 181 for MED1) revealed that, on average, BRD4 and MED1 fluorescence intensities are most enriched at the center of FISH foci (Fig. 1, E and G); this trend was not observed for average images centered at randomly selected nuclear positions (Fig. 1, E and G). Radial distribution functions of the averaged images for FISH and IF pairs show a significant correlation (Spearman correlation

coefficients > 0.6; P values < 1×10^{-16}), with both BRD4 and MED1 having their highest signal intensities at the center of the FISH focus; signals decay with distance from this center (fig. S3). The radial distributions of FISH and IF at randomly selected nuclear positions are not correlated (Spearman correlation coefficients < 0.2) (fig. S3). Similar results were obtained when we performed IF for BRD4 or MED1 together with nascent RNA-FISH for the SE-regulated genes *Klf4*, *Mir290*, and *Trim28* (figs. S3 and S4, A to F). When a similar experiment was conducted for two genes expressed in mESCs but not associated with a SE (*Fam168b* and *Zfp606*), there was no evident overlap between FISH foci

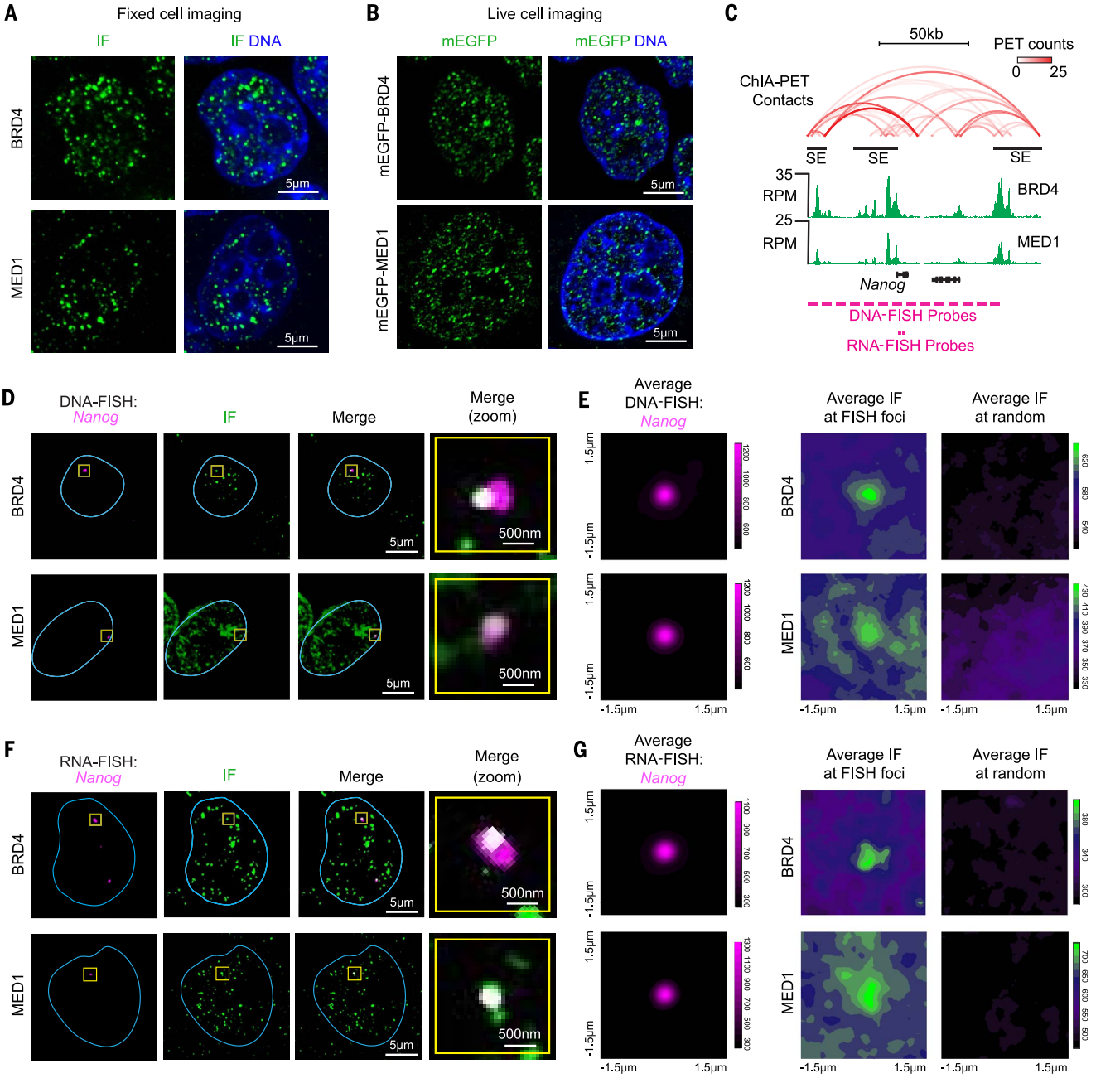


Fig. 1. BRD4 and MED1 form puncta at super-enhancers (SEs). (**A**) Immunofluorescence (IF) imaging of BRD4 and MED1 in mouse embryonic stem cells (mESCs). Fluorescence signal is shown alone (left) and merged with Hoechst stain (right). (**B**) Live imaging of endogenously tagged mEGFP-BRD4 and mEGFP-MED1 in mESCs. (**C**) Depiction of *Nanog* locus, associated SEs (black bars), DNA contacts (red arcs), BRD4 and MED1 ChIP-seq (green histograms), and location of FISH probes. ChIA-PET, chromatin interaction analysis with paired-end tag; RPM, reads per million. (**D**) Colocalization between BRD4 or MED1 and the *Nanog* locus by IF and DNA-FISH in fixed mESCs. Separate images of the indicated IF and FISH are shown, along with an image showing the merged channels (overlapping signal in white). The blue line highlights the nuclear periphery, determined by Hoechst staining (not shown). The rightmost column shows the area in the yellow box in greater detail. (**E**) Averaged signal of (left) DNA-FISH for *Nanog* and (right) IF for BRD4 or MED1 centered at *Nanog* DNA-FISH foci or randomly selected nuclear positions. (**F**) Colocalization between BRD4 or MED1 and the nascent RNA of *Nanog*, determined by IF and RNA-FISH in fixed mESCs. Data are shown as in (D). (**G**) Averaged signal of (left) RNA-FISH for *Nanog* and (right) IF for BRD4 or MED1 centered at *Nanog* RNA-FISH foci or randomly selected nuclear positions.

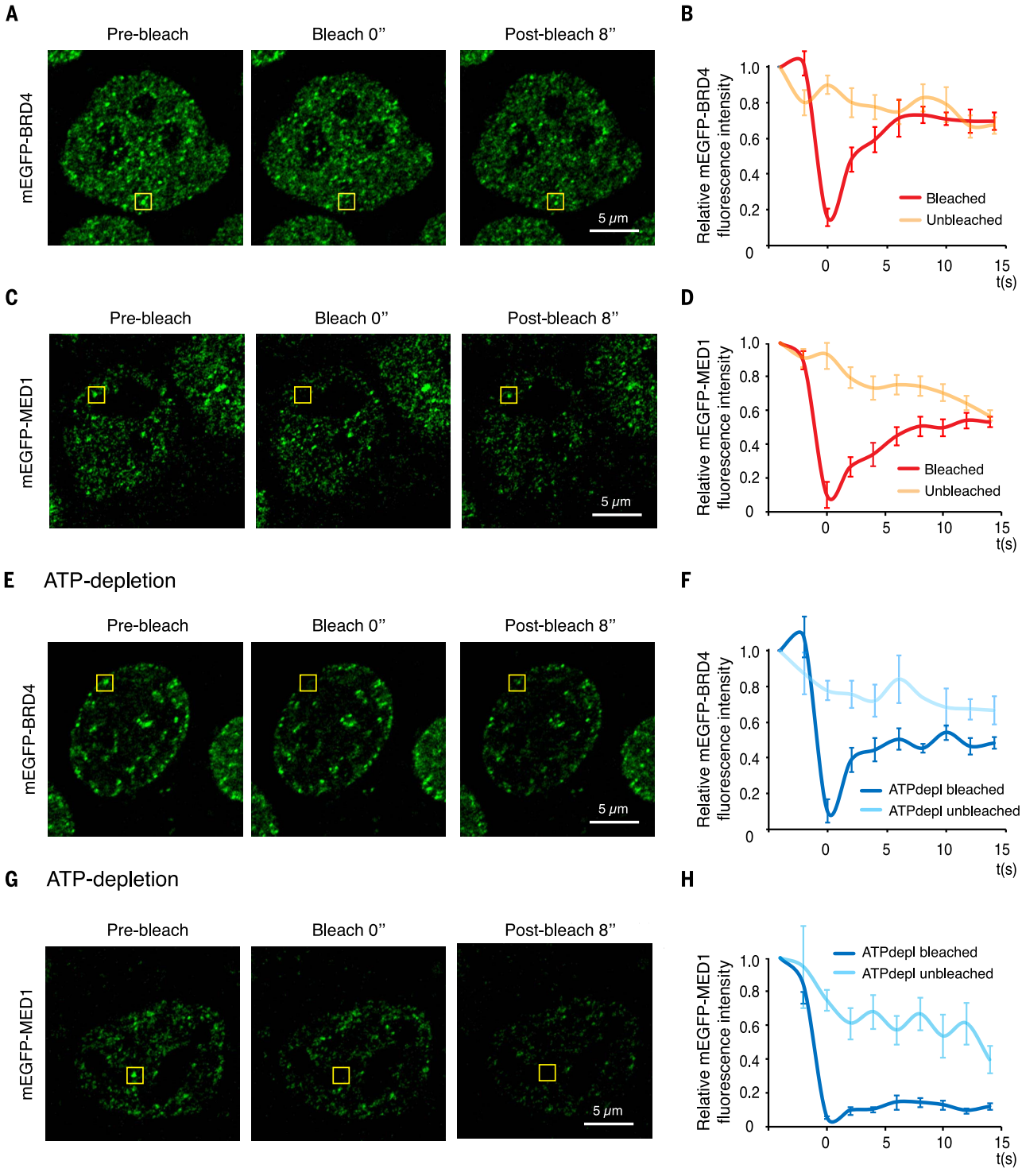


Fig. 2. BRD4 and MED1 nuclear puncta exhibit properties expected for biomolecular condensates. (**A**) Representative images of the FRAP experiment with mEGFP-BRD4–engineered mESCs (where " indicates time in seconds). The yellow box highlights the punctum undergoing targeted bleaching. (**B**) Quantification of FRAP data for mEGFP-BRD4 puncta. The bleaching event occurs at t = 0 s. For the bleached area and the unbleached control, background-subtracted fluorescence intensities are plotted relative to a prebleach time point (t = −4 s). Data are plotted as means ± SEM (n = 9). (**C**) Same as (A), but with mEGFP-MED1–engineered mESCs. (**D**) Same as (B), but for mEGFP-MED1 puncta (n = 9). (**E**) Representative images of the FRAP experiment with mEGFP-BRD4–engineered mESCs upon ATP depletion. (**F**) Quantification of FRAP data for mEGFP-BRD4 puncta upon ATP depletion (n = 8), as in (B). (**G**) Representative images of the FRAP experiment with mEGFP-MED1–engineered mESCs upon ATP depletion. (**H**) Quantification of FRAP data for mEGFP-MED1 puncta upon ATP depletion (n = 8), as in (B). Images were taken using the Zeiss LSM 880 confocal microscope with an Airyscan detector and a 63× objective at 37°C.

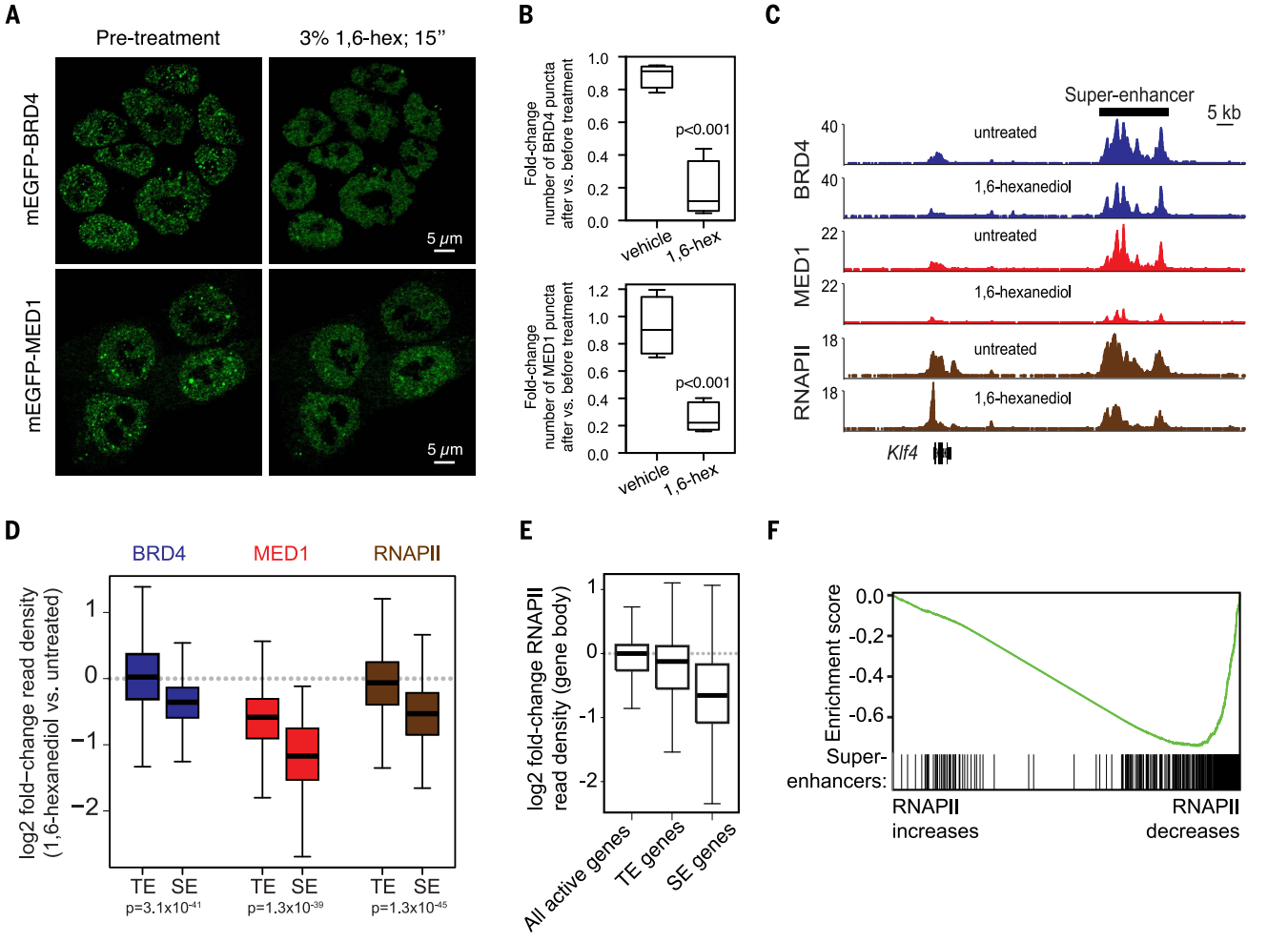


Fig. 3. 1,6-hexanediol disrupts BRD4 and MED1 puncta and disrupts BRD4, MED1, and RNA Pol II occupancy at SEs and SE-driven genes. (**A**) Representative images of mEGFP-BRD4– or mEGP-MED1–engineered mESCs before and after treatment with 3% hexanediol for 15 s. (**B**) Box-plot representation of the fold change in the number of mEGFP-BRD4 or mEGFP-MED1 puncta observed after addition of vehicle or 1,6-hexanediol to a final concentration of 3%. (**C**) Genome browser view of BRD4 (blue), MED1 (red), and RNA Pol II (RNAPII, brown) ChIP-seq data from untreated or 1,6-hexanediol–treated (1.5% for 30 min) mESCs at the *Klf4* locus. The *y* axis shows reads per million. (**D**) Box-plot representation of the log2 fold change in BRD4 (blue), MED1 (red), and RNA Pol II (brown) ChIP-seq read density (1,6-hexanediol–treated versus untreated) for regions defined as SEs or typical enhancers (TEs) (methods and table S2). (**E**) Box-plot representation of the log2 fold change in RNA Pol II ChIP-seq density (1,6-hexanediol–treated versus untreated) within the gene body (transcription start site to transcription end site) of all active genes (reads per kilobase per million > 1), TE-associated genes, or SE-associated genes. (**F**) Gene set enrichment analysis, with genes ranked by their log2 fold change in RNA Pol II ChIP-seq density within the gene body and annotated against the set of SE-associated genes. Enrichment score profile and the position of SE-associated genes are shown.

and BRD4 puncta (fig. S4G). These results indicate that both BRD4 and MED1 puncta are present at SEs.

Coactivator puncta exhibit liquid-like rates of fluorescence recovery after photobleaching

We next sought to examine whether BRD4 and MED1 puncta exhibit features characteristic of liquid-like condensates. A hallmark of liquid-like condensates is internal dynamical reorganization and rapid exchange kinetics (*1*–*3*), which can be interrogated by measuring the rate of fluorescence recovery after photobleaching (FRAP). To study the dynamics of BRD4 and MED1 foci in live cells, we performed FRAP experiments on endogenously tagged mEGFP-BRD4 or mEGFP-MED1 cell lines. After photobleaching, mEGFP-BRD4 and mEGFP-MED1 puncta recovered fluorescence on a time scale of seconds (Fig. 2, A to D), with apparent diffusion coefficients of ~0.37 ± 0.13 and ~0.14 ± 0.04 $\mu m^2/s$, respectively. These values are similar to those previously reported for components of liquid-like condensates (*36*, *37*). Adenosine triphosphate (ATP) has been implicated in promoting condensate fluidity by driving energy-dependent processes and/or through its intrinsic hydrotrope activity (*38*, *39*). Depletion of cellular ATP by glucose deprivation and oligomycin treatment altered fluorescence recovery after photobleaching for both mEGFP-BRD4 and mEGFP-MED1 foci; the rate of recovery for MED1 was reduced, and the extent of recovery for BRD4 was diminished (Fig. 2, E to H). These results indicate that puncta containing BRD4 and MED1 have liquid-like properties in cells, consistent with previously described phase-separated condensates.

Coactivator puncta and SE occupancy are sensitive to condensate perturbation

To further investigate the biophysical properties of BRD4 and MED1 puncta, we investigated their sensitivity to 1,6-hexanediol, a compound known to disrupt liquid-like condensates, possibly by disruption of hydrophobic interactions (*40*). We found that treatment of mESCs expressing endogenously tagged mEGFP-BRD4 or mEGFP-MED1

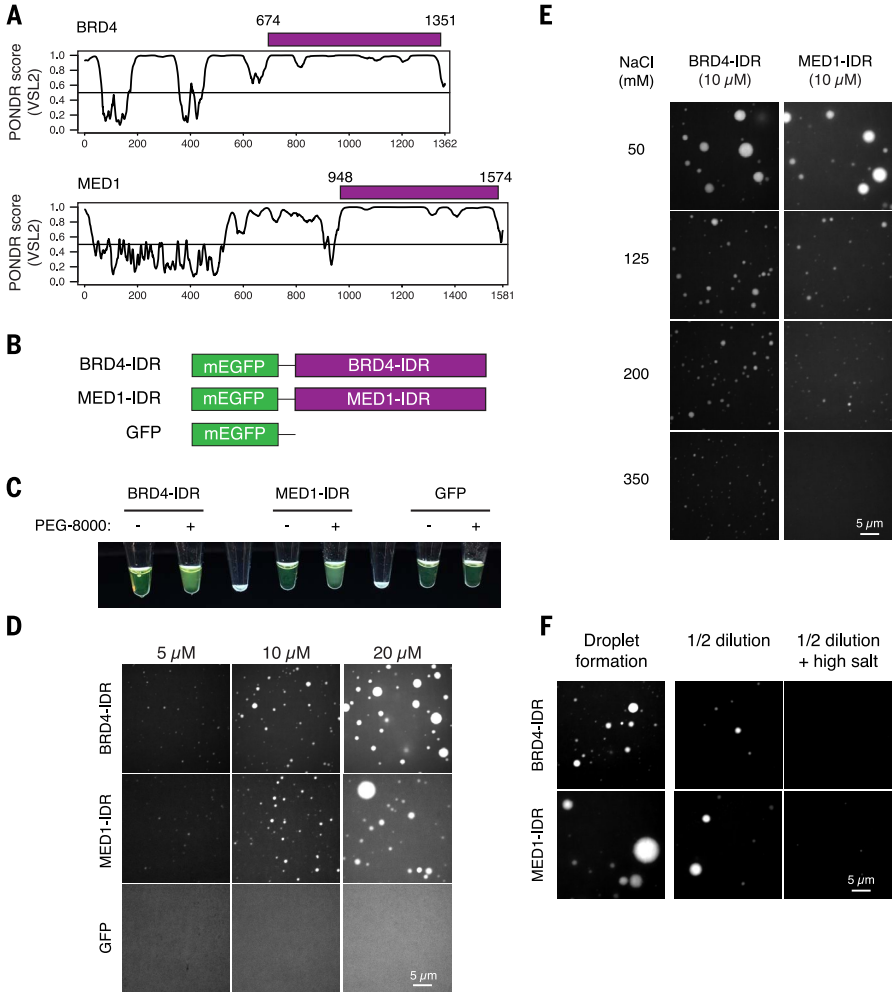


Fig. 4. Intrinsically disordered regions (IDRs) of BRD4 and MED1 phase-separate in vitro. (**A**) Graphs plotting intrinsic disorder for BRD4 and MED1. PONDR (Predictor of Natural Disordered Regions) VSL2 scores are shown on the *y* axis, and amino acid positions are shown on the *x* axis. The purple bar designates the IDR under investigation. (**B**) Schematic of recombinant mEGFP fusion proteins used in this study. Purple boxes indicate the IDRs of BRD4 and MED1 shown in (A). (**C**) Visualization of turbidity associated with droplet formation. Tubes containing BRD4-IDR (left pair), MED1-IDR (middle pair), or GFP (right pair) in the presence (+) or absence (−) of PEG-8000 are shown. Blank tubes are included between pairs for contrast. (**D**) Representative images of droplet formation at different protein concentrations. BRD4-IDR, MED1-IDR, or mEGFP were added to the droplet formation buffer to the final concentrations indicated. (**E**) Representative images of droplet formation at different salt concentrations. BRD4-IDR or MED1-IDR was added to droplet formation buffer to achieve 10 μM protein concentration with a final NaCl concentration as indicated. (**F**) Representative images of the droplet reversibility experiment with BRD4-IDR (top row) or MED1-IDR (bottom row) [20 μM protein and 75 mM NaCl (initial), followed by a 1:1 dilution (1/2 dilution) or a 1:1 dilution with an increase to 425 mM NaCl (1/2 dilution + NaCl)].

with 1,6-hexanediol caused a reduction in the number of BRD4 and MED1 puncta (Fig. 3, A and B).

To determine the effect of 1,6-hexanediol on BRD4, MED1, and RNA polymerase II (RNA Pol II) occupancy at enhancers and genes, ChIP-seq was performed with antibodies against these proteins in untreated or 1,6-hexanediol–treated mESCs. Treatment with 1,6-hexanediol caused a reduction in all three proteins at enhancers, with the most profound effects occurring at SEs (Fig. 3, C and D, and fig. S5A). For example, at the *Klf4* SE, the levels of BRD4 were reduced by 44%, those of MED1 by 80%, and those of RNA Pol II by 56% (Fig. 3C). Similar effects were observed genome-wide, where reductions in BRD4, MED1, and RNA Pol II were substantially larger at SEs than at typical enhancers (Fig. 3D), and the degrees to which BRD4 and MED1 were lost from SEs were positively correlated (fig. S5B). These results are consistent with the notion that BRD4 and MED1 form condensates at SEs that are sensitive to 1,6-hexanediol.

The level of RNA Pol II occupancy across gene bodies can be used as a measure of transcriptional output (*41*). The ChIP-seq data revealed that the reduction in BRD4 and MED1 occupancy at SEs was associated with a loss of RNA Pol II occupancy across SE-associated gene bodies (Fig. 3, C and E, and fig. S5A). When genes were ranked by the extent to which RNA Pol II was lost upon 1,6-hexanediol treatment, SE-associated genes were highly enriched among those that lost the most RNA Pol II (Fig. 3F). These results are consistent with the idea that BRD4 and MED1 condensates are associated with SEs and that loss of condensate integrity adversely affects transcription.

IDRs of BRD4 and MED1 phase-separate in vitro

BRD4 and MED1 contain large IDRs (Fig. 4A) and share features with the IDRs of several proteins known to facilitate condensate formation (*2*, *3*), including high proline and glutamine content (BRD4), high serine content (MED1), and acidic and basic regions (BRD4 and MED1). The purified IDRs of several proteins involved in

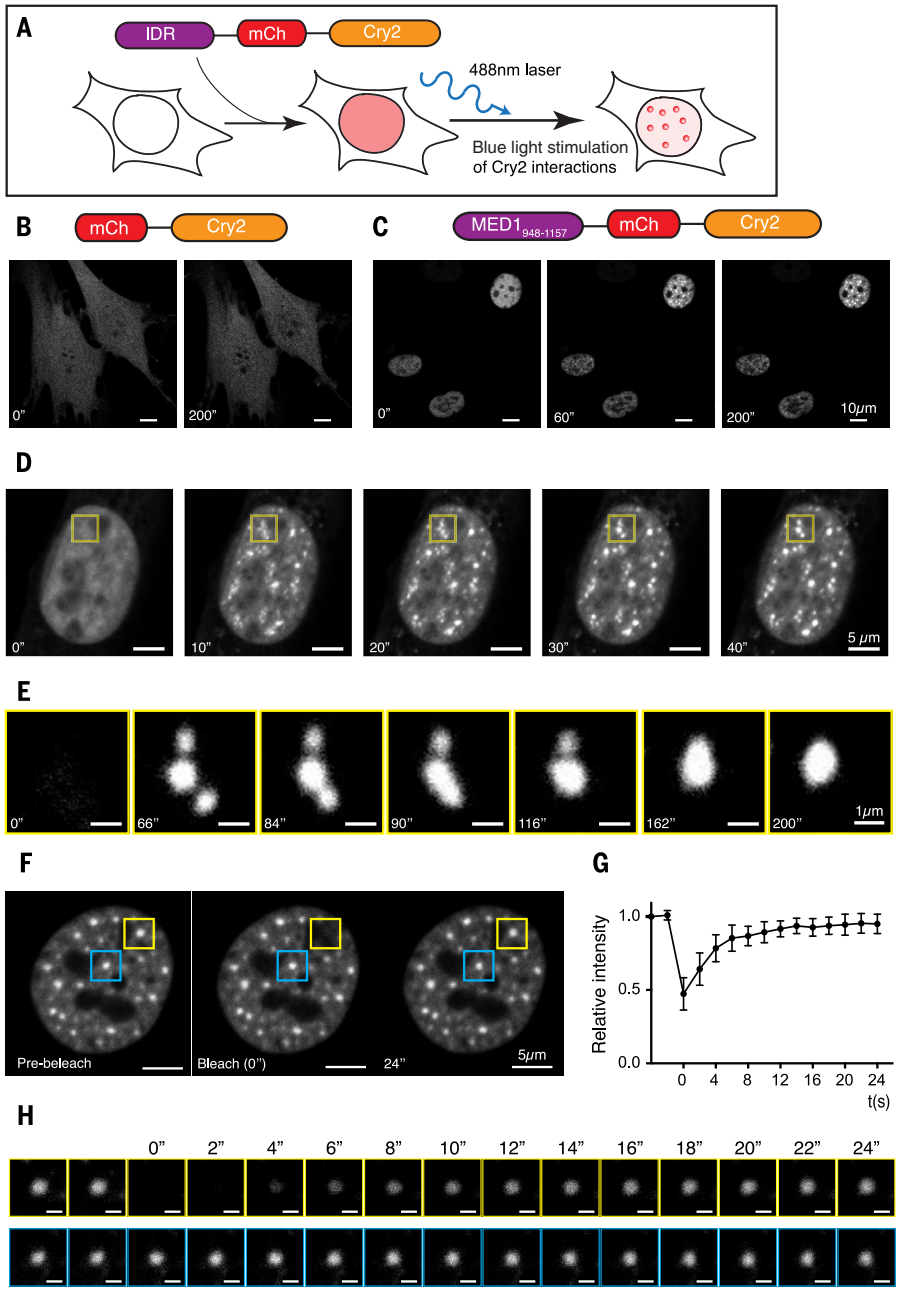


Fig. 5. The IDR of MED1 participates in phase separation in cells. (**A**) Schematic of the optoIDR assay, depicting recombinant protein with an IDR (purple), mCherry (red), and Cry2 (orange) expressed in cells exposed to blue light. (**B** and **C**) Images of NIH3T3 cells expressing either (B) mCherry-Cry2 or (C) a portion of the MED1-IDR (amino acids 948 to 1157) fused to mCherry-Cry2 (MED1-optoIDR). Cells were subjected to laser excitation every 2 s for the indicated times. (**D**) Time-lapse images of the nucleus of an NIH3T3 cell expressing MED1-optoIDR subjected to laser excitation every 2 s for the times indicated. A droplet fusion event occurs in the region highlighted by the yellow box. (**E**) The droplet fusion event highlighted in (D) at higher resolution and extended times as indicated. (**F**) Image of a MED1-optoIDR optoDroplet (yellow box) before (left), during (middle), and after (right) photobleaching. The blue box highlights an unbleached region for comparison. Time relative to photobleaching (0 s) is indicated. (**G**) Signal intensity relative to the prebleaching signal (*y* axis) versus time relative to photobleaching (*x* axis). Data are shown as average relative intensity ± SD (n = 15). (**H**) Time-lapse and close-up view of droplet recovery for regions highlighted in (F). Times relative to photobleaching are indicated. Scale bars, 1 μm.

condensate formation form phase-separated droplets in vitro (*7*, *36*, *37*, *42*), so we investigated whether the IDRs of BRD4 or MED1 form such droplets in vitro. Purified recombinant mEGFP-IDR fusion proteins (BRD4-IDR and MED1-IDR) (Fig. 4B) were added to buffers containing 10% PEG-8000 (polyethylene glycol, molecular weight 8000; materials and methods), turning the solution opaque, whereas equivalent solutions with mEGFP alone remained clear (Fig. 4C). Fluorescence microscopy of the opaque MED1-IDR and BRD4-IDR solutions revealed GFP-positive, micron-sized spherical droplets freely moving in solution (Movies 1 and 2) and falling onto and wetting the surface of the glass coverslip, where they remained stationary (Movie 3). As determined by aspect ratio analysis, the MED1-IDR and BRD4-IDR droplets were highly spherical (fig. S6A), a property expected for liquid-like droplets (*1*–*3*).

Phase-separated droplets typically scale in size according to the concentration of components in the system (*43*). We performed the droplet formation assay with varying concentrations of BRD4-IDR, MED1-IDR, and mEGFP, ranging from 0.625 to 20 μM. BRD4-IDR and MED1-IDR formed droplets with concentration-dependent size distributions, whereas mEGFP remained diffuse in all conditions tested (Fig. 4D and fig. S6B). Although these droplets were smaller at lower concentrations, we observed BRD4-IDR and MED1-IDR droplets at the lowest concentration tested (0.625 μM) (fig. S6C).

To investigate the biophysical properties of these droplets, we tested their ability to form under varying salt concentrations (to probe the contribution of electrostatic interactions) or upon 1,6-hexanediol treatment (to probe the contribution of hydrophobic interactions). The size distributions of both BRD4-IDR and MED1-IDR droplets shifted toward smaller droplets with increasing NaCl concentration (from 50 to 350 mM) (Fig. 4E and fig. S6D), and opacity was reduced with 10% 1,6-hexanediol treatment (fig. S7A). These results demonstrate that a variety of molecular interactions contribute to BRD4-IDR and MED1-IDR droplet formation.

We next sought to test whether the droplets are irreversible aggregates or reversible phase-separated condensates. To do this, BRD4-IDR and MED1-IDR were allowed to form droplets in an initial solution. The protein concentration was then diluted by half in equimolar salt or in

a higher salt solution (Fig. 4F). The preformed BRD4-IDR and MED1-IDR droplets were reduced in size and number with dilution and even further reduced with elevated salt concentration (Fig. 4F and fig. S7B). These results show that the BRD4-IDR and MED1-IDR droplets form a distribution of sizes that is dependent on the conditions of the system and, once formed, respond to changes in the system, with rapid adjustments in size. These features are characteristic of phase-separated condensates formed by networks of weak protein-protein interactions (*1*–*3*).

MED1-IDR participates in liquid-liquid phase separation in cells

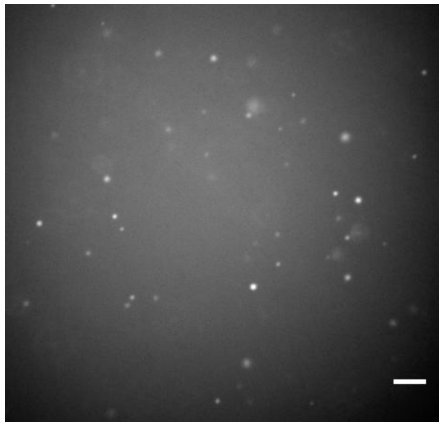
To investigate whether the coactivator IDRs facilitate phase separation in cells, we used a previously developed assay to manipulate local protein concentrations within the cell; this optoIDR assay tests IDR-dependent, light-inducible droplet formation in vivo (*44*). Briefly, the photoactivatable, self-associating Cry2 protein was labeled with mCherry and fused to an IDR of interest. This fusion mediates a blue light–inducible increase in the local concentration of selected IDRs within the cell (Fig. 5A) (*44*). In this assay, IDRs known to promote phase separation enhance the photoresponsive clustering properties of Cry2 (*45*, *46*), causing rapid formation of liquid-like spherical droplets under stimulation by blue light. Fusion of a portion of the MED1-IDR to Cry2-mCherry facilitated the rapid formation of micron-sized spherical droplets upon blue light stimulation (optoDroplets) (Fig. 5, B and C, and fig. S8). During stimulation, proximal droplets were observed to fuse (Fig. 5, D and E, and Movie 4). The fusions exhibited characteristic liquid-like fusion properties of necking and relaxation to spherical shape (Fig. 5E). The MED1-IDR droplets persisted after blue light stimulation and exhibited liquid-like FRAP recovery rates in the absence of blue light stimulation (Fig. 5, F to H). The rapid FRAP kinetics in the absence of light-activated Cry2 interactions suggests that the MED1-IDR optoDroplets established by blue light are dynamic assemblies exchanging with the dilute phase.

Conserved serine bias in the MED1-IDR is necessary for phase separation

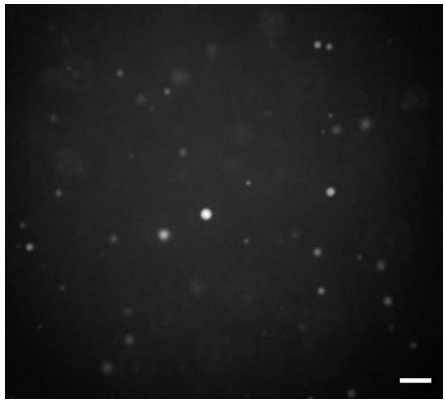
Previous studies have implicated low-complexity IDRs of proteins in liquid-liquid phase separation (*7*, *36*, *37*, *42*). An examination of the amino acid content of MED1 revealed that the IDR contains a compositional bias for serine (Fig. 6A). This serine compositional bias is conserved among vertebrates (Fig. 6B). To investigate whether the serine bias is necessary for the MED1-IDR's capacity to phase-separate, we mutated all the serine (S) residues to alanine (A) and investigated the ability of this mutated IDR to form phase-separated droplets in vitro. The MED1-IDR S-to-A mutant was incapable of forming phase-separated droplets under conditions in which the wild-type IDR readily formed droplets (Fig. 6C), indicating that the conserved serine bias in the MED1-IDR is necessary for droplet formation.

MED1-IDR droplets can incorporate proteins necessary for transcription

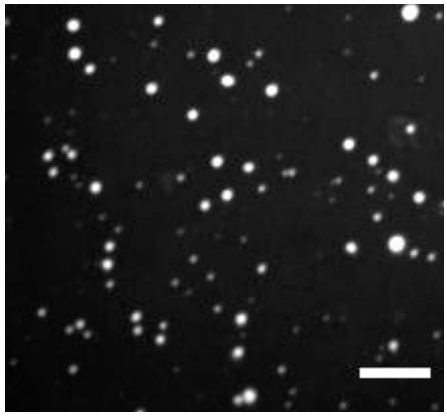
A proposed function of phase separation at SEs is the ability to compartmentalize and concentrate factors within a biomolecular condensate, so we sought to test whether MED1-IDR droplets could recapitulate this compartmentalization function in vitro. We identified conditions under which the MED1-IDR could form droplets but the BRD4-IDR could not (fig. S9). We then investigated whether the MED1-IDR droplets could compartmentalize BRD4-IDR protein under these conditions (Fig. 7A). Using differentially labeled proteins (mCherry–MED1-IDR and mEGFP–BRD4-IDR), we found that the MED1-IDR droplets could incorporate, and thus concentrate, the BRD4-IDR protein (Fig. 7A). The MED1-IDR droplets did not incorporate mEGFP (Fig. 7A). To probe the approximate mesh size of the MED1-IDR droplets (*47*), we incubated them with fluorescently labeled dextrans with average molecular weights of 4, 10, and 40 kDa. We found that the 4-kDa dextrans were incorporated into the MED1-IDR droplets, the 10-kDa dextrans were incorporated with less efficiency, and the 40-kDa dextrans were excluded (fig. S10). These results suggest that the incorporation of mEGFP–BRD4-IDR (105 kDa) into the MED1-IDR droplet is due to attractive molecular interactions, as opposed to passive diffusion through the droplet mesh.



Movie 1. BRD4-IDR droplets in solution. Each frame represents 1 s. The movie is rendered at 12 frames per second. 20 μM protein, 125 mM NaCl. Scale bar, 5 μm.



Movie 2. MED1-IDR droplets in solution. Each frame represents 1 s. The movie is rendered at 12 frames per second. 20 μM protein, 125 mM NaCl. Scale bar, 5 μm.



Movie 3. MED1-IDR droplets settling onto a glass coverslip. Each frame represents 1 s. The movie is rendered at 12 frames per second. 10 μM protein, 125 mM NaCl. Scale bar, 5 μm.

We next investigated whether the MED1-IDR, introduced into a transcription-competent nuclear extract, would form droplets that might incorporate BRD4 or other transcriptional components. We found that the wild-type MED1-IDR, but not the MED1-IDR S-to-A mutant, formed droplets in these extracts (Fig. 7B). The MED1-IDR phase-separated droplets were denser than the surrounding extract and thus could be purified from solution by centrifugation. Immunoblot analysis revealed that BRD4 and the largest subunit of RNA Pol II (RPB1) were enriched in pelleted droplets in a MED1-IDR dose–dependent manner (Fig. 7C). These results indicate that the MED1-IDR droplets can incorporate BRD4 and RNA Pol II.

The ability of the MED1-IDR protein to incorporate BRD4 and RNA Pol II into an artificial phase-separated compartment suggests that it sequesters key components of the transcription apparatus and might thus "squelch" transcription in the nuclear extract. We carried out an in vitro transcription assay with these extracts and found that the wild-type MED1-IDR protein does squelch transcription, correlating with the amount of material separated from solution by the MED1-IDR droplets (Fig. 7D). We did not observe these effects with equivalent concentrations of mEGFP or with the MED1-IDR S-to-A mutant (Fig. 7D). These results demonstrate that the MED1-IDR has the capacity to compartmentalize and concentrate transcriptional machinery from a complex nuclear extract.

Discussion

SEs regulate genes with prominent roles in healthy and diseased cellular states (*14*, *15*, *19*–*25*, *48*, *49*). SEs and their components have been proposed to form phase-separated condensates (*30*), but with no direct evidence. Here we demonstrate that two key components of SEs, BRD4 and MED1, form nuclear condensates at sites of SE-driven

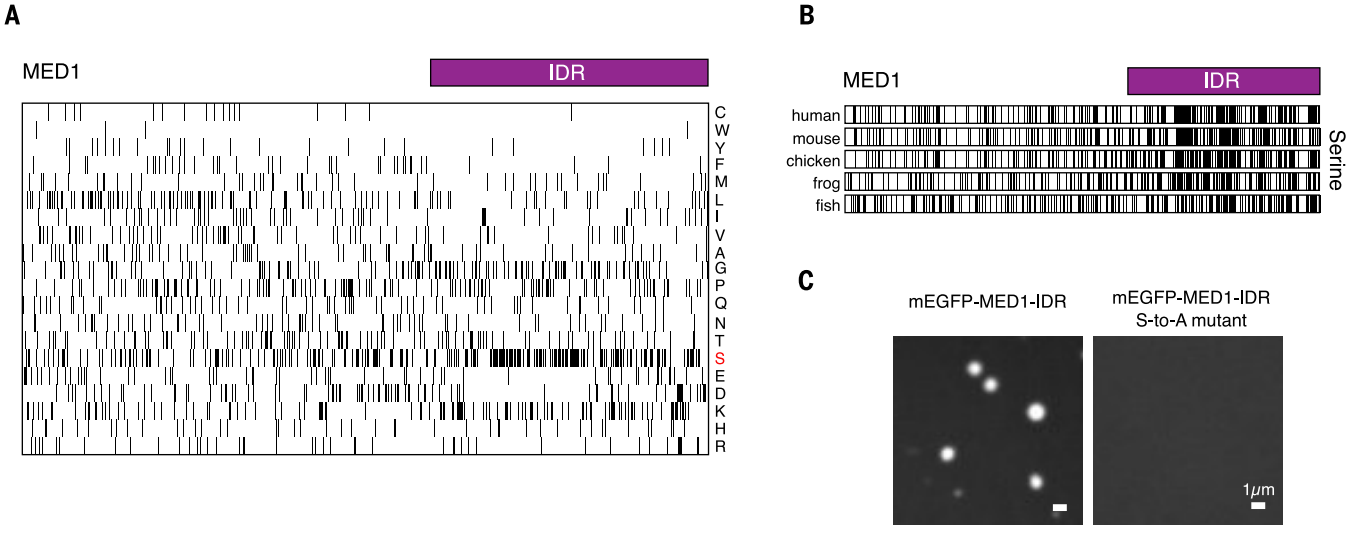
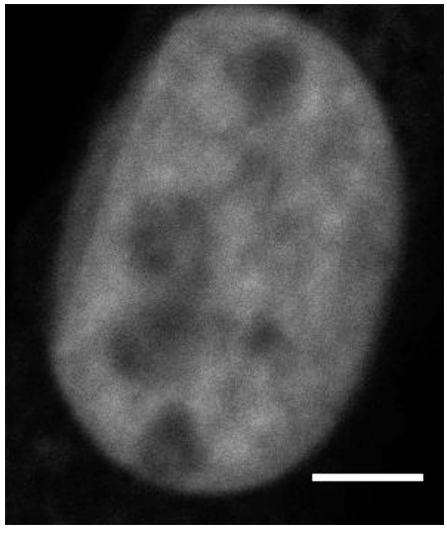


Fig. 6. Conserved serine bias is necessary for MED1-IDR phase separation. (**A**) Amino acid composition of the MED1 protein. Each row represents information for a single amino acid. Single-letter amino code abbreviations (right) are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; 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. The length of the row corresponds to the length of the MED1 protein. Black bars represent the occurrence of the indicated amino acid at that position in MED1. The purple bar represents the IDR of MED1 under investigation. (**B**) Serine composition of MED1 protein from indicated organisms, presented as in (A). (**C**) Mutating all serines to alanine (S to A) disrupts phase separation. Representative images of wild-type MED1-IDR or MED1-IDR S-to-A mutant fused to mEGFP in the droplet formation assay (10 μM protein, 125 mM NaCl, and 10% Ficoll-400).

transcription. Within these condensates, BRD4 and MED1 exhibit apparent diffusion coefficients similar to those previously reported for other proteins in phase-separated condensates in vivo (*36*, *37*). The IDRs of both BRD4 and MED1 are sufficient to form phase-separated droplets in vitro, and the MED1-IDR facilitates phase separation in living cells. Droplets formed by MED1-IDR are capable of concentrating transcriptional machinery in a transcriptionally competent nuclear extract. These results support a model in which transcriptional coactivators form phase-separated condensates that compartmentalize and concentrate the transcription apparatus at SE-regulated genes and identify SE components that likely play a role in phase separation.

SEs are established by the binding of master TFs to enhancer clusters (*14*, *15*). These TFs typically consist of a structured DNA-binding domain and an intrinsically disordered transcriptional activation domain (*50*–*52*). The activation domains of these TFs recruit high densities of many transcription proteins, which, as a class, are enriched for IDRs (*53*). Although the exact client-scaffold relationship (*54*) between these components remains unknown, it is likely that these protein sequences mediate weak multivalent interactions, thereby facilitating condensation. We propose that condensation of such high-valency factors at SEs creates a reaction crucible within the separated dense phase, where high local concentrations of the transcriptional machinery ensure robust gene expression.

The nuclear organization of chromosomes is likely influenced by condensates at SEs. DNA interaction technologies indicate that the individual enhancers within the SEs have exceptionally high interaction frequencies with one another (*16*–*18*), consistent with the idea that condensates draw these elements into close proximity in the dense phase. Several recent studies suggest that SEs can interact with one another and may also contribute in this fashion to chromosome organization (*55*, *56*). Cohesin, an SMC (structural maintenance of chromosomes) protein complex, has been implicated in constraining SE-SE interactions because its loss causes extensive fusion of SEs within the nucleus (*56*). These SE-SE interactions may be due to a tendency of liquid-phase condensates to undergo fusion (*1*–*3*).



Movie 4. Formation of MED1-IDR optoDroplets upon stimulation with blue light. NIH3T3 cells expressing the MED1-optoIDR construct were subjected to 488-nm laser light in 2-s intervals. Each frame represents 2 s. The movie is rendered at 12 frames per second. Scale bar, 5 μm.

The model whereby phase separation of coactivators compartmentalizes and concentrates the transcription apparatus at SEs and their regulated genes, described here and corroborated by (*57*), raises many questions. How does condensation contribute to regulation of transcriptional output? A study of RNA Pol II clusters, which may be phase-separated condensates, suggests a positive correlation between condensate lifetime and transcriptional output (*58*). What components drive formation and dissolution of transcriptional condensates? Our studies indicate that BRD4 and MED1 likely participate, but the roles of DNA-binding TFs, RNA Pol II, and regulatory RNAs require further study. Why do some proteins, such as HP1a, contribute to phase-separated heterochromatin condensates (*59*, *60*) and others contribute to euchromatic condensates? The rules that govern partitioning into specific types of condensates have begun to be studied (*61*–*65*) and will need to be defined for proteins involved in transcriptional condensates. Does condensate misregulation contribute to pathological processes in disease, and will new insights into condensate behaviors present new opportunities for therapy? Mutations within IDRs and misregulation of phase separation have already been implicated in a number of neurodegenerative diseases (*66*–*68*). Tumor cells have exceptionally large SEs at driver oncogenes that are not found in their cell of origin, and some of these are exceptionally sensitive to drugs that target SE components (*22*–*25*). How do we take advantage of phase separation

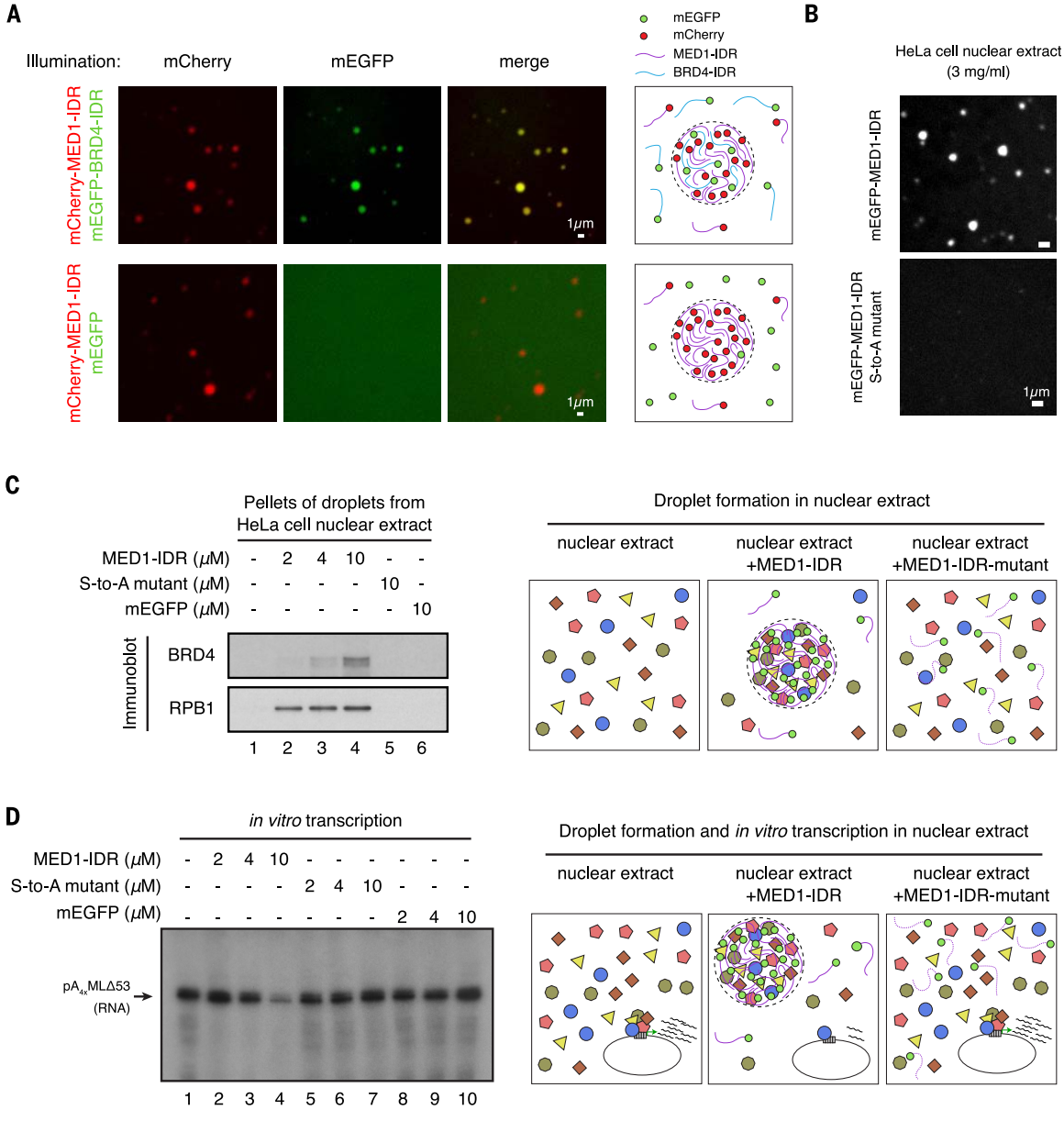


Fig. 7. MED1-IDR droplets compartmentalize and concentrate proteins necessary for transcription. (**A**) MED1-IDR droplets incorporate BRD4-IDR protein in vitro. The indicated mEGFP or mCherry fusion proteins were mixed at 10 μM each in buffer D containing 10% Ficoll-400 and 125 mM NaCl. Indicated fluorescence channels are presented for each mixture. Illustrations summarizing results are shown on the right. (**B**) MED1-IDR forms droplets in an in vitro transcription reaction containing HeLa cell nuclear extract, whereas the MED1-IDR S-to-A mutant does not. Shown are representative images of the indicated mEGFP-fusion protein when added to an in vitro transcription reaction containing HeLa cell nuclear extract at a final concentration of 3 mg/ml (a complete list of components is given in the methods). (**C**) MED1-IDR droplets compartmentalize transcriptional machinery from a nuclear extract. Shown are immunoblots of the pellet fraction of the indicated protein added to in vitro transcription reactions [as described in (B)]. A proposed model of molecular interactions taking place within MED1-IDR droplets in the nuclear extract is illustrated on the right. (**D**) MED1-IDR droplets compartmentalize machinery necessary for the in vitro transcription reaction. An autoradiograph of radiolabeled RNA products of in vitro transcription reactions under indicated conditions is shown on the left. The arrow indicates the expected RNA product. Reactions were conducted as in (*68*) with minor modifications (full details are given in the methods). A proposed model of molecular interactions taking place within MED1-IDR droplets in nuclear extract and the impact on the in vitro transcription reaction is illustrated on the right.

principles established in physics and chemistry to more effectively improve our understanding of this form of regulatory biology? Addressing these questions at the crossroads of physics, chemistry, and biology will require collaboration across these diverse sciences.

Methods summary

Immunofluorescence against BRD4 and MED1, coupled with DNA-FISH or RNA-FISH against SEs or SE-driven nascent transcripts, was performed in mESCs to visualize the colocalization between BRD4 or MED1 puncta and SEs. BRD4 and MED1 were endogenously tagged with mEGFP in mESCs to visualize the organization of BRD4 and MED1 and to study their dynamics by FRAP and drug treatments in live cells. ChIP-seq was performed to investigate the effect of 1,6-hexanediol treatment on the chromatin occupancy of BRD4, MED1, and RNA Pol II. Recombinant BRD4-IDR and MED1-IDR were purified to test their capacity to phase-separate in vitro. The optoIDR assay (*45*) was implemented to test the capacity of a section of MED1-IDR to phase-separate in live cells. Mutations were introduced into MED1-IDR to study the sequence determinants of MED1-IDR phase separation. BRD4-IDR and MED1-IDR fused to different fluorescent tags were used to demonstrate the capacity of MED1-IDR droplets to compartmentalize and concentrate BRD4-IDR. Formation of MED1-IDR droplets in a transcriptionally competent nuclear extract was used to study the ability of MED1-IDR droplets to compartmentalize and concentrate BRD4 and RNA Pol II from a complex extract. In vitro transcription assays were used to measure the effect of synthetic droplet formation on transcription. All procedures are described in detail in the supplementary materials.

REFERENCES AND NOTES

1. A. A. Hyman, C. A. Weber, F. Jülicher, Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58 (2014). doi: 10.1146/annurev-cellbio-100913-013325; pmid: 25288112
2. S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017). doi: 10.1038/nrm.2017.7; pmid: 28225081
3. Y. Shin, C. P. Brangwynne, Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017). doi: 10.1126/science.aaf4382; pmid: 28935776
4. C. P. Brangwynne *et al*., Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009). doi: 10.1126/science.1172046; pmid: 19460965
5. M. Kato *et al*., Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012). doi: 10.1016/j.cell.2012.04.017; pmid: 22579281
6. P. Li *et al*., Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–340 (2012). doi: 10.1038/nature10879; pmid: 22398450
7. Y. Lin, D. S. W. Protter, M. K. Rosen, R. Parker, Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol. Cell* **60**, 208–219 (2015). doi: 10.1016/j.molcel.2015.08.018; pmid: 26412307
8. K. Adelman, J. T. Lis, Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012). doi: 10.1038/nrg3293; pmid: 22986266
9. M. Bulger, M. Groudine, Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339 (2011). doi: 10.1016/j.cell.2011.01.024; pmid: 21295696
10. E. Calo, J. Wysocka, Modification of enhancer chromatin: What, how, and why? *Mol. Cell* **49**, 825–837 (2013). doi: 10.1016/j.molcel.2013.01.038; pmid: 23473601
11. F. Spitz, E. E. M. Furlong, Transcription factors: From enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012). doi: 10.1038/nrg3207; pmid: 22868264
12. W. Xie, B. Ren, Developmental biology. Enhancing pluripotency and lineage specification. *Science* **341**, 245–247 (2013). doi: 10.1126/science.1236254; pmid: 23869010
13. M. Levine, C. Cattoglio, R. Tjian, Looping back to leap forward: Transcription enters a new era. *Cell* **157**, 13–25 (2014). doi: 10.1016/j.cell.2014.02.009; pmid: 24679523
14. W. A. Whyte *et al*., Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013). doi: 10.1016/j.cell.2013.03.035; pmid: 23582322
15. D. Hnisz *et al*., Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013). doi: 10.1016/j.cell.2013.09.053; pmid: 24119843
16. J. M. Dowen *et al*., Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014). doi: 10.1016/j.cell.2014.09.030; pmid: 25303531
17. D. Hnisz *et al*., Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016). doi: 10.1126/science.aad9024; pmid: 26940867
18. X. Ji *et al*., 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* **18**, 262–275 (2016). doi: 10.1016/j.stem.2015.11.007; pmid: 26686465
19. M. R. Mansour *et al*., An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014). doi: 10.1126/science.1259037; pmid: 25394790
20. J. D. Brown *et al*., NF-κB directs dynamic super enhancer formation in inflammation and atherogenesis. *Mol. Cell* **56**, 219–231 (2014). doi: 10.1016/j.molcel.2014.08.024; pmid: 25263595
21. B. Chapuy *et al*., Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell* **24**, 777–790 (2013). doi: 10.1016/j.ccr.2013.11.003; pmid: 24332044
22. J. Lovén *et al*., Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013). doi: 10.1016/j.cell.2013.03.036; pmid: 23582323
23. E. Chipumuro *et al*., CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell* **159**, 1126–1139 (2014). doi: 10.1016/j.cell.2014.10.024; pmid: 25416950
24. N. Kwiatkowski *et al*., Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature* **511**, 616–620 (2014). doi: 10.1038/nature13393; pmid: 25043025
25. Y. Wang *et al*., CDK7-dependent transcriptional addiction in triple-negative breast cancer. *Cell* **163**, 174–186 (2015). doi: 10.1016/j.cell.2015.08.063; pmid: 26406377
26. D. Hnisz *et al*., Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell* **58**, 362–370 (2015). doi: 10.1016/j.molcel.2015.02.014; pmid: 25801169
27. T. Jiang *et al*., Identification of multi-loci hubs from 4C-seq demonstrates the functional importance of simultaneous interactions. *Nucleic Acids Res.* **44**, 8714–8725 (2016). doi: 10.1093/nar/gkw568; pmid: 27439714
28. C. Proudhon *et al*., Active and Inactive Enhancers Cooperate to Exert Localized and Long-Range Control of Gene Regulation. *Cell Rep.* **15**, 2159–2169 (2016). pmid: 27239026
29. H. Y. Shin *et al*., Hierarchy within the mammary STAT5-driven *Wap* super-enhancer. *Nat. Genet.* **48**, 904–911 (2016). doi: 10.1038/ng.3606; pmid: 27376239
30. D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, P. A. Sharp, A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017). doi: 10.1016/j.cell.2017.02.007; pmid: 28340338
31. Z. Yang *et al*., Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol. Cell* **19**, 535–545 (2005). doi: 10.1016/j.molcel.2005.06.029; pmid: 16109377
32. M. K. Jang *et al*., The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol. Cell* **19**, 523–534 (2005). doi: 10.1016/j.molcel.2005.06.027; pmid: 16109376
33. R. Di Micco *et al*., Control of embryonic stem cell identity by BRD4-dependent transcriptional elongation of super-enhancer-associated pluripotency genes. *Cell Rep.* **9**, 234–247 (2014). pmid: 25263550
34. J. Soutourina, S. Wydau, Y. Ambroise, C. Boschiero, M. Werner, Direct interaction of RNA polymerase II and mediator required for transcription in vivo. *Science* **331**, 1451–1454 (2011). doi: 10.1126/science.1200188; pmid: 21415355
35. J. Soutourina, Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol.* **19**, 262–274 (2018). pmid: 29209056
36. T. J. Nott *et al*., Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015). doi: 10.1016/j.molcel.2015.01.013; pmid: 25747659
37. C. W. Pak *et al*., Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. *Mol. Cell* **63**, 72–85 (2016). doi: 10.1016/j.molcel.2016.05.042; pmid: 27392146
38. C. P. Brangwynne, T. J. Mitchison, A. A. Hyman, Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4334–4339 (2011). doi: 10.1073/pnas.1017150108; pmid: 21368180
39. A. Patel *et al*., ATP as a biological hydrotrope. *Science* **356**, 753–756 (2017). doi: 10.1126/science.aaf6846; pmid: 28522535
40. S. Kroschwald, S. Maharana, A. Simon, Hexanediol: A chemical probe to investigate the material properties of membrane-less compartments. *Matters* 10.19185/matters.201702000010 (2017).
41. C. Y. Lin *et al*., Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012). doi: 10.1016/j.cell.2012.08.026; pmid: 23021215
42. S. Elbaum-Garfinkle *et al*., The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7189–7194 (2015). doi: 10.1073/pnas.1504822112; pmid: 26015579
43. C. P. Brangwynne, Phase transitions and size scaling of membrane-less organelles. *J. Cell Biol.* **203**, 875–881 (2013). doi: 10.1083/jcb.201308087; pmid: 24368804
44. Y. Shin *et al*., Spatiotemporal Control of Intracellular Phase Transitions Using Light-Activated optoDroplets. *Cell* **168**, 159–171.e14 (2017). doi: 10.1016/j.cell.2016.11.054; pmid: 28041848
45. I. Ozkan-Dagliyan *et al*., Formation of Arabidopsis Cryptochrome 2 photobodies in mammalian nuclei: Application as an optogenetic DNA damage checkpoint switch. *J. Biol. Chem.* **288**, 23244–23251 (2013). doi: 10.1074/jbc.M113.493361; pmid: 23833191
46. X. Yu *et al*., Formation of nuclear bodies of Arabidopsis CRY2 in response to blue light is associated with its blue light-dependent degradation. *Plant Cell* **21**, 118–130 (2009). doi: 10.1105/tpc.108.061663; pmid: 19141709
47. M.-T. Wei *et al*., Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* **9**, 1118–1125 (2017). doi: 10.1038/nchem.2803; pmid: 29064502
48. H. I. Suzuki, R. A. Young, P. A. Sharp, Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. *Cell* **168**, 1000–1014.e15 (2017). doi: 10.1016/j.cell.2017.02.015; pmid: 28283057
49. J. E. Bradner, D. Hnisz, R. A. Young, Transcriptional Addiction in Cancer. *Cell* **168**, 629–643 (2017). doi: 10.1016/j.cell.2016.12.013; pmid: 28187285
50. M. Ptashne, How eukaryotic transcriptional activators work. *Nature* **335**, 683–689 (1988). doi: 10.1038/335683a0; pmid: 3050531
51. P. J. Mitchell, R. Tjian, Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989). doi: 10.1126/science.2667136; pmid: 2667136
52. J. Liu *et al*., Intrinsic disorder in transcription factors. *Biochemistry* **45**, 6873–6888 (2006). doi: 10.1021/bi0602718; pmid: 16734424
53. H. Xie *et al*., Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* **6**, 1882–1898 (2007). doi: 10.1021/pr060392u; pmid: 17391014
54. S. F. Banani *et al*., Compositional Control of Phase-Separated Cellular Bodies. *Cell* **166**, 651–663 (2016). doi: 10.1016/j.cell.2016.06.010; pmid: 27374333
55. R. A. Beagrie *et al*., Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017). pmid: 28273065
56. S. S. P. Rao *et al*., Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305–320.e24 (2017). doi: 10.1016/j.cell.2017.09.026; pmid: 28985562
57. W.-K. Cho *et al*., Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 (2018).
58. W.-K. Cho *et al*., RNA Polymerase II cluster dynamics predict mRNA output in living cells. *eLife* **5**, 1123 (2016). doi: 10.7554/eLife.13617; pmid: 27138339
59. A. G. Larson *et al*., Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **547**, 236–240 (2017). doi: 10.1038/nature22822; pmid: 28636604
60. A. R. Strom *et al*., Phase separation drives heterochromatin domain formation. *Nature* **547**, 241–245 (2017). doi: 10.1038/nature22989; pmid: 28636597
61. M. Feric *et al*., Coexisting Liquid Phases Underlie Nucleolar Subcompartments. *Cell* **165**, 1686–1697 (2016). doi: 10.1016/j.cell.2016.04.047; pmid: 27212236
62. T. S. Harmon, A. S. Holehouse, M. K. Rosen, R. V. Pappu, Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife* **6**, e30294 (2017). doi: 10.7554/eLife.30294; pmid: 29091028
63. J. A. Riback *et al*., Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**, 1028–1040.e19 (2017). doi: 10.1016/j.cell.2017.02.027; pmid: 28283059
64. S. Boeynaems *et al*., Phase Separation of C9orf72 Dipeptide Repeats Perturbs Stress Granule Dynamics. *Mol. Cell* **65**, 1044–1055.e5 (2017). doi: 10.1016/j.molcel.2017.02.013; pmid: 28306503
65. J. P. Brady *et al*., Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8194–E8203 (2017). doi: 10.1073/pnas.1706197114; pmid: 28894006
66. A. Patel *et al*., A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **162**, 1066–1077 (2015). doi: 10.1016/j.cell.2015.07.047; pmid: 26317470
67. A. Molliex *et al*., Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015). doi: 10.1016/j.cell.2015.09.015; pmid: 26406374
68. A. Jain, R. D. Vale, RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017). pmid: 28562589

ACKNOWLEDGMENTS

We thank W. Salmon of the W. M. Keck Microscopy Facility; D. Richardson and S. Terclavers of the Harvard Center for Biological Imaging; and T. Volkert, D. Reynolds, S. Mraz, and S. Gupta of the Whitehead Genome Technologies Core for technical

assistance. We thank the Imaging Platform at the Broad Institute for assistance with CellProfiler. **Funding:** The work was supported by NIH grants GM123511 (R.A.Y.) and P01-CA042063 (P.A.S.), NSF grant PHY-1743900 (A.K.C., R.A.Y., and P.A.S.), Koch Institute Support (core) grant P30-CA14051 from the NCI (P.A.S.), Damon Runyon Cancer Research Foundation Fellowship 2309-17 (B.R.S.), Swedish Research Council Postdoctoral Fellowship VR 2017-00372 (A.B.), a Hope Funds for Cancer Research fellowship (B.J.A.), an NSF Graduate Research Fellowship (A.V.Z.), a Cancer Research Institute Irvington Fellowship (Y.E.G.), American Cancer Society New England Division Postdoctoral Fellowship PF-16-146-01-DMC (D.S.D.), and a NWO Rubicon Fellowship (J.S.). **Author contributions:** B.R.S., A.D., and R.A.Y. conceptualized and organized the project and wrote the manuscript. A.D., A.B., J.C.M., and Y.E.G. performed cell-imaging experiments and image analysis. I.A.K. and A.V.Z. generated endogenously tagged cell lines. B.R.S. and A.B. performed ChIP-seq. B.R.S. and E.L.C. performed in vitro droplet assays and optoIDR experiments. K.S. and B.J.A. developed and performed image analysis and produced visualizations. B.J.A. performed ChIP-seq analysis and produced visualizations. N.M.H. produced and purified recombinant proteins. A.V.Z. helped with biochemical experiments. C.H.L. performed protein amino acid analysis. D.S.D. performed ChIA-PET analysis and visualization. B.R.S., I.A.K., E.L.C., J.S., and A.V.Z. generated constructs. S.M. performed in vitro transcription assays. D.H., E.V., T.I.L., I.I.C., R.G.R., P.A.S., A.K.C., and R.A.Y. provided input into experimental design and interpretation. P.A.S., A.K.C., and R.A.Y. acquired funding for this study. R.A.Y. supervised the project with help from T.I.L. and A.K.C. All authors contributed to editing the manuscript. **Competing interests:** The Whitehead Institute filed a patent application based on this paper. R.A.Y. is a founder and shareholder of Syros Pharmaceuticals, Camp4 Therapeutics, and Omega Therapeutics. B.J.A. and T.I.L. are shareholders of Syros Pharmaceuticals, and T.I.L. is a consultant to Camp4 Therapeutics. All other authors declare no competing interests. **Data and materials availability:** Datasets generated in this study have been deposited in the Gene Expression Omnibus under accession number GSE112808.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/eaar3958/suppl/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 to S3
References (*69*–*82*)
Data S1

4 November 2017; resubmitted 9 April 2018
Accepted 6 June 2018
Published online 21 June 2018
10.1126/science.aar3958

RESEARCH ARTICLE SUMMARY

BIOTECHNOLOGY

Three-dimensional intact-tissue sequencing of single-cell transcriptional states

Xiao Wang*, William E. Allen*, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan†, Felice-Alessio Bava†, Karl Deisseroth†

INTRODUCTION: Single-cell RNA sequencing has demonstrated that both stable cell types and transient cell states can be discovered and defined by transcriptomes. In situ transcriptomic methods can map both RNA quantity and position; however, it remains challenging to simultaneously satisfy key technological requirements such as efficiency, signal intensity, accuracy, scalability to large gene numbers, and applicability to three-dimensional (3D) volumes. Well-established single-molecule fluorescence in situ hybridization (FISH) approaches (such as MERFISH and seqFISH) have high detection efficiency but require long RNA species (more than 1000 nucelotides) and yield lower intensity than that of enzymatic amplification methods (tens versus thousands of fluorophores per RNA molecule). Other pioneering in situ sequencing methods (via padlock probes and fluorescent in situ sequencing) use enzymatic amplification, thus achieving high intensity but with room to improve on efficiency.

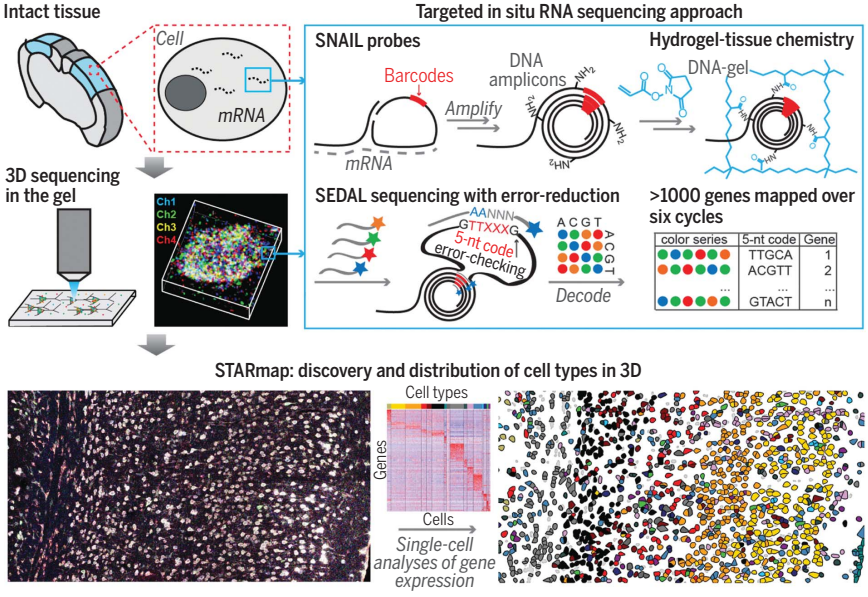
RATIONALE: We have developed, validated, and applied STARmap (spatially-resolved transcript amplicon readout mapping). STARmap begins with labeling of cellular RNAs by pairs of DNA probes followed by enzymatic amplification so as to produce a DNA nanoball (amplicon), which eliminates background caused by mislabeling of single probes. Tissue can then be transformed into a 3D hydrogel DNA chip by anchoring DNA amplicons via an in situ–synthesized polymer network and removing proteins and lipids. This form of hydrogel-tissue chemistry replots amplicons onto an optically transparent hydrogel coordinate system; then, to identify and quantify RNA species-abundance manifested by DNA amplicons, the identity of each species is encoded as a five-base barcode and read out by means of an in situ sequencing method that decodes DNA sequence in multicolor fluorescence. Using a new two-base sequencing scheme (SEDAL), STARmap was found to simultaneously detect more than 1000 genes over six imaging cycles, in which sequencing errors in any cycle cause misdecoding and are effectively rejected.

RESULTS: We began by (i) detecting and quantifying a focused 160-gene set (including cell type markers and activity-regulated genes) simultaneously in mouse primary visual cortex; (ii) clustering resulting per-cell gene expression patterns into a dozen distinct inhibitory, excitatory, and non-neuronal cell types; and (iii) mapping the spatial distribution of all of these cell types across layers of cortex. For validation, per-cell-type gene expression was found to correlate well both with in situ hybridization results and with single-cell RNA sequencing, and widespread up-regulation of activity-regulated genes was observed in response to visual stimulation. We next applied STARmap to a higher cognitive area (the medial prefrontal cortex) and discovered a more complex distribution of cell types. Last, we extended STARmap to much larger numbers of genes and spatial scales; we measured 1020 genes simultaneously in sections—obtaining results concordant with the 160-gene set—and measured 28 genes across millimeter-scale volumes encompassing ~30,000 cells, revealing 3D patterning principles that jointly characterize a broad and diverse spectrum of cell types.

CONCLUSION: STARmap combines hydrogel-tissue chemistry and in situ DNA sequencing to achieve intact-tissue single-cell measurement of expression of more than a thousand genes. In the future, combining this intact-system gene expression measurement with complementary cellular-resolution methodologies (with which STARmap is designed to be compatible)—including in vivo activity recording, optogenetic causal tests, and anatomical connectivity in the same cells—will help bridge molecular, cellular, and circuit scales of neuroscience. ■

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aat5691



STARmap for 3D transcriptome imaging and molecular cell typing. STARmap is an in situ RNA-sequencing technology that transforms intact tissue into a 3D hydrogel-tissue hybrid and measures spatially resolved single-cell transcriptomes in situ. Error- and background-reduction mechanisms are implemented at multiple layers, enabling precise RNA quantification, spatially resolved cell typing, scalability to large gene numbers, and 3D mapping of tissue architecture.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: deissero@stanford.edu (K.D.); alessio.bava@gmail.com (F.-A.B.); gnolan@stanford.edu (G.P.N.)
Cite this article as X. Wang *et al.*, *Science* **361**, eaat5691 (2018). DOI: 10.1126/science.aat5691

RESEARCH ARTICLE

BIOTECHNOLOGY

Three-dimensional intact-tissue sequencing of single-cell transcriptional states

Xiao Wang[1]*, William E. Allen[1,2]*, Matthew A. Wright[1,3], Emily L. Sylwestrak[1], Nikolay Samusik[4], Sam Vesuna[1], Kathryn Evans[1], Cindy Liu[1], Charu Ramakrishnan[1], Jia Liu[5], Garry P. Nolan[4]†, Felice-Alessio Bava[4]†‡, Karl Deisseroth[1,3,6]†

Retrieving high-content gene-expression information while retaining three-dimensional (3D) positional anatomy at cellular resolution has been difficult, limiting integrative understanding of structure and function in complex biological tissues. We developed and applied a technology for 3D intact-tissue RNA sequencing, termed STARmap (spatially-resolved transcript amplicon readout mapping), which integrates hydrogel-tissue chemistry, targeted signal amplification, and in situ sequencing. The capabilities of STARmap were tested by mapping 160 to 1020 genes simultaneously in sections of mouse brain at single-cell resolution with high efficiency, accuracy, and reproducibility. Moving to thick tissue blocks, we observed a molecularly defined gradient distribution of excitatory-neuron subtypes across cubic millimeter–scale volumes (>30,000 cells) and a short-range 3D self-clustering in many inhibitory-neuron subtypes that could be identified and described with 3D STARmap.

In biological tissues, diversity of function arises from diversity of form—in part, via the complexity of cell-specific gene expression, which defines the distinct three-dimensional (3D) molecular anatomy and cellular properties of each tissue. In situ transcriptomic tools for the spatial mapping of gene expression with subcellular resolution have emerged that may be applicable to probing these tissue structure–function relationships, including both multiplexed in situ RNA hybridization and in situ RNA sequencing (*1*–*10*). Current in situ sequencing approaches face the challenge of implementing enzymatic reactions in the dense, complex tissue environment and currently suffer from low efficiency (*2*), but the potential value of such intact-tissue sequencing could be enormous; in comparison with hybridization-based multiplexing/readout, which uses multiple polynucleotide probes to encode gene identity (*3*–*5*), sequencing operates with single-nucleotide resolution and thus inherently provides greater information. In addition, in situ sequencing methods typically use signal amplification, which is important for the detection of short transcripts (such as neuropeptides) and for high-quality imaging in thick tissue blocks. However, current sequencing methods have not yet been successfully applied to 3D volumes of intact tissue because of fundamental limitations in requisite sensitivity, fidelity, and scalability for throughput in tissues such as the mammalian brain.

Hydrogels have been widely used for extracellular 3D scaffolding in applications across biology and medicine (*11*–*13*). Recently developed hydrogel-tissue chemistry (HTC) methodologies (*14*), beginning with CLARITY (*15*), physically link in situ–synthesized polymers with selected intracellular biomolecules. This process transforms the tissue, from within its constituent cells, into a new state suitable for high-resolution volumetric imaging and analysis compatible with many kinds of molecular phenotyping for proteins, nucleic acids, and other targets (*15*). HTC-based hydrogel-embedding strategies have been extended to nucleic acid analyses in the form of in situ hybridization for RNA (*16*–*19*), but these have not yet been extended to in situ RNA sequencing—which would have the potential to reveal the full molecular complexity of the transcriptome. In nontissue environments, however, purely synthetic hydrogels have been used to accommodate enzymatic reactions that include DNA sequencing (*20*), and if biological tissue could be converted into a hydrogel-embedded form compatible with creation, retention, and functional presentation of RNA-derived or hybridized complementary DNA (cDNA), it might be possible to perform 3D in situ sequencing within such a tissue-hydrogel formulation—leveraging the crucial attendant properties of optical transparency, reduced background, elevated diffusion rate, and greater mechanical stability. We achieved this goal with the development and application of a sequencing-based method (spatially-resolved transcript amplicon readout mapping, or STARmap) for targeted 3D in situ transcriptomics in intact tissue (Fig. 1A); using STARmap, we were able to identify organizational principles of a full spectrum of cell types, which would not have been otherwise accessible for identification in the adult mammalian brain.

Results

Design and validation of STARmap principles

One component is an efficient approach for in situ amplification of a library of cDNA probes hybridized with cellular RNAs (this approach is termed SNAIL, for specific amplification of nucleic acids via intramolecular ligation). Reverse transcription may be the major efficiency-limiting step for in situ sequencing (*7*, *21*), and SNAIL bypasses this step with a pair of primer and padlock probes (fig. S1A) designed so that only when both probes hybridize to the same RNA molecule can the padlock probe be circularized and rolling-circle-amplified to generate a DNA nanoball (amplicon) that contains multiple copies of the cDNA probes (Fig. 1, A to D). This mechanism ensures target-specific signal amplification and excludes noise that invariably otherwise arises from nonspecific hybridization of single probes. The outcome includes much higher absolute intensity and signal-to-noise ratio (SNR) as compared with those of commercial single-molecule fluorescent in situ hybridization (smFISH) probes (fig. S1, B to F) and substantial improvement of detection efficiency (comparable with that of single-cell RNA sequencing), with simplified experimental procedures compared with previous in situ RNA sequencing methods (fig. S1, G to I).

To enable cDNA amplicon embedding in the tissue-hydrogel setting, amine-modified nucleotides were spiked into the rolling-circle amplification reaction, functionalized with an acrylamide moiety by using acrylic acid *N*-hydroxysuccinimide esters, and copolymerized with acrylamide monomers so as to form a distinct kind of hydrogel-DNA amplicon network (Fig. 1A and fig. S2A). The resulting tissue-hydrogel was then subjected to protein digestion and lipid removal in order to enhance transparency (fig. S2, B to E). This design chemistry dictates that amplicons are covalently linked with the hydrogel network, and such cross-linking is essential to maintain the position and integrity of the amplicons through many cycles of detection (fig. S2, F to H).

A five-base barcode (library size of 1024) was designed and built into each padlock probe as a gene-unique identifier to be sequenced, thus enabling multiplexed gene detection (Fig. 1A). Sequencing-by-synthesis paradigms were avoided because these require elevated reaction temperatures, which in turn are problematic for

[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA. [2]Neuroscience Program, Stanford University, CA 94305, USA. [3]Department of Psychiatry and Behavioral Sciences, Stanford University, CA 94305, USA. [4]Baxter Laboratory, Department of Microbiology and Immunology, Stanford University, Stanford, CA 94305, USA. [5]Department of Chemical Engineering, Stanford University, CA 94305, USA. [6]Howard Hughes Medical Institute, Stanford University, CA 94305, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: deissero@stanford.edu (K.D.); alessio.bava@gmail.com (F.-A.B.); gnolan@stanford.edu (G.P.N.)
‡Present address: Institut Curie, PSL Research University, Université Paris Sud, Université Paris-Saclay, Centre Universitaire, CNRS UMR 3348, Orsay 91405, France.

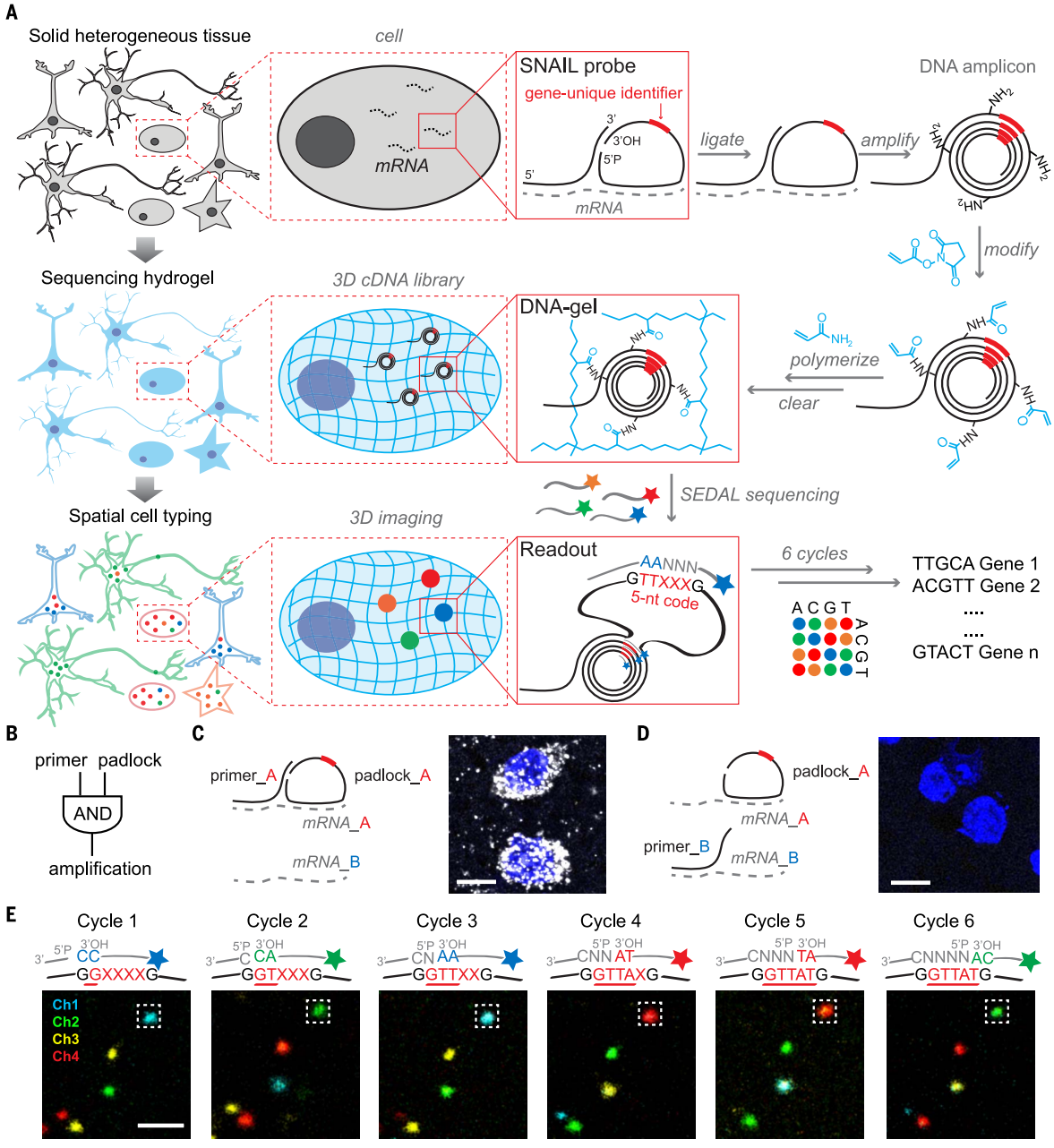


Fig. 1. STARmap principles: in situ RNA sequencing for spatial transcriptomics within the 3D tissue environment. (**A**) STARmap overview schematic. After brain tissue is prepared (mouse brain protocols are available in the supplementary materials, materials and methods), the custom SNAIL probes that encounter and hybridize to intracellular mRNAs (dashed lines) within the intact tissue are enzymatically replicated as cDNA amplicons. The amplicons are constructed in situ with an acrylic acid *N*-hydroxysuccinimide moiety modification (blue) and then copolymerized with acrylamide to embed within a hydrogel network (blue wavy lines), followed by clearance of unbound lipids and proteins (fig. S2). Each SNAIL probe contains a gene-specific identifier segment (red) that is read-out through in situ sequencing with two-base encoding for error correction (SEDAL) (fig. S3). Last, highly multiplexed RNA quantification in three dimensions reveals gene expression and cell types in space. (**B**) SNAIL logic. A pair of primer and padlock probes amplifies target-specific signals and excludes noise known to commonly arise from nonspecific hybridization of a single probe. (**C** and **D**) Only adjacent binding of primer and padlock probes leads to signal amplification. mRNA A represents *Gapdh*, and mRNA B represents *Actb*. Both fluorescent images show *Gapdh* (gray) mRNA and cell nuclei (blue) labeling in mouse brain slice; there is an absence of labeling with mismatched primer and padlock (right). Scale bar, 10 μm. (**E**) In situ sequencing of DNA amplicons in the tissue-hydrogel complex via SEDAL, the sequencing-by-ligation method devised for STARmap. For each cycle, the reading probes (gray line without star-symbol label) contain an incrementally increasing-length run of degenerate bases (N representing an equal mixture of A, T, C, and G) with phosphate at the 5′ end (5′P) to set the reading position; the decoding probes (gray line with star-symbol label) are labeled by fluorophores with color coding for the dinucleotide at the 3′ end. Only if both probes are perfectly complementary to the DNA template (black lower sequence) can the two kinds of probes then be ligated to form a stable product with a high melting temperature, allowing later imaging after unligated probes are washed away. After each imaging cycle, probes are stripped away from the robust tissue-hydrogel by using 60% formamide so that the next cycle can begin. X, unknown base to be read; red underline, decoded sequence; Ch1 to Ch4, fluorescence channels. Scale bar, 2 μm.

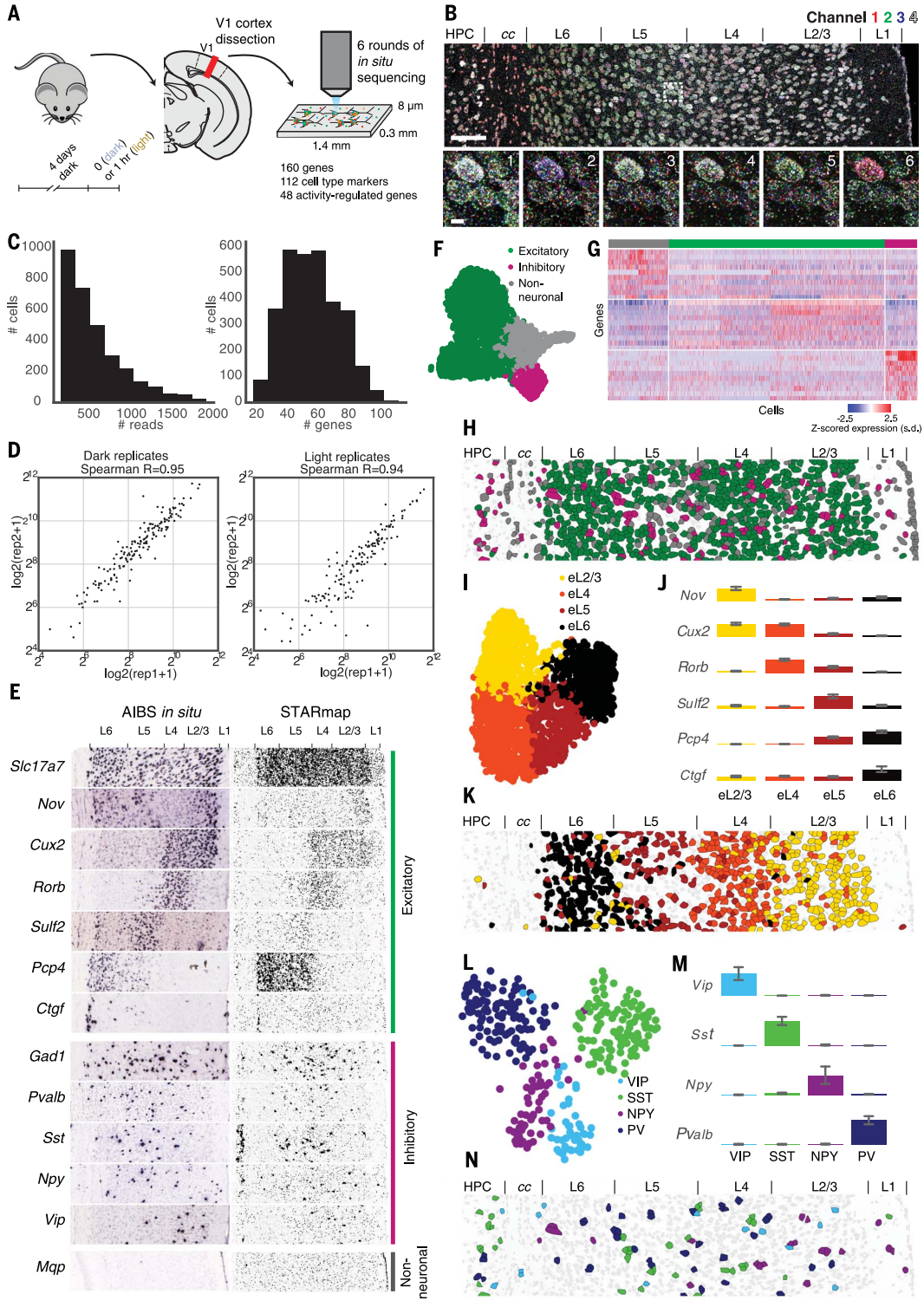


Fig. 2. STARmapping cell types in V1. (**A**) Experimental design. Mice were dark housed, before sacrifice, for 4 days and then either kept in the dark or exposed to light for 1 hour. V1 was coronally sectioned, and RNAs of 112 cell type markers and 48 activity-regulated genes were quantified by means of STARmap. (**B**) Raw fluorescence images of in-process STARmap with the full view of cycle 1 (top) and zoomed views across all six cycles (bottom). Full field: 1.4 by 0.3 mm, scale bar, 100 µm; zoomed region: 11.78 by 11.78 µm, scale bar, 2 µm; Channel, color code for the four fluorescence channels; L1 to L6, the six neocortical layers; *cc*, corpus callosum; HPC, hippocampus. (**C**) Histograms. Shown are detected reads (DNA amplicons) per cell (left), and genes per cell (right). (**D**) Quantitative reproducibility of biological replicates, whether in the light or dark condition: $\log_2$ (amplicon quantity) for 160 genes across the whole imaging region plotted. Rep1, expression value in first replicate; rep2, expression value in second replicate. (**E**) Validation of STARmap. (Left) in situ images from Allen Institute of Brain Science (AIBS). (Right) RNA pattern of individual genes extracted from 160-gene STARmap, which reliably reproduced the spatial gene expression pattern from AIBS. (**F**) Uniform manifold approximation plot (UMAP), a nonlinear dimensionality reduction technique used to visualize the similarity of cell transcriptomes in two dimensions, showing consistent clustering of major cell types across 3142 cells pooled from four biological replicates: 2199 excitatory neurons, 324 inhibitory neurons, and 619 non-neuronal cells. (**G**) Gene expression heatmap for 112 cell-type markers aligned with each cell cluster, showing clustering by inhibitory, excitatory, or non-neuronal cell types. Expression for each gene is *z*-scored across all genes in each cell. (**H**) Representative cell-resolved spatial map in neocortex and beyond. Cell types are color-coded as in (F). (**I** to **N**) Clustering of excitatory and inhibitory subtypes. [(I) and (L)] UMAP plots, [(J) and (M)] bar plots of representative genes (mean ± 95% confidence interval expression across all cells in that cluster, with each bar scaled to the maximum mean expression across all clusters), and [(K) and (N)] in situ spatial distribution of [(I) to (K)] excitatory and [(L) to (N)] inhibitory neurons. The number of cells in each cluster was as follows: eL2/3, 589; eL4, 649; eL5, 393; eL6, 368; PV neurons, 111; VIP neurons, 46; SST neurons, 46; and NPY neurons, 56. Inclusion of cells in clusters was guided entirely by amplicon representation in each cell without using spatial information; excitatory cell clusters were then named according to the spatial layering observed for that cluster, whereas inhibitory cell clusters were named according to the dominant cell-type amplicon based on the strong segregation of amplicon markers.

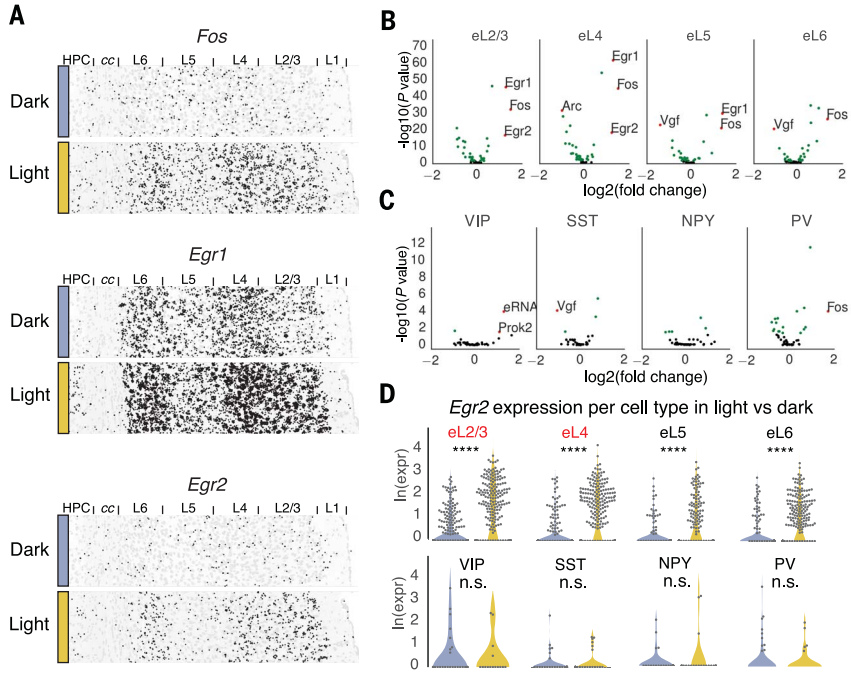


Fig. 3. STARmapping behavioral experience: Detecting and quantifying cell type–specific regulation of ARGs. (**A**) Validation. Shown is spatial expression pattern in the visual cortex of prototypical ARGs known as immediate early genes. Sacrifice was in darkness or after 1 hour of light exposure. (**B** and **C**) Volcano plots of log fold-change in gene expression between light and dark conditions in inhibitory and excitatory cell types. Genes with significantly increased or decreased expression (false discovery rate–adjusted $P < 0.05$, Wilcoxon rank-sum test) are labeled in green, and the most significantly changed genes ($P < 0.05$ and fold change > 2) are labeled in red. Many ARGs showed cell-type specificity, pointing to discovery of unanticipated cell type–specific logic of excitation-transcription coupling. (**D**) Violin plot of *Egr2* expression by cell type. ****$P < 0.0001$, n.s. not significant, Wilcoxon rank-sum test; red-labeled cell types, fold change >2.

high-resolution imaging and sample stability (*16*) in comparison with sequencing-by-ligation methods that can be implemented at room temperature. However, none of the reported or commercially available sequencing-by-ligation methods exhibit the necessary SNR or accuracy for this challenging intact-tissue application: Supported Oligo Ligation Detection (SOLiD) sequencing causes strong background fluorescence in biological samples (*10*), whereas combinatorial probe anchor ligation (cPAL) sequencing (*22*) lacks an error-rejection mechanism (fig. S3). For this reason, an approach we term sequencing with error-reduction by dynamic annealing and ligation (SEDAL) was devised specifically for STARmap (fig. S3).

SEDAL uses two kinds of short, degenerate probes: reading probes to decode bases, and fluorescence probes to transduce decoded sequence information into fluorescence signals. The two short probes only transiently bind to the target DNA and ligate to form a stable product for imaging only when a perfect match occurs; after each cycle corresponding to a base readout, the fluorescent products are stripped with formamide, which eliminates error accumulation as sequencing proceeds (Fig. 1E and fig. S3B). In contrast to SOLiD, SEDAL exhibits minimal background (fig. S3, C to F). A two-base encoding scheme was designed and implemented in order to mitigate any residual errors related to imaging high densities of spots (fig. S3, G and H). On the basis of a panel of four very highly expressed test genes in mouse brain (to mimic amplicon crowdedness as would be encountered in highly multiplexed gene-detection), we found that the error rate of STARmap was more than an order of magnitude lower than prior cPAL methods (~1.8 versus 29.4%) (fig. S3, I to L) (*17*).

Spatial cell typing in primary visual cortex with 160-gene STARmapping

To test whether STARmap could deliver on the initial goal of high-content 3D intact-tissue sequencing of single-cell transcriptional states with the necessary sensitivity and accuracy, we applied STARmap to a pressing challenge in neuroscience: detecting and classifying cell types and corresponding tissue-organization principles in the neocortex of the adult mouse brain. The anatomy and function of the mouse primary visual neocortex have been extensively studied (*23*), a setting which here allows validation of our results by comparison with prior findings that span multiple papers, methodologies, and data sources (but the full diversity of deeply molecularly defined cell types within the visual cortex has not yet been spatially resolved in a single experiment, precluding identification of potentially fundamental joint statistics and organizational principles across 3D volumes). Among many examples of the experimental leverage such information could provide, joint 3D cell-typology mapping might be used to help decode the spatiotemporal logic of neural activity–triggered gene expression as a function of cell type and spatial location.

We therefore used five-base barcoded SNAIL probes over six rounds of in situ SEDAL sequencing in coronal mouse brain slices (Figs. 1A and 2, A and B) to survey a large but focused and curated gene set [160 genes including 112 putative cell-type markers collated from mouse cortical single-cell RNA sequencing (*24*, *25*) and 48 activity-regulated genes (ARGs) (*26*, *27*)]. In one arm of the experiment, visually evoked neural activity was provided to a cohort of mice via 1 hour of light exposure after 4 days of housing in the dark; other mice were kept continuously in the dark (*27*, *28*). Eight-μm-thick volumes containing up to 1000 cells covering all cortical layers were imaged. After six rounds of sequencing, fluorescent Nissl staining was used to segment cell bodies, allowing attribution of amplicons to individual cells (fig. S4, A and B). The values corresponding to amplicons-per-cell and genes-per-cell varied substantially (Fig. 2C), whereas the 160-gene expression pattern was consistent between biological replicates [correlation coefficient (r) = 0.94 to 0.95] (Fig. 2D), revealing reliable detection of transcript diversity at the single-cell level. Because only 160 genes were encoded out of the 1024 possible barcodes from five bases, we were able to quantify sequencing errors that resulted in sequences being corrupted from the 160 true barcodes to the 864 invalid barcodes, which was remarkably low at 1 to 4%. We found that this 160-gene pilot faithfully reproduced the spatial distribution of known cortical layer markers and interneurons, illustrated here via comparison of in situ images from paired public atlases (*29*) and STARmap results (Fig. 2E).

We next performed cell classification using expression data of the 112 cell-type markers. First, >3000 cells pooled from four biological replicates were clustered into three major cell types (excitatory neurons, inhibitory neurons, and non-neuronal cells) by using graph-based clustering after principal-component decomposition (*30*) and then further subclustered under each category (Fig. 2, F to H, and fig. S4C). The richly defined excitatory neurons segregated into four major types (here denoted eL2/3, eL4, eL5, and eL6) by spatial correspondence with anatomic cortical layers and expression profiles of known layer-specific gene markers (Fig. 2, I to K, and fig. S5, A and B). Although spatial organization of the four excitatory types exhibited a layered pattern, there was extensive intermixing among different cell types within each layer. Inhibitory neurons were also clustered into four major types, here denoted by the dominant interneuron marker of each subtype [VIP, SST, NPY, and PV (Pvalb)] (Fig. 2, L to N, and fig. S5,

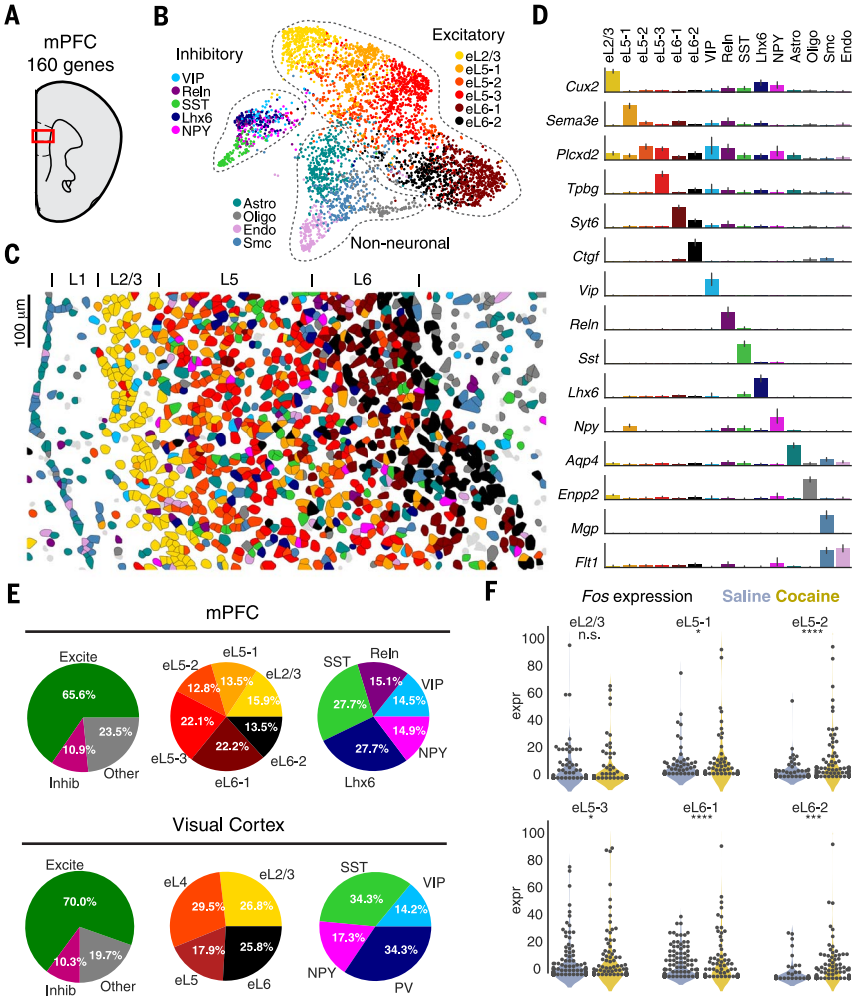


Fig. 4. STARmapping cell types and neural activity in mPFC. (**A**) Diagram of targeted region (red box) containing primarily prelimbic cortex (PrL) within mPFC. (**B**) UMAP visualization of all inhibitory (VIP, Reln, SST, Lhx6, and NPY), excitatory (eL2/3, eL5-1, eL5-2, eL5-3, eL6-1, and eL6-2), and non-neuronal (Astro, Oligo, Smc, and Endo) cell types. (**C**) Spatial visualization of cell type layout in mPFC, using the same color scheme as in (B). (**D**) Barplot of representative genes per cluster (mean ± 95% confidence interval), with each bar scaled to the maximum mean expression for that gene across clusters. (**E**) Piecharts showing the relative proportion of each major and minor cell type in both mPFC and visual cortex. (**F**) Violin plots of *Fos* gene induction in different excitatory cell types in mPFC in response to cocaine. The mice were sacrificed after 1 hour of cocaine or saline injection. Expr, normalized expression; n.s., not significant; *P < 0.05, ***P < 0.001, ****P < 0.0001, likelihood ratio test. Astro, astrocytes; Oligo, oligodendrocytes; Smc, smooth muscle cells; Endo, endothelial cells.

C and D); the VIP and NPY type were observed to distribute more to the upper layers (L1 to L3), whereas SST and PV types were found more commonly in the lower layers (L4 to L6). We also detected non-neuronal cell types, including astrocytes, oligodendrocytes, endothelial cells, and smooth muscle cells (fig. S6). The number of major cell types illustrated here (12 in total) can be further broken down (single-cell RNA sequencing can lead to classification into 40 or more subtypes, which is consistent with the readily apparent heterogeneity of gene expression within each type) (figs. S5 and S6). With our targeted 112-gene set and at the size of 600 to 800 cells per sample, all 12 major cell types could be reliably detected without batch effects with highly similar spatial patterning among four biological replicates (defined as samples prepared from different animals) (fig. S7) and matched with published single-cell RNA sequencing results (fig. S8).

We next sought to take advantage of STARmap's quantitative capabilities at the single-cell level, in order to test differential gene expression analyses across experimental conditions, in molecularly defined cell types. To this end, we assessed visual stimulus–dependent gene expression patterns (via 48 defined ARGs with single-cell resolution in situ). Further developing the single-cell RNA sequencing procedure, mouse brains were flash-frozen with minimal handling time after sacrifice (<5 min), for maximal preservation of native transcriptional signatures. Consistent with prior reports (*26*–*28*), we observed global induction of known immediate-early genes (*Fos*, *Egr1*, and *Egr2*) (Fig. 3A) in the primary visual cortex (V1) upon 1 hour of light exposure. At single-cell resolution, the quantitative extent (fold change in expression) of ARG changes exhibited striking diversity across neuronal cell types (Fig. 3, B and C, and fig. S9) (*28*). In general, ARG expression programs in excitatory neurons across different layers were highly similar, whereas ARG expression programs in inhibitory cells exhibited much more distinct cell type–specific characteristics (fig. S9C); for example, *Egr2* exhibited light-induction across excitatory neurons (Fig. 3D) but not in inhibitory neurons, whereas *Prok2* was up-regulated in *Vip* inhibitory neurons (Fig. 3C) (*22*). Last, because neural activity can trigger cotranscription of noncoding RNAs from within enhancers of ARGs (*26*, *31*), we also studied exemplars of these enhancer RNAs (eRNAs) (here, eRNAs 1 to 5 of the *Fos* gene); these transcripts, not polyadenylated, would be very difficult to measure with current single-cell RNA sequencing. eRNA3 was identified as the most notable and consistent ARG marker (fig. S9B).

Comparing spatial cell-type distributions in frontal and sensory cortices

We then investigated to what extent the cell types of the higher cognitive cortex resemble those of the sensory cortex, as exemplified by V1. We applied the same 160-gene set to STARmapping the cell populations of the medial prefrontal cortex (mPFC) (Fig. 4A), which is involved in high-level cognitive functions such as attention and memory and is thought to be dysregulated in major psychiatric disorders (*32*). We identified 15 distinct molecular cell types, including six excitatory neuron subtypes (eL2/3, eL5-1, eL5-2, eL5-3, eL6-1, and eL6-2, annotated by anatomic cortical layers), five inhibitory neuron subtypes (VIP, Reln, SST, Lhx6 and NPY, annotated by dominant gene markers), and four non-neuronal types (astrocytes, oligodendrocytes, endothelial cells, and smooth-muscle cells) (Fig. 4B and fig. S10).

The spatial organization of broad cell types in mPFC resembled that of V1 with intermixed excitatory neuronal layers and sparsely distributed inhibitory neurons (Fig. 4C); however, the nature and composition of neuronal subtypes in mPFC and V1 strikingly differed (Fig. 4, D and E). For excitatory subtypes, mPFC lacks eL4 (which is consistent with previous reports) (*33*) and exhibits reduced eL2/3 and vast expansion of eL5 and eL6 compared with that of V1 (Fig. 4E). Many new types of cell were discovered, including three eL5 subtypes and two eL6 subtypes, as characterized by gene markers *Sema3e*, *Plcxd2*, *Tpbg*, *Syt6*, and *Ctgf*, respectively (Fig. 4D).

Substantially different tissue organization by cell type was also observed for inhibitory subclusters. *Sst*-, *Vip*-, and *Npy*-positive subtypes in mPFC were represented similarly among all inhibitory neurons compared with those in V1, whereas *Pvalb*-positive cells were comparatively

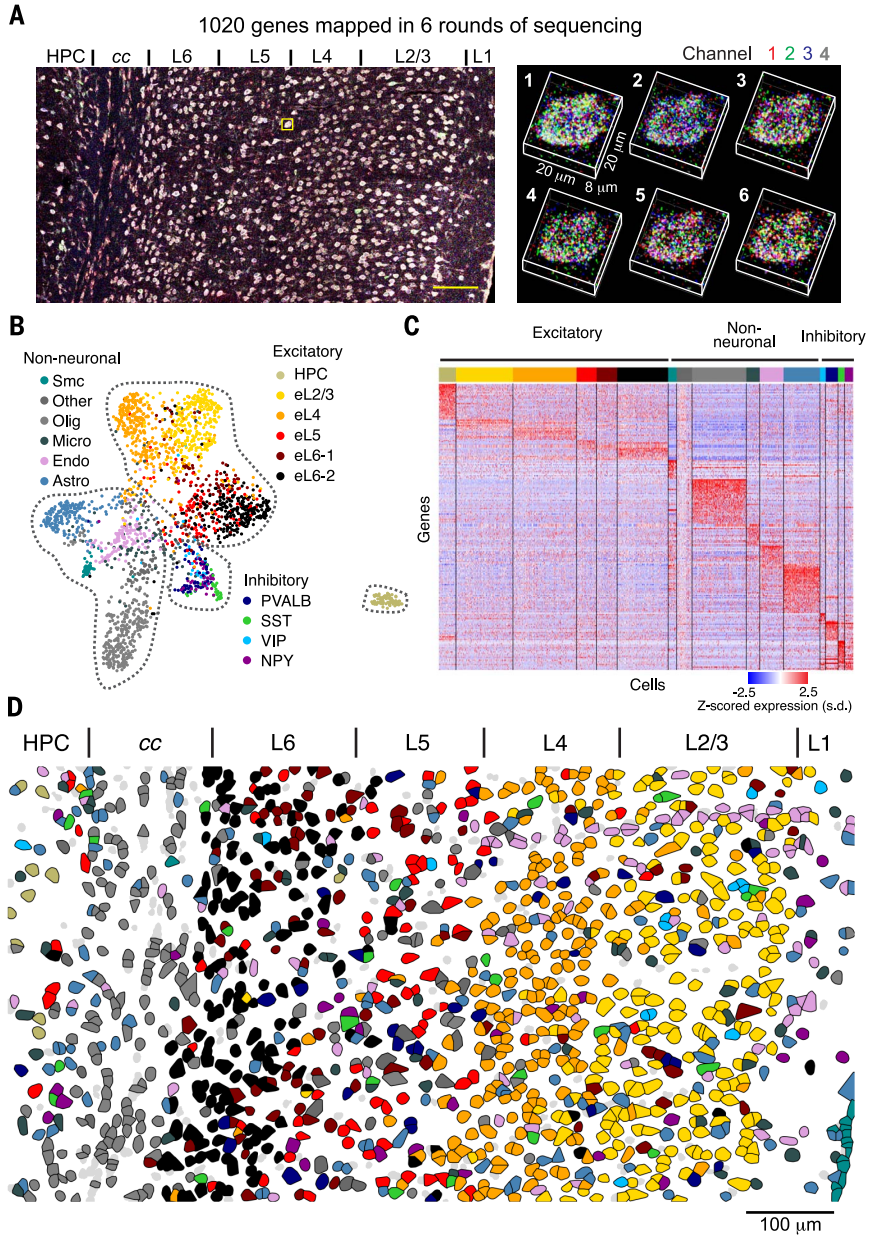


Fig. 5. Simultaneous mapping of 1020 genes in V1 by STARmap. (**A**) Input fluorescence data. (Left) Maximum-intensity projection of the first sequencing round for 1020 gene experiment, showing all four channels simultaneously. Yellow square, zoom region. Scale bar, 100 μm. (Right) Zoom into a single cell showing spatial arrangement of amplicons in three dimensions across six sequencing rounds. (**B**) Joint UMAP plot showing all excitatory (HPC, eL2/3, eL4, eL5, eL6-1, and eL6-2), non-neuronal (Smc, Other, Olig, Micro, Endo, and Astro), and inhibitory (PVALB, SST, VIP, and NPY) cell types. (**C**) Plot of all differentially expressed genes across every cluster, with $P < 10^{-12}$ and log fold change > 1.5. (**D**) Spatial map of all excitatory, non-neuronal, and inhibitory cell types in visual cortex using the same color code of (B). HPC, hippocampus; Smc, smooth muscle cells; Other, other unclassified cells; Oligo, oligodendrocytes; Micro, microglia; Endo, endothelia cells; Astro, astrocytes.

much sparser. In V1, *Reln*-positive neurons co-exist with *Sst* and *Npy*, whereas in mPFC, these segregate as a single cluster, with ~50% co-marked by *Ndnf*; we also discovered a new inhibitory subtype labeled by *Lhx6*, which in fact constitutes the most abundant inhibitory subtype in mPFC (Fig. 4E). Although the 5-HT(3A) receptor (*Htr3a*) expression has been reported in cortical inhibitory neurons (*34*), *Htr3a* has not been ranked as a critical genetic marker of inhibitory subtypes in V1. In mPFC, however, we found that *Htr3a* distinguishably marks a large fraction of Vip^+ neurons and a subset of $Reln^+$ neurons (fig. S10D).

Superficial layers (L1 to L3) were found to contain *Vip*, *Reln*, and *Npy* subtypes, whereas deeper layers (L5 to L6) were found to contain all of the inhibitory subtypes. All of the 15 cell types with tissue-level spatial organization could be reliably detected with STARmap across four biological replicates (fig. S11). The capability of STARmap for multidimensional cell typing in mPFC was further demonstrated in the setting of activity dependence, supporting the possibility of defining cell types in part by communication properties, including activity during behavior (*35*, *36*). One hour after cocaine injection (*37*), a specific subpopulation of deep-layer excitatory neurons (such as *Tpbg* labeled eL5-2) in mPFC was activated compared with that in saline-injected control mice (Fig. 4F), revealing STARmap capability for identifying functional segregation of neuronal subtypes in mPFC.

Scaling STARmap to more than 1000 genes

To further test the scalability of STARmap, we extended our gene list from 160 to 1020 genes, leveraging previously published single-cell RNA sequencing data (*24*). The 1020-gene set was first validated in mouse hippocampal neuron culture, with successful resolution of neuronal and glial cells (fig. S12). We then probed mouse V1 neocortex with the 1020-gene set in order to evaluate performance in spatial cell typing in comparison with that of the 160-gene set. Amplicons obtained in the 1020-gene experiment were much denser in cells as compared with those in 160-gene experiments but were optically resolvable in 3D with high-resolution imaging and postimaging deconvolution (Fig. 5A).

We observed that a higher percentage (40%) of amplicons were filtered out in the 1020-gene experiments by our error-rejection mechanism (fig. S3H) in comparison with the four-gene experiments (20%) (fig. S3L), indicating that a more frequent initial color-misassignment potentially resulted from amplicon merging or optical resolution and further demonstrating the importance of our designed error-rejection mechanism. Crucially, despite the read loss, we successfully clustered single cells of the imaging area into 15 annotated cell types and one unclassified type using 1020 genes and the same data analysis pipeline from the focused 160 gene probe set (Fig. 5, B and C, and fig. S13). Three new cell types were identified in addition to the 12 cell types detected by 160 genes (Fig. 5B): eL6 was resolved into two subtypes (eL6-1 and eL6-2), a previously uncharacterized hippocampal excitatory subtype (HPC) was identified, and microglial cells were cleanly identified with an expansion of non-neuronal cell type markers in the 1020-gene set.

Beyond those advances, the 1020-gene findings also successfully reproduced the cell types (and their spatial patterning) from the 160-gene findings and further allowed discovery of multiple new gene markers for each cell type (for example, *3110035E14Rik* for deep layers, *Cnot6l* for *Sst* neurons, and *Cplx1* for *Pvalb* neurons) (Fig. 5D and fig. S13). These molecularly defined cell types were highly reproducible between biological replicates for 1020-gene detection and were

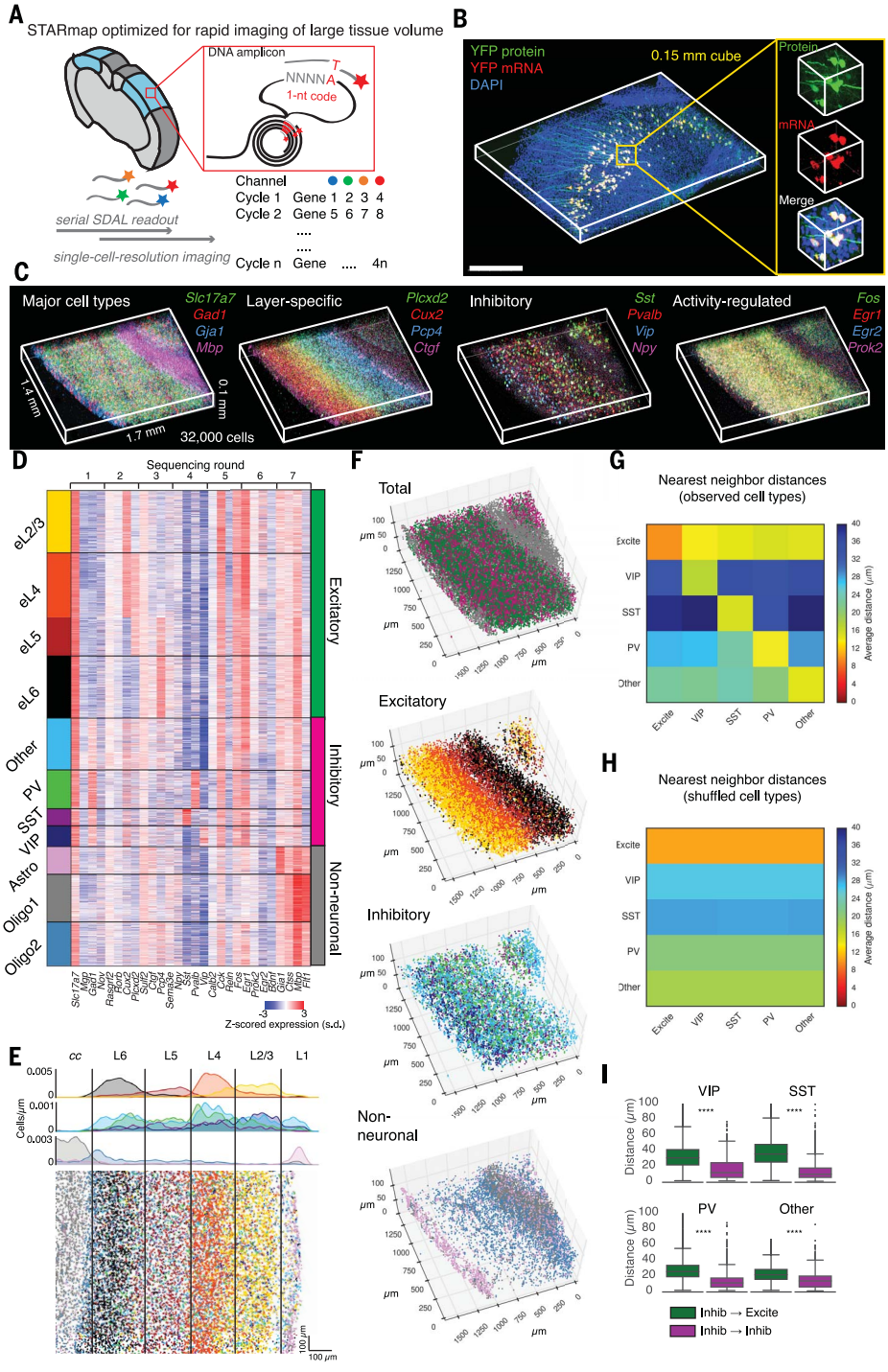


Fig. 6. 3D architecture of cell types in visual cortex volumes. (**A**) Volumetric STARmapping via sequential SEDAL gene readout. Using a modified STARmap procedure (fig. S16) and cyclic gene readout (four genes in each cycle), large tissue volumes can be rapidly mapped at single-cell resolution without oversampling each amplicon. (**B**) Validation showing specific STARMAP labeling of YFP-expressing neurons (from transgenic Thy1::YFP mouse line) in 3D cortical volume. Scale bar, 0.5 mm. (**C**) Representative labeling of (left) major cell types, (left center) layer-specific markers, (right center) inhibitory markers, and (right) activity-regulated genes acquired over multiple rounds in visual cortex STARmap volumes. (**D**) Per-cell expression matrix of 28 genes from 32,845 single cells from one volume clustered into multiple excitatory, inhibitory, and non-neuronal cell types, *z*-scored across genes for each cell in order to normalize for mean differences in total signal between cells. Columns are sorted by order of sequencing rounds as conducted, in groups of four. (**E**) (Top) Spatial histograms of excitatory, inhibitory, and non-neuronal cell types, using same color labels as (D). Cells were counted in 5-μm bins in a 2D max-projection and plotted in cell-count-per-micrometer units as a function of distance from the corpus callosum (cc) to the pia, averaged across the bins perpendicular to the cortical layers. (Bottom) Plot of max-projected cell locations color-coded by cluster as in (D). (**F**) Spatial distribution of each cell type (excitatory, inhibitory, and non-neuronal) and subtypes in three dimensions. Each dot represents a single cell; spatial dimensions are in micrometers. (**G**) Average nearest-neighbor distances computed in three dimensions between all excitatory cells (Excite) and each inhibitory cell type. For self-comparisons, the nearest neighbor was defined as the closest nonidentical cell; persistent self-correlation reveals self-clustering of inhibitory subtypes. (**H**) Same distances as (H) but using shuffled (randomized) cell-type labels. (**I**) Nearest-neighbor distances computed in three dimensions between each inhibitory cell of a certain type and any member of the same type (Inhib → Inhib, eg VIP → VIP) or any excitatory neuron (Inhib → Excite). **** $P < 0.0001$, Wilcoxon rank-sum test.

concordant with published single-cell RNA sequencing results (fig. S14). We further assessed the possibility of scaling up STARmap to accommodate higher gene numbers; although the STARmap scheme can encode and decode more than 1 million codes and the physical volume of mammalian cells is not limiting for amplification of more than 1000 genes (fig. S15), the 1020-gene experiments approached the upper limit of the optical volume of cells (fig. S15E); for those cases in which more genes are needed, STARmap may cover the whole transcriptome with more sequencing rounds of serial 1000-gene detection, or via optical resolution enhanced with super-resolution microscopy (*38*, *39*) or the physical swelling typical of the hydrogel-tissue chemistries (*14*, *19*).

Adapting STARmap to thick tissue blocks for 3D analyses

In neuroscience, addressing the 3D complexity of both neurons and neural circuits has generally required the development and use of thick tissue blocks or fully intact brains for functional and structural readouts, including electrophysiology, imaging of activity, and analysis of morphology and connectivity. Therefore, for linking these readout measures from intact or semi-intact tissue preparations with cellular-resolution gene expression readouts from the very same preparations, methods of 3D spatial transcriptomic analysis in thick tissues have long been sought in order to achieve datastream registration as well as preserve 3D morphology and

to obtain readouts from very much larger cell numbers (*2*). The initial experiments were carried out in brain slices no more than one cell body thick; we therefore next developed and tested STARmap to overcome limitations in diffusional access and imaging throughput for large tissue volumes, with a modified strategy for linearly reading out gene expression at cellular resolution so as to enable high-throughput molecular analysis in tissue volumes (Fig. 6A and fig. S16). Specificity and penetration depth of large-volume STARmap were tested initially by using Thy1::YFP mouse brains, in which STARmap successfully detected yellow fluorescent protein (YFP) mRNA across 150 μm of tissue thickness and specifically colocalized YFP protein and mRNA at single-cell resolution (Fig. 6B) without labeling the tens of thousands of interspersed neighboring cells.

We then extended the spatial cell-typing of mouse V1 to more than 30,000 cells across volumes spanning all six layers and the corpus callosum. Using a curated gene set including 23 cell-type markers and five ARGs read out over seven cycles of linear SEDAL sequencing (Fig. 6, C and D, and fig. S17), we applied *K*-means clustering of marker genes (supplementary materials, materials and methods) for each cell type (recovering 11 cell types corresponding to the majority of those extracted by the 160-gene experiment). We found that 3D patterning of the 11 cell types (Fig. 6, E and F) was consistent with the 160-gene thin-section tissue findings but provided an accurate and quantitative profiling of cellular distribution across space, with much larger cell numbers. As reflected by both spatial-histogram (Fig. 6E) and correlational analyses (fig. S17B), excitatory subtypes exhibited a layered gradient distribution, with the spatial density of each subtype decaying across space into adjacent layers. By contrast, inhibitory subtypes were dispersed, albeit with layer preferences exhibited by the *Vip* subtype (largely located in layer 2/3) and the *Sst* and *Pvalb* subtypes (in layers 4 and 5). Non-neuronal cells were largely seen in layer 1 and white matter.

To discover yet-finer volumetric patterns, we further analyzed the distribution of distances from each individual cell from each sequencing-defined subtype to its nearest neighbors, finding unexpectedly that the nearest neighbor of any inhibitory neuron tended to be its own subtype, rather than excitatory neurons or other inhibitory subtypes (Fig. 6G). If inhibitory neurons were randomly dispersed among the more abundant excitatory neurons in a purely salt-and-pepper distribution, the distance between inhibitory neurons would be larger than that from inhibitory to excitatory neurons (Fig. 6H). Instead, the actual intrasubtype distance of inhibitory neurons was much shorter (~15 μm, which is equivalent to the size of a single neuron, indicating direct somatic juxtaposition) (Fig. 6I), revealing a short-range self-clustering organization of inhibitory subtypes across volumes that could only be accurately measured in three but not in two dimensions (fig. S18A). When guided by this initial STARmap observation, evidence for such patterning could be also obtained in transgenic mouse lines (fig. S18, B and C). This discovery bears considerable relevance to previous functional work; for example, electrophysiological studies have suggested that inhibitory neurons in spatial proximity tend to be connected by electric (gap) junctions important for setting up synchronized firing patterns (*40*, *41*), and in vivo imaging has suggested that inhibitory-neuron groupings in visual cortex could sharpen visual responses (*42*).

Discussion

STARmap defines a platform for 3D in situ transcriptomics, enabled by state-of-the-art DNA library preparation and sequencing in an HTC formulation. Here, STARmap was shown to be applicable to the study of molecularly defined cell types and activity-regulated gene expression in mouse cortex and to be scalable to larger 3D tissue blocks so as to visualize short- and long-range spatial organization of cortical neurons on a volumetric scale not previously accessible. In future work, STARmap may also be adapted to longer sequencing lengths or higher gene numbers; there is no intrinsic limit to the number of genes or RNA species that can be simultaneously and quantitatively accessed with STARmap (fig. S15); STARmap may also be capable of integrating cell-type information with single-neuron morphology and projection anatomy (for example, by means of Brainbow and MAPseq) (*43*, *44*) as well as with in vivo neural activity imaging and electrophysiology. This platform can also be generalized to study other heterogeneous cell populations in diverse tissues across the body, although the brain poses special challenges well suited to STARmap analysis. For example, the polymorphic ARG expression observed across different cell types is likely to depend on both intrinsic cell-biological properties (such as signal transduction pathway-component expression) and on extrinsic properties such as neural circuit anatomy that routes external sensory information to different cells (here, in visual cortex). In general, it may not be possible to fully define brain cell typology independent of such 3D anatomy as well as activity patterns exhibited and experienced by cells during behavior; the nature of input and output communication pathways for the cells in question in fact can form the foundation for defining cell types (*35*, *36*). Toward this end, in situ transcriptomics exemplified by STARmap can effectively link this imaging-based molecular information with complementary cellular-resolution datastreams describing anatomy, natural activity, and causal importance, thus promising to fundamentally deepen our understanding of brain function and dysfunction (*2*).

Methods summary

All animal procedures followed animal care guidelines approved by Stanford University's Administrative Panel on Laboratory Animal Care (APLAC) and guidelines of the National Institutes of Health. For thin sections, animals were anesthetized and rapidly decapitated; the brain tissues were sliced by use of a cryostat. For thick sections, animals were anesthetized and transcardially perfused with paraformaldehyde; the brain tissues were sliced by use of a vibratome. In STARmap experiments, tissues were hybridized with SNAIL probes, enzymatically amplified, hydrogel embedded, and sequentially imaged by using the SEDAL process and a confocal microscope. The resulting image datasets were registered across multiple cycles by using the positions of all amplicons in each cycle and decoded. For cell typing and single-cell gene expression analyses, the amplicons were attributed to individual cells based on segmentation images of fluorescent Nissl staining. All the detailed procedures for the experiments and data analyses are described in the supplementary materials.

REFERENCES AND NOTES

1. N. Crosetto, M. Bienko, A. van Oudenaarden, Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.* **16**, 57–66 (2015). doi: 10.1038/nrg3832; pmid: 25446315
2. E. Lein, L. E. Borm, S. Linnarsson, The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017). doi: 10.1126/science.aan6827; pmid: 28983044
3. E. Lubeck, L. Cai, Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9**, 743–748 (2012). doi: 10.1038/nmeth.2069; pmid: 22660740
4. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015). doi: 10.1126/science.aaa6090; pmid: 25858977
5. S. Shah, E. Lubeck, W. Zhou, L. Cai, *In situ* transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016). doi: 10.1016/j.neuron.2016.10.001; pmid: 27764670
6. R. Ke *et al.*, In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013). doi: 10.1038/nmeth.2563; pmid: 23852452
7. J. H. Lee *et al.*, Highly multiplexed subcellular RNA sequencing *in situ*. *Science* **343**, 1360–1363 (2014). doi: 10.1126/science.1250212; pmid: 24578530
8. A. P. Frei *et al.*, Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* **13**, 269–275 (2016). doi: 10.1038/nmeth.3742; pmid: 26808670
9. X. Chen, Y.-C. Sun, G. M. Church, J. H. Lee, A. M. Zador, Efficient *in situ* barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* **46**, e22 (2018). doi: 10.1093/nar/gkx1206; pmid: 29190363
10. M. Nagendran, D. P. Riordan, P. B. Harbury, T. J. Desai, Automated cell-type classification in intact tissues by single-cell molecular profiling. *eLife* **7**, e30510 (2018). doi: 10.7554/eLife.30510; pmid: 29319504
11. N. A. Peppas, J. Z. Hilt, A. Khademhosseini, R. Langer, Hydrogels in biology and medicine: From molecular principles to bionanotechnology. *Adv. Mater.* **18**, 1345–1360 (2006). doi: 10.1002/adma.200501612
12. A. M. Rosales, K. S. Anseth, The design of reversible hydrogels to capture extracellular matrix dynamics. *Nat. Rev. Mater.* **1**, 1–15 (2016). pmid: 29214058
13. R. Y. Tam, L. J. Smith, M. S. Shoichet, Engineering cellular microenvironments with photo- and enzymatically responsive hydrogels: Toward biomimetic 3D cell culture models. *Acc. Chem. Res.* **50**, 703–713 (2017). doi: 10.1021/acs.accounts.6b00543; pmid: 28345876
14. V. Gradinaru, J. Treweek, K. Overton, K. Deisseroth, Hydrogel-tissue chemistry: Principles and applications. *Annu. Rev. Biophys.* **47**, 355–376 (2018). doi: 10.1146/annurev-biophys-070317-032905; pmid: 29792820
15. K. Chung *et al.*, Structural and molecular interrogation of intact biological systems. *Nature* **497**, 332–337 (2013). doi: 10.1038/nature12107; pmid: 23575631
16. E. L. Sylwestrak, P. Rajasethupathy, M. A. Wright, A. Jaffe, K. Deisseroth, Multiplexed intact-tissue transcriptional analysis

at cellular resolution. *Cell* **164**, 792–804 (2016). doi: 10.1016/j.cell.2016.01.038; pmid: 26871636
17. S. Shah *et al*., Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. *Development* **143**, 2862–2867 (2016). doi: 10.1242/dev.138560; pmid: 27342713
18. J. R. Moffitt *et al*., High-performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14456–14461 (2016). doi: 10.1073/pnas.1617699113; pmid: 27911841
19. F. Chen *et al*., Nanoscale imaging of RNA with expansion microscopy. *Nat. Methods* **13**, 679–684 (2016). doi: 10.1038/nmeth.3899; pmid: 27376770
20. J. Shendure *et al*., Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005). doi: 10.1126/science.1117389; pmid: 16081699
21. J. H. Lee *et al*., Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–458 (2015). doi: 10.1038/nprot.2014.191; pmid: 25675209
22. R. Drmanac *et al*., Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010). doi: 10.1126/science.1181498; pmid: 19892942
23. L. L. Glickfeld, R. C. Reid, M. L. Andermann, A mouse model of higher visual cortical function. *Curr. Opin. Neurobiol.* **24**, 28–33 (2014). doi: 10.1016/j.conb.2013.08.009; pmid: 24492075
24. B. Tasic *et al*., Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016). doi: 10.1038/nn.4216; pmid: 26727548
25. A. Zeisel *et al*., Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015). doi: 10.1126/science.aaa1934; pmid: 25700174
26. T. K. Kim *et al*., Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010). doi: 10.1038/nature09033; pmid: 20393465
27. A. R. Mardinly *et al*., Sensory experience regulates cortical inhibition by inducing IGF1 in VIP neurons. *Nature* **531**, 371–375 (2016). doi: 10.1038/nature17187; pmid: 26958833
28. S. Hrvatin *et al*., Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018). doi: 10.1038/s41593-017-0029-5; pmid: 29230054
29. E. S. Lein *et al*., Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007). doi: 10.1038/nature05453; pmid: 17151600
30. K. Shekhar *et al*., Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323.e30 (2016). doi: 10.1016/j.cell.2016.07.054; pmid: 27565351
31. J. Y. Joo, K. Schaukowitch, L. Farbiak, G. Kilaru, T.-K. Kim, Stimulus-specific combinatorial functionality of neuronal *c-fos* enhancers. *Nat. Neurosci.* **19**, 75–83 (2016). doi: 10.1038/nn.4170; pmid: 26595656
32. M. Carlén, What constitutes the prefrontal cortex? *Science* **358**, 478–482 (2017). doi: 10.1126/science.aan8868; pmid: 29074767
33. H. B. Uylings, H. J. Groenewegen, B. Kolb, Do rats have a prefrontal cortex? *Behav. Brain Res.* **146**, 3–17 (2003). doi: 10.1016/j.bbr.2003.09.028; pmid: 14643455
34. S. Lee, J. Hjerling-Leffler, E. Zagha, G. Fishell, B. Rudy, The largest group of superficial neocortical GABAergic interneurons expresses ionotropic serotonin receptors. *J. Neurosci.* **30**, 16796–16808 (2010). doi: 10.1523/JNEUROSCI.1869-10.2010; pmid: 21159951
35. A. Paul *et al*., Transcriptional architecture of synaptic communication delineates GABAergic neuron identity. *Cell* **171**, 522–539.e20 (2017). doi: 10.1016/j.cell.2017.08.032; pmid: 28942923
36. T. N. Lerner, L. Ye, K. Deisseroth, Communication in neural circuits: Tools, opportunities, and challenges. *Cell* **164**, 1136–1150 (2016). doi: 10.1016/j.cell.2016.02.027; pmid: 26967281
37. L. Ye *et al*., Wiring and molecular features of prefrontal ensembles representing distinct experience. *Cell* **165**, 1776–1788 (2016). doi: 10.1016/j.cell.2016.05.010; pmid: 27238022
38. M. J. Rust, M. Bates, X. Zhuang, Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006). doi: 10.1038/nmeth929; pmid: 16896339
39. E. Betzig *et al*., Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006). doi: 10.1126/science.1127344; pmid: 16902090
40. J. R. Gibson, M. Beierlein, B. W. Connors, Two networks of electrically coupled inhibitory neurons in neocortex. *Nature* **402**, 75–79 (1999). doi: 10.1038/47035; pmid: 10573419
41. Y. Amitai *et al*., The spatial dimensions of electrically coupled networks of interneurons in the neocortex. *J. Neurosci.* **22**, 4142–4152 (2002). doi: 10.1523/JNEUROSCI.22-10-04142.2002; pmid: 12019332
42. T. Ebina *et al*., 3D clustering of GABAergic neurons enhances inhibitory actions on excitatory neurons in the mouse visual cortex. *Cell Reports* **9**, 1896–1907 (2014). doi: 10.1016/j.celrep.2014.10.057; pmid: 25464846
43. J. Livet *et al*., Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56–62 (2007). doi: 10.1038/nature06293; pmid: 17972876
44. J. M. Kebschull *et al*., High-throughput mapping of single-neuron projections by sequencing of barcoded RNA. *Neuron* **91**, 975–987 (2016). doi: 10.1016/j.neuron.2016.07.036; pmid: 27545715

ACKNOWLEDGMENTS

We are grateful for the efforts and support of Deisseroth laboratory members A. Crow, S. Quirin, A. Chibukhchyan, C. Lee, M. Lo, and C. Delacruz. We thank M. Lovett-Barron for helping with image registration and J. H. Lee (Cold Spring Harbor Laboratory), J. Yan (Princeton University), N. D. Donoghue (Brown University), and Q. Dai (University of Chicago) for advice. We thank Deisseroth laboratory members L. Ye, A. Andalman, H. Wang, and B. Hsueh for discussion. We also thank L. Luo for suggestions on the manuscript. **Funding:** X.W. is supported by a Life Science Research Foundation fellowship and the Gordon and Betty Moore Foundation. W.E.A. is supported by a Fannie and John Hertz Foundation Fellowship and an NSF Graduate Research Fellowship. E.L.S. is supported by a NIMH Ruth L. Kirschstein NRSA fellowship (1F32MH110144-01). M.A.W. is supported by an NIMH Career Development Award (1K08MH113039). J.L. is supported by a Bio-X Interdisciplinary Initiatives Seed Grant. G.P.N., N.S., and F.-A. B. were supported by the Parker Institute for Cancer Immunotherapy, the FDA, and the NIH; F.-A.B. was supported by the Human Frontiers Science Program. K.D. is supported by NIMH (R01MH099647), NIDA (P50DA042012), the DARPA NeuroFAST program W911NF-14-2-0013, the NSF NeuroNex program, the Gatsby Foundation, the AE Foundation, the NOMIS Foundation, the Fresenius Foundation, the Wiegers Family Fund, the James Grosfeld Foundation, the Sam and Betsy Reeves Foundation, and the H. L. Snyder Foundation. **Author contributions:** X.W. and K.D. initiated the STARmap project to integrate HTC with in situ sequencing. X.W. developed the STARmap HTC, SEDAL sequencing, and STARmap hardware, designed all the DNA probes, and conducted the experiments. W.E.A. developed the STARmap software pipeline and analyzed the sequencing data. N.S., F.-A.B., and G.P.N. designed the initial versions of the SNAIL system. X.W. and F.-A.B. developed the SNAIL process for mouse brain tissue and compared SNAIL with other in situ methods. M.A.W. and E.L.S. conducted animal behavior and preparation of mouse brain tissue and contributed valuable advice. M.A.W. compared Nissl staining with other segmentation methods. W.E.A. and E.L.S. designed the visual-stimulus and cocaine-stimulus procedures. S.V. generated the CLARITY data with PV transgenicmice. K.E. and C.R. contributed to preparation of cell cultures. C.L. assisted with experiments. J.L. assisted with experiments and graphic design. K.D. supervised all aspects of the work. X.W., W.E.A., and K.D. interpreted the STARmap data, designed and prepared the figures, and wrote the manuscript, with edits from all authors. **Competing interests:** The design, steps, and applications of STARmap are covered in pending patent application material from Stanford University; all methods, protocols, and sequences are freely available to nonprofit institutions and investigators. **Data and materials availability:** All data are available in the main text or the supplementary materials. Computational tools and code and other materials are available at http://clarityresourcecenter.org. References for the open-source software packages used here can be found in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/eaat5691/suppl/DC1
Materials and Methods
Figs. S1 to S18
References (*45*–*49*)
Tables S1 and S2

13 March 2018; resubmitted 1 June 2018
Accepted 13 June 2018
10.1126/science.aat5691

RESEARCH ARTICLE

ORGANIC CHEMISTRY

Teaching an old carbocation new tricks: Intermolecular C–H insertion reactions of vinyl cations

Stasik Popov[1], Brian Shao[1], Alex L. Bagdasarian[1], Tyler R. Benton[1], Luyi Zou[1,2], Zhongyue Yang[1]*, K. N. Houk[1]†, Hosea M. Nelson[1]†

Vinyl carbocations have been the subject of extensive experimental and theoretical studies over the past five decades. Despite this long history in chemistry, the utility of vinyl cations in chemical synthesis has been limited, with most reactivity studies focusing on solvolysis reactions or intramolecular processes. Here we report synthetic and mechanistic studies of vinyl cations generated through silylium–weakly coordinating anion catalysis. We find that these reactive intermediates undergo mild intermolecular carbon-carbon bond–forming reactions, including carbon-hydrogen (C–H) insertion into unactivated sp^3 C–H bonds and reductive Friedel-Crafts reactions with arenes. Moreover, we conducted computational studies of these alkane C–H functionalization reactions and discovered that they proceed through nonclassical, ambimodal transition structures. This reaction manifold provides a framework for the catalytic functionalization of hydrocarbons using simple ketone derivatives.

For more than a century, carbocations have played a central role in the chemical sciences, inspiring the development of broadly applied chemical reactions and a greater understanding of the fundamental properties of molecules (*1*). This is pointedly true for vinyl cations. Beginning with Jacobs' proposal of a vinyl cation intermediate (*2*) and continuing through the investigations by Grob *et al.* (*3*), Hanack (*4*), Rappoport and Gal (*5*), Schleyer and others (*6*), Stang and Rappoport (*7*), and others (*8*–*11*), this class of reactive intermediates has been the subject of extensive theoretical and experimental studies. The structure and bonding of these cations has fascinated theorists, driving the development of computational techniques over the past five decades (*12*–*15*). Vinyl cation intermediates have also been implicated in modern transition metal-catalyzed processes (*16*) and in classical acid-catalyzed reactions (*17*). Despite this rich history and broad scientific relevance, most reactivity studies have focused on solvolysis reactions where the reactive vinyl cation is intercepted by heteroatom-containing solvent molecules (*18*, *19*).

Recently, our group reported that putative phenyl cations (**1**; Fig. 1A), a related class of dicoordinated carbocations, undergo insertion into sp^3 C–H bonds (*20*). Inspired by this result, we hypothesized that vinyl cations may display analogous C–H insertion reactivity. Challenging this hypothesis was a lack of substantial precedent for this mode of reactivity, despite the extensive experimental studies of vinyl cations over the past five decades. However, a few examples of vinyl cations engaging in C–C bond–forming reactions have been reported. Hanack and others reported the formation of fused products **2** and **3** from the solvolysis of cyclononenyl triflate **4** (Fig. 1B) (*21*, *22*). In this reaction, and in analogous reports

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA. [2]Laboratory of Theoretical and Computational Chemistry, Institute of Theoretical Chemistry, Jilin University, Changchun 130023, People's Republic of China. *Present address: Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. †Corresponding author. Email: hosea@chem.ucla.edu (H.M.N.); houk@chem.ucla.edu (K.N.H.)

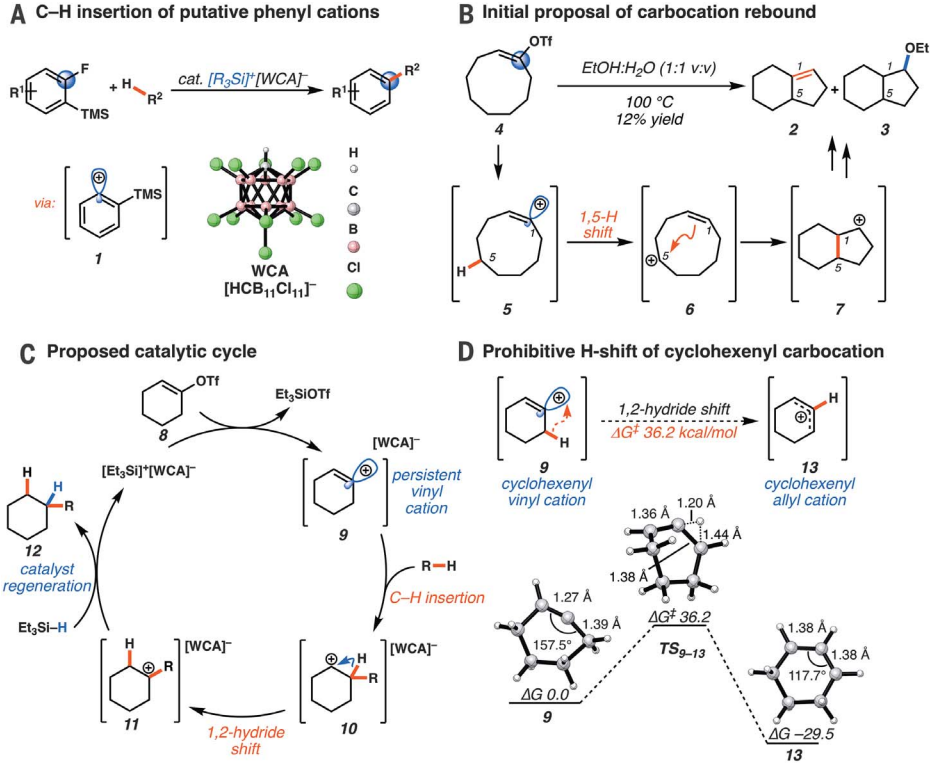


Fig. 1. C–C bond–forming reactions of dicoordinated carbocations. (**A**) Phenyl cation insertion into sp^3 C–H bonds. (**B**) Initial proposal of carbocation rebound mechanism in ring-contraction reactions of cyclononenyl triflates. (**C**) Proposed catalytic cycle for reductive alkane alkylation. (**D**) Identification of cyclohexenyl triflate as a persistent vinyl cation precursor owing to the prohibitive barrier to 1,2-hydride shift. Calculations performed at M062X/6-311+G(d,p) level of theory; $[WCA]^-$ omitted for clarity. R, alkyl; R^1, aryl, alkyl, halide, or silyl ether; R^2, alkyl or aryl; cat., catalytic; TMS, trimethylsilyl; OTf, trifluoromethanesulfonate; Et, ethyl; EtOH, ethanol; ΔG, Gibbs free energy; $\Delta G^{\ddagger}$, Gibbs free energy of activation.

from Olah and Mayr (*23*) and Caple and others (*24*), this unusual C–C bond–forming process was rationalized via a carbocation rebound mechanism. Here the vinyl cation **5** is quenched via 1,5-hydride shift, and the neutralized alkene (**6**) attacks the newly formed alkyl carbocation. This sequence forges the ring fusion of the bicyclononane cation **7**, leading to products **2** and **3**. The insertion of vinyl cations into methane C–H bonds has also been observed in the gas phase (*25*). More recently, Metzger and others (*26*) and Brewer and others (*27*) have independently proposed the intramolecular C–H insertion of vinyl cations generated from alkynes or diazo compounds in solution.

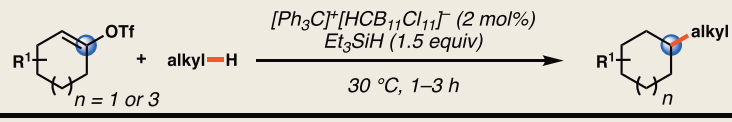
The solvolysis studies described above relied on polar solvents to facilitate thermal ionization of vinyl species bearing halide or pseudohalide leaving groups (for example, Fig. 1B). This approach almost invariably leads to attack of solvent on the cationic carbon center, leaving reactivity beyond the S_N1 manifold largely undiscovered. We drew on our phenyl cation studies to overcome this limitation by using silylium–monocarba-*closo*-dodecaborate salts (formed in situ through the reaction of triethylsilane with $[Ph_3C]^+[HCB_{11}Cl_{11}]^-$) in hydrophobic solvents to generate vinyl cations. Here the nonnucleophilic and nonbasic properties of weakly coordinating anions (WCA), such as undecachlorinated monocarba-*closo*-dodecaborate anion ($HCB_{11}Cl_{11}^-$) (*28*), would enhance the Lewis acidity of the cationic silicon center, allowing for mild ionization of alkene derivatives. In this scenario, silylium-mediated ionization of a vinyl triflate (**8**) would generate a kinetically persistent vinyl cation–WCA ion pair (**9**; Fig. 1C). Insertion of this reactive dicoordinated cation (**9**) into an alkane C–H bond would lead to formation of alkyl carbocation **10**. A 1,2-hydride shift would lead to the more stable tertiary cation (**11**) that, upon reduction by a sacrificial silane (*29*), would generate the functionalized hydrocarbon product **12** and regenerate the silylium-carborane initiator. In its entirety, this process would enable the direct C–H alkylation of alkanes and arenes by simple ketone derivatives. We now report the successful execution of this mechanistic hypothesis, entailing intermolecular, reductive C–H alkylation mediated by silane and catalytic quantities of carborane salts. Moreover, we describe the unexpected discovery of an ambimodal transition state and post–transition state bifurcation in this reaction that leads to several products directly through a nonclassical mechanism (*30*–*32*). Ambimodal transition states have characteristics of two or more classical transition states, just as nonclassical ions simultaneously possess the features of two classical ions.

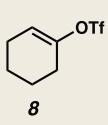
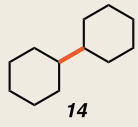
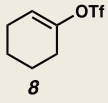
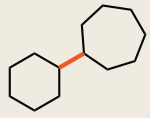
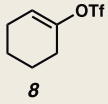
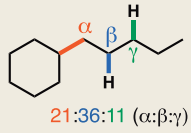
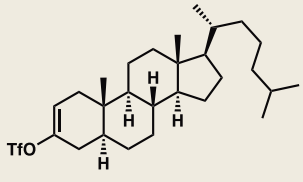
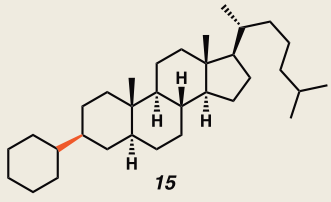
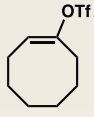
Vinyl cation generation from cyclohexenyl triflates

Given the known propensity of substituted cyclic vinyl cations to undergo rapid intramolecular rearrangements (*33*), we identified prospective cation precursors through density functional theory calculations. First, we verified that a 1,2-hydride shift in cyclohexenyl cation **9**, which would generate the more stable cyclic allyl cation **13**, is kinetically unfavorable (Fig. 1D). The calculated high barrier, despite the impressive exothermicity of the rearrangement, results from the geometrical distortion required to achieve overlap of the migrating H 1s orbital with the nearly orthogonal orbitals at the origin and terminus of the migration. On the basis of this computational result, we initiated our studies with cyclohexenyl triflate (**8**). We were pleased to find that exposure of cyclohexenyl triflate (**8**) to 1.5 equivalents of triethylsilane and 2 mole % (mol %) $[Ph_3C]^+[HCB_{11}Cl_{11}]^-$ in dried cyclohexane solvent at 30°C in a nitrogen-filled glove box resulted in the formation of cyclohexylcyclohexane (**14**; Table 1, entry 1) in 87% yield, after 90 min.

Intrigued by the remarkably mild conditions used in this alkane alkylation reaction, we undertook a brief study of scope to further elucidate potential synthetic applications and to gain mechanistic insight. Cycloheptane and *n*-pentane also reacted efficiently with the cyclohexenyl cation, albeit with poor regioselectivity in the latter case (Table 1, entries 2 and 3). Although cyclohexenyl triflates bearing substituents at the carbons in the 2- or 6-position led to complex mixtures of products, presumably the result of nonproductive unimolecular decompositions (*33*), other positions of the cyclohexenyl ring were tolerant of substitution. For example, exposure of the enol triflate derived from 5α-cholestan-3-one (Table 1, entry 4) to our reaction conditions led to formation of the alkylated steroid **15** in 88% yield and 15:1 diastereomeric ratio (d.r.) (relative stereochemistry determined by x-ray crystallography). Analogous to the previously reported ring-contraction reactions of medium-sized cyclic vinyl triflates (*21*), exposure of cyclooctenyl triflate to our optimized reaction conditions led to rapid transannular C–H insertion to yield bicyclo[3.3.0]octane (Table 1, entry 5).

Table 1. Reductive alkylation of cyclic vinyl triflates. Reactions performed at 0.1 M in hydrocarbon solvent. *Yield determined by GC-FID with nonane as an internal standard. †Isolated yield. R^1, alkyl.



Entry	Substrate	Solvent	Product	Yield (%), Time (h)
1	 *8*	C_6H_{12}	 *14*	87*, 1.5
2	 *8*	C_7H_{14}		88*, 2
3	 *8*	*n*-C_5H_{12}	 21:36:11 (α:β:γ)	68*, 1.5
4		C_6H_{12}	 *15*	88†, 3 (15:1 *d.r.*)
5		C_6H_{12}		91*, 1

Evidence of a nonclassical mechanism

To probe the possibility of a hydride-abstraction, carbocation-rebound mechanism as proposed in early examples (*22*) (Fig. 1B), we studied the insertion reactions of unsymmetrical cyclohexenyl triflate derivatives **16** and **17** (Fig. 2A). If a stepwise rebound mechanism were operative, we would expect 4,4-dimethylcyclohexenyl triflate (**16**) and 5,5-dimethylcyclohexenyl triflate (**17**) to yield identical product distributions, as they would both proceed through alkene intermediate

18 (Fig. 2A). In the event, we found that exposure of triflate **16** to our reaction conditions led to formation of C1-cyclohexylated product **19** in 68% yield and C2-cyclohexylated product **20** in 6% yield. However, triflate **17**, the 5,5-disubstituted analog, provided C1-cyclohexylated product **20** in 78% yield and C2-cyclohexylated product **19** in 6% yield. Although the C–H insertion event occurred primarily at the C1 that originally bore the triflate group, in both systems, minor products were observed that correspond to insertion at the C2 adjacent to the triflate (>10:1 C1:C2 insertion in both cases). Although these experimental results support a nonrebound, concerted C–H insertion mechanism, they do not provide mechanistic insight into the formation of C2-alkylation products. Therefore, we undertook quantum mechanical calculations and molecular dynamics (MD) simulations to probe the mechanism in more detail (movies S1 to S4).

Calculations on the reaction of the cyclohexen-1-yl (**9**) cation with cyclohexane gives a very early ambimodal transition structure (TS), wherein the

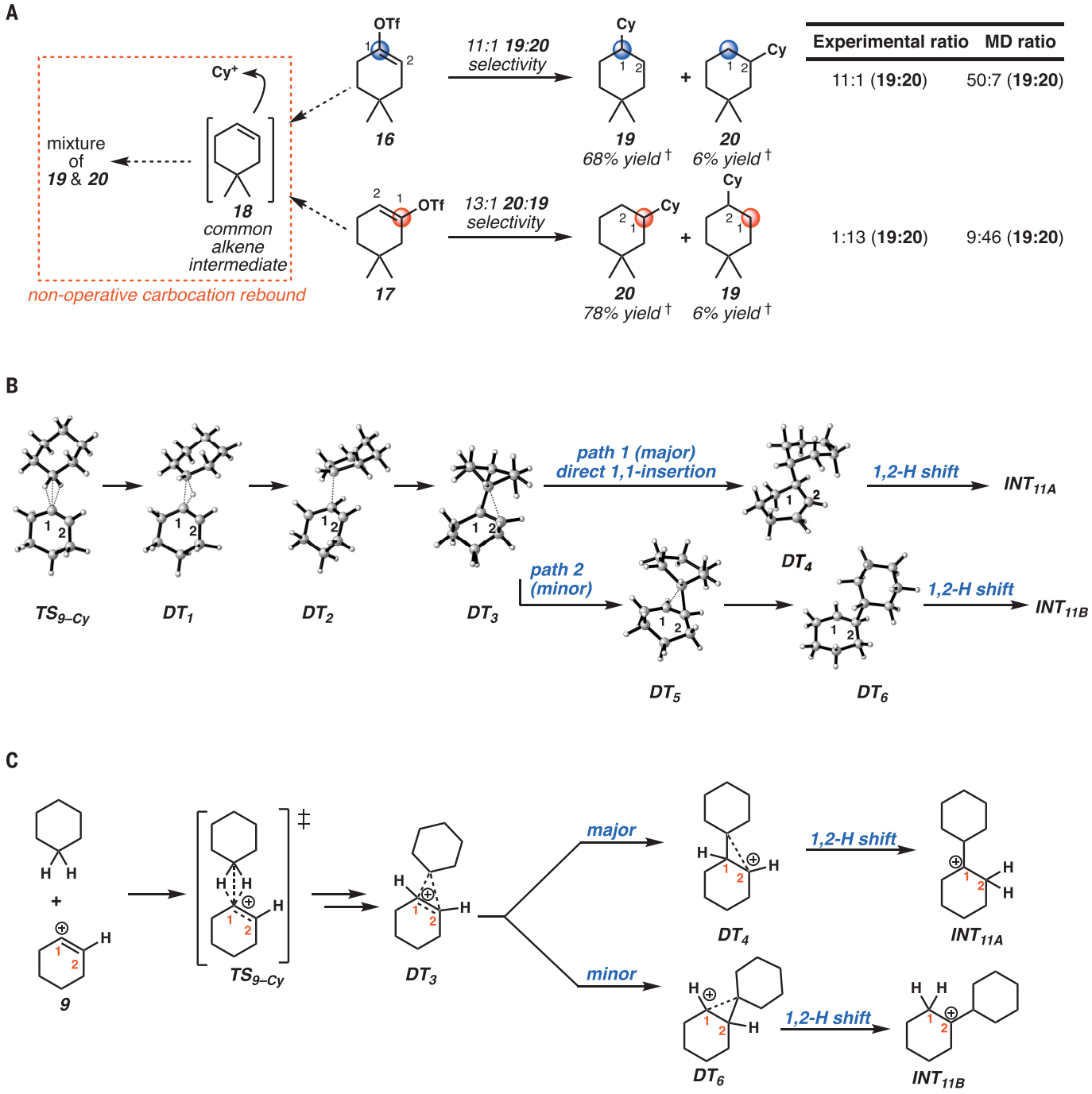
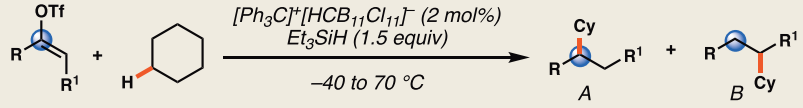
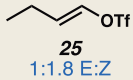


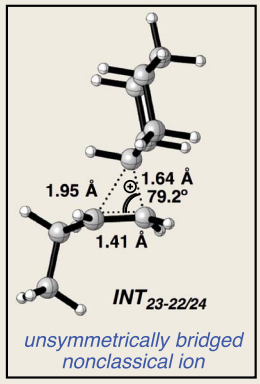
Fig. 2. Discovery of an ambimodal transition state. (**A**) Probing the possibility of a rebound mechanism. Reactions were performed under the same conditions as described in Table 1. (**B**) Snapshots from MD simulations representing structures taken from multiple trajectories starting from the ambimodal $\mathbf{TS_{9\text{-}Cy}}$, illustrating the most common reaction pathways, the post–transition state bifurcation, and resulting products. (**C**) Line drawing representations of the three-dimensional representations in (B). †Yield determined by gas chromatography–flame ionization detector (GC-FID) with nonane as an internal standard. Calculations performed at M062X/6-311+G(d,p) level of theory. Cy, cyclohexyl.

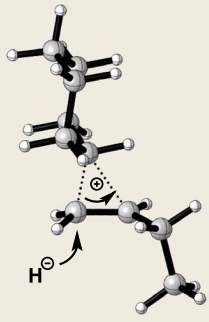
geminal cyclohexane C–H bonds interact with the vinyl cation vacant orbital ($\mathbf{TS_{9-Cy}}$, Fig. 2B). Quasi-classical MD trajectories were initiated from normal mode–sampled structures (a Boltzmann distribution) in the region of transition structure $\mathbf{TS_{9-Cy}}$. Taken together, these represent the ambimodal transition state, a single saddle-point region that can give rise to several different products (*30*–*32*). Two typical trajectories are summarized in Fig. 2B, with several snapshots of structures illustrating the bifurcation of the reaction pathway. The common trajectory path from ambimodal $\mathbf{TS_{9-Cy}}$ begins with hydride transfer from cyclohexane to the vinyl cation via $\mathbf{DT_1}$ (where DT stands for dynamical transient, not an intermediate or transition state) (Fig. 2B). Unlike a rebound mechanism, the cyclohexene and cyclohexyl cation moieties fail to separate; the newly formed cyclohexyl cation coordinates ($\mathbf{DT_2}$) with the alkene of the neutralized cyclohexene, in most trajectories, at the carbon that previously bore the cation (C1) (Fig. 2B). The resulting bridging nonclassical species ($\mathbf{DT_3}$) continues to different products through different trajectories (Fig. 2B). $\mathbf{DT_3}$ is an unsymmetrical nonclassical carbocation, with the cyclohexyl bound more strongly at C1, but still provides the three-center two-electron hypervalent bonding (very strong hyperconjugation) to stabilize the carbocation. In most of the trajectories, $\mathbf{DT_3}$ collapses to the classical secondary carbocation structure, $\mathbf{DT_4}$, that undergoes a barrierless 1,2-hydride shift to form the tertiary cyclohexylcyclohexyl cation as a potential energy minimum at –40.9 kcal/mol ($\mathbf{INT_{11A}}$) (movies S2 and S4). In fewer trajectories, $\mathbf{DT_3}$ goes on to the opposite asymmetric nonclassical species $\mathbf{DT_5}$, favoring C2 functionalization (Fig. 2B). $\mathbf{DT_5}$ leads to $\mathbf{DT_6}$ and eventually $\mathbf{INT_{11B}}$ (movies S1 and S3). Through the stabilizing effects and geometric constraints (fig. S40) (*34*) of these persistent nonclassical intermediates, energetically favorable hydride shifts are delayed, allowing for the experimentally observed dynamic distribution of products in our ring-substituted cyclohexen-1-yl cation substrates (Fig. 2A). The free energies of various stationary points on the surface between ambimodal $\mathbf{TS_{9-Cy}}$ and final stable carbocation $\mathbf{INT_{11A}}$ show that minima are very shallow and are, therefore, not equilibrated during the MD trajectories (fig. S39) (*34*).

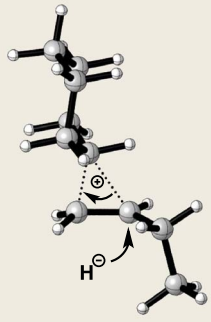
Table 2. Reactions of acyclic vinyl triflates and calculated nonclassical $\mathbf{INT_{23-22/24}}$ for insertion of cyclohexane into butenyl cation. *Reaction performed at 0.1 M and 30°C, with triisopropylsilane used instead of triethylsilane. †Reaction performed at 0.005 M and 30°C. ‡Reaction performed at 0.1 M and 70°C. §Reaction performed at 0.008 M (1:1 v:v), 5 mol % catalyst loading, and –40°C. ‖Yield determined by GC-FID with nonane as an internal standard. R, alkyl or H; R^1, alkyl or H.



Entry	Substrate	Solvent	Temp. (°C)	Product	
1*	 *21* 2.5:1 E:Z	C_6H_{12}	30	 *22*	85‖
2†	 *23*	C_6H_{12}	30	 *22* +  *24*	40, 39‖
3‡	 *25* 1:1.8 E:Z	C_6H_{12}	70	 *22* +  *24*	16, 19‖
4§	 *23*	$CHCl_3/C_6H_{12}$	–40	 *22* +  *24*	17, 34‖







Reactivity of acyclic vinyl triflates

Having established the catalytic, intermolecular C–H alkylation reactions of cyclic vinyl triflates, and having come to a deeper understanding of the nonclassical reactivity of cyclic vinyl cations, we extended our investigations to the reactivities of acyclic vinyl triflates. Acyclic vinyl cations have constituted most of the previous experimental and theoretical studies of dicoordinated cations, as cyclic variants have traditionally proven difficult to generate under solvolytic conditions (*35*). Moreover, depending on substitution, several studies support nonclassical ground-state bonding in vinyl cations, structures reminiscent of protonated alkynes (*36*–*38*). Subjection of butenyl triflate **21** to the reaction conditions led to high-yielding reductive alkylation of cyclohexane, providing 2-cyclohexylbutane (**22**) in 85% yield (Table 2, entry 1). We found that the treatment of a cyclohexane solution of but-1-en-2-yl triflate (**23**) with 2 mol % of $[Ph_3C]^+[HCB_{11}Cl_{11}]^-$ and 1.5 equivalents of triethylsilane led to formation of a 1:1 mixture of 1-cyclohexylbutane (**24**) and 2-cyclohexylbutane (**22**) (Table 2, entry 2). Moreover, use of a 1:1 mixture of C_6D_{12}:C_6H_{12} in this system led to exclusive formation of D_{12} and D_0 isomers, without any detectable amount of crossover products (that is, $C_{10}H_{8+n}D_{12-n}$ or $C_{10}H_{20-n}D_n$, $n \neq 0$) that would be expected in a stepwise process (fig. S24) (*34*, *39*). Use of terminal triflate **25** (Table 2, entry 3) resulted in an identical product distribution to that of entry 2, albeit requiring extended reaction times. Although primary vinyl cations are difficult to generate under solvolytic conditions (*11*), here the abstraction of the terminal triflate by silylium cation may provide a lower-energy transition through favorable Si–O bond formation. Moreover, a favorable pathway for a concerted triflate abstraction and hydride migration leading directly to a secondary vinyl cation has been calculated (fig. S43). Carrying out the reaction at –40°C in chloroform solvent allowed for selective formation of 1-cyclohexylbutane (**24**) (about 2:1) (Table 2, entry 4). These results are consistent with our view of a nonclassical intermediate (Table 2, $\mathbf{INT_{23-22/24}}$) (movie S5), where hydride delivery from silane determines product distributions. In the absence of coordinating

solvent (entries 1 and 2), hyperconjugative stabilization may lead to a greater degree of bridging. As we increase the amount of chloroform solvent to attenuate hyperconjugative effects, stabilization of a more classical 2° carbocation, instead of the energetically prohibitive 1° carbocation, would lead to internal Si–H quenching and the terminal product **24**. The important observation of modulated product distributions in polar solvent (entry 4) implies that the product ratios of reactions that proceed through unsymmetrical nonclassical species can be dictated by experimental factors that influence the degree of bridging.

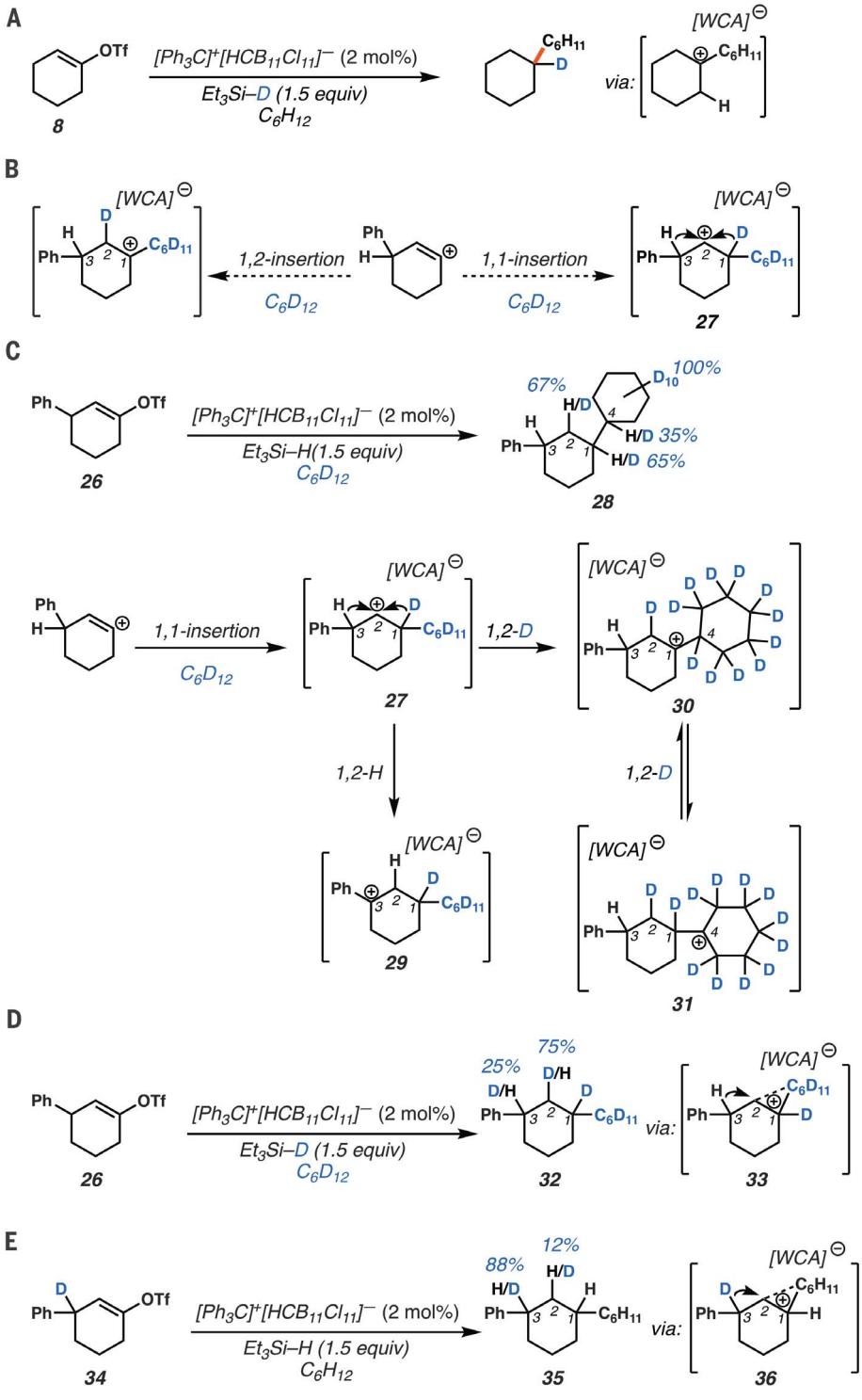


Fig. 3. Deuterium labeling studies of cyclohexenyl triflate derivatives. (**A**) Deuterium labeling experiment to support charge propagation step. (**B**) Experimental design to probe 1,1- versus 1,2-C–H insertion. (**C** to **E**) Mechanistic experiments suggesting benzylic 1,2-hydride and deuteride shifts of the resulting bridged cation. Ph, phenyl.

Isotopic labeling studies

We also carried out deuterium labeling studies to further probe the mechanism. First, competition experiments with a 1:1 mixture of C_6H_{12}:C_6D_{12} provided a 1:1 mixture of cyclohexylcyclohexane-d_0 and cyclohexylcyclohexane-d_{12}, supporting a concerted reaction pathway with C–H insertion being a diffusion-controlled event (fig. S22). The use of triethylsilane-d_1 with cyclohexane and cyclohexenyl triflate (**8**) resulted in high levels of deuterium incorporation at the tertiary carbon, supporting the formation of a persistent 3° carbocation (Figs. 1C and 3A). However, this experiment does not differentiate a 1,1- from a 1,2-insertion mechanism; the 2° carbocation (**10**) formed via 1,1 insertion will always proceed to the more stable 3° carbocation **11** (Fig. 1C). To tease apart these different C–H insertion mechanisms, we prepared 3-phenylcyclohexenyl triflate **26** in the hope that migration of the benzylic C3 hydride of carbocation intermediate **27** would be competitive with C1-deuteride migration, allowing for direct experimental evidence of a 1,1-insertion event (Fig. 3B). Upon exposure of triflate **26** to our reaction conditions, using cyclohexane-d_{12} solvent and triethylsilane, we observed 67% deuterium incorporation at C2, 65% at C1, and 35% at C4 (Fig. 3C). In total, we observed 1.6 deuterides incorporated at these three carbons (C1, C2, and C4), accounting for 80% of the two cyclohexane-d_{12} deuterides prone to migration. Here we posit that following 1,1 insertion, C3-hydride migration produces benzylic cation **29**, whereas C1-deuteride migration leads to tertiary C1 and C4 carbocations **30** and **31**. All three of these species (**29** to **31**) can be quenched by Si–H reduction. Further confirming migration of the C3-benzylic H, using C_6D_{12} and triethylsilane-d_1, we observed 25% deuterium incorporation at C3 and 75% at C2 (**32**, Fig. 3D). Here D incorporation at C3 supports the formation of a benzylic carbocation that results from 1,2-hydride migration to quench a transient nonclassical bridged cation (**33**). Moreover, use of C3-deuterated substrate **34** (Fig. 3E) with C_6H_{12} leads to D incorporation at C2 (**35**, about 12%), further supporting the migration of the benzylic deuteride to quench the carbocation (**36**) generated upon insertion. The lower levels of C2-deuteride incorporation in product **35** can be attributed to kinetic isotope effects, as the C3 deuteride (**36**) will migrate slower than the C3 hydride in cation **33**, whereas the C1 hydride of cation **36** will migrate faster in this system than the deuteride of transient **33** (*40*). The fact that less migration of the benzylic C3 hydride is observed over the tertiary C1 hydride supports the nonclassical nature of the reaction, where a discrete classical secondary cation is not formed (Fig. 3D).

Reactivity of vinyl cations with arenes

Having established that vinyl triflates are competent vinyl cation precursors under silylium

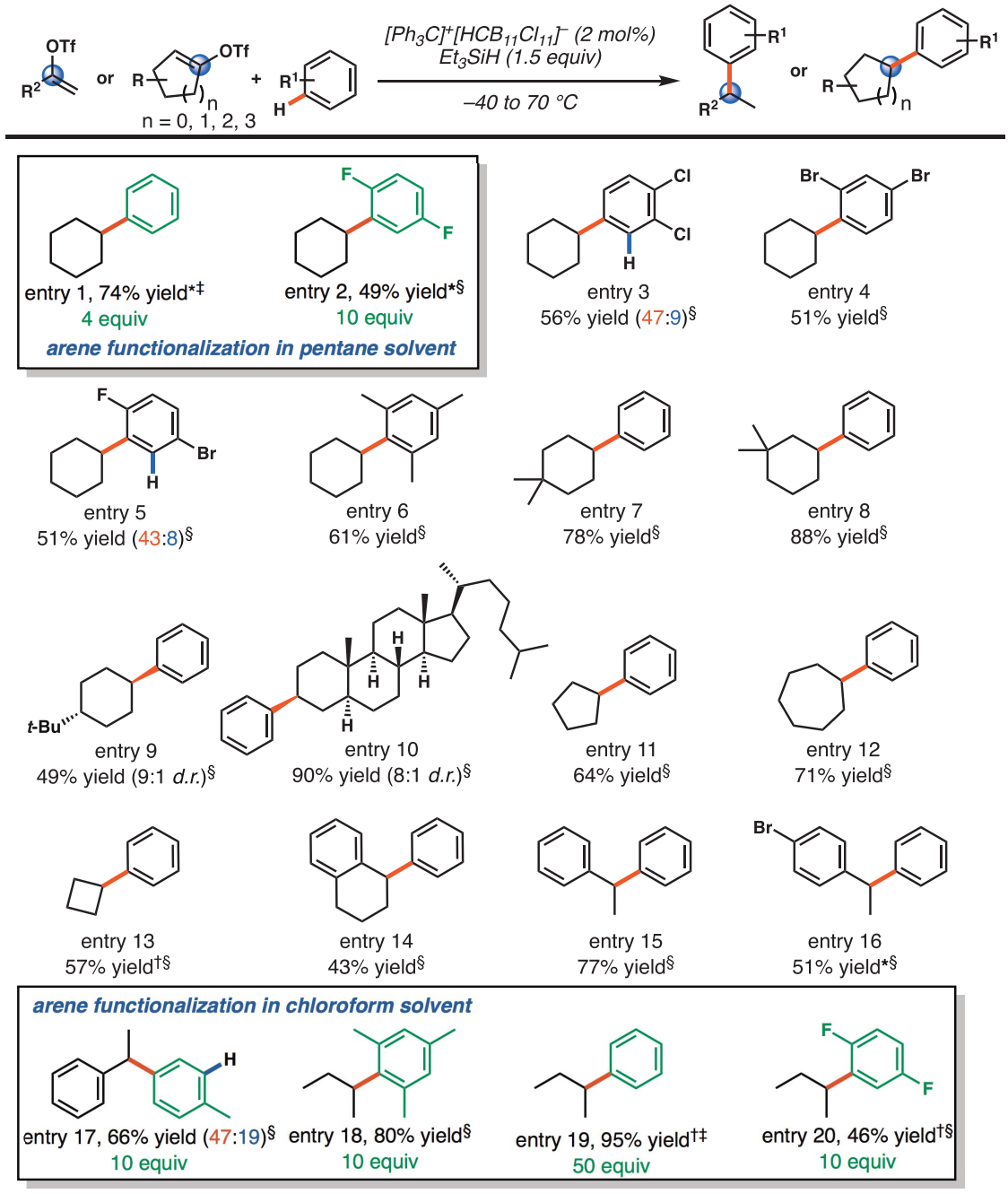


Fig. 4. Reductive arylation of vinyl triflates. *Reaction performed with 1.2 equivalents triethylsilane. †Triisopropylsilane used instead of triethylsilane. ‡Yield determined by GC-FID, with nonane as an internal standard. §Yield determined by nuclear magnetic resonance with an internal standard. R, alkyl, aryl, or H; R^1, alkyl, halide, or H; R^2, phenyl or ethyl; *t*-Bu, *tert*-butyl.

catalysis conditions and that these reactive intermediates undergo efficient sp^3 C–H functionalization reactions, we sought to investigate their reactivity with arenes. It has been reported that cyclic vinyl triflates are poor electrophiles in Friedel-Crafts arylation reactions (*41*). This finding has been attributed to the poor electrophilicity of vinyl cations (*15*) and to the slow ionization of vinyl cation precursors in arene solvents (*41*). In early studies, cyclohexenyl triflates were completely unreactive toward arene nucleophiles, whereas triflates with larger ring sizes participated in Friedel-Crafts reactions at increased temperatures, albeit in poor yields (*41*). We posited that use of silylium-carborane salts would allow for mild ionization of cyclic vinyl triflates in nonpolar solvents, allowing for facile Friedel-Crafts arylation reactions. We were pleased to find that with four equivalents of benzene in pentane solvent (1:22 molar ratio of benzene: pentane), cyclohexenyl triflate (**8**) underwent smooth reductive arylation to yield phenylcyclohexane in 79% yield in 2 hours at room temperature (Fig. 4). In this reaction, the mass balance is composed of cyclohexylpentane isomers. Although we posit that triflate abstraction is rate limiting, we hypothesize that the product-determining step is the C–C bond-forming event, with the barrier for attack of the arene π-system falling nearly 20 kcal/mol below the barrier for C–H insertion (fig. S44). In addition to benzene, electron-poor haloarenes such as difluorobenzene and dichlorobenzene underwent smooth C–C bond formation to yield cyclohexylated haloarenes in synthetically useful yields. Likewise, electron-rich arenes, such as

mesitylene, were competent nucleophiles. Cyclohexenyl triflates bearing substituents at the 4- or 5-position could also be arylated, including the enol triflate derived from 5α-cholestan-3-one, which yielded an arylated steroid core in 90% yield and 8:1 d.r. Various ring sizes were also competent under these reaction conditions, with cyclopentenyl triflate and cycloheptenyl triflate undergoing smooth reductive alkylation with benzene reaction partners in 64 and 71% yields, respectively. Cyclobutenyl triflate participated in this reductive Friedel-Crafts alkylation, as did aromatic alkenes. The triflate derived from α-tetralone was reductively arylated in 43% yield, and acetophenone-derived acyclic triflates were also arylated in 51 to 77% yield (Fig. 4, entries 15 and 16). Simple acyclic vinyl triflates were competent electrophiles for arylation by both electron-poor and electron-rich arenes, requiring as little as 10 equivalents of arene in chloroform solvent at −40°C (Fig. 4, entries 17 to 20).

Outlook

We have shown that vinyl cations, the subject of numerous computational and experimental studies, are now accessible synthetically from simple vinyl triflates using WCA salts under mild conditions. The nonnucleophilic nature of the WCA allows these unstabilized vinyl cations to engage in C–C bond–forming reactions with alkanes and a variety of arenes, modes of reactivity that have been largely unreported despite extensive previous work. We find that the C–H insertion reactions of vinyl cations proceed through mechanisms that feature post–transition state bifurcation and nonclassical ions, mechanistic features that are common in terpene biosynthesis (*32*) but rarely found in synthetic methodology. These findings lay the conceptual and experimental groundwork for further discoveries in the field of alkane C–H bond functionalization using ketone derivatives and WCA catalysis.

REFERENCES AND NOTES

1. G. A. Olah, *J. Org. Chem.* **66**, 5943–5957 (2001).
2. T. L. Jacobs, S. Searles Jr., *J. Am. Chem. Soc.* **66**, 686–689 (1944).
3. C. A. Grob, J. Csapilla, G. Cseh, *Helv. Chim. Acta* **47**, 1590–1602 (1964).
4. M. Hanack, *Angew. Chem. Int. Ed. Engl.* **17**, 333–341 (1978).
5. Z. Rappoport, A. Gal, *J. Am. Chem. Soc.* **91**, 5246–5254 (1969).
6. L. Radom, P. C. Hariharan, J. A. Pople, P. R. Schleyer, *J. Am. Chem. Soc.* **95**, 6531–6544 (1973).
7. P. J. Stang, Z. Rappoport, *Dicoordinated Carbocations* (Wiley, 1997).
8. S. A. Sherrod, R. G. Bergman, *J. Am. Chem. Soc.* **91**, 2115–2117 (1969).
9. T. Müller, M. Juhasz, C. A. Reed, *Angew. Chem. Int. Ed.* **43**, 1543–1546 (2004).
10. W. M. Jones, F. W. Miller, *J. Am. Chem. Soc.* **89**, 1960–1962 (1967).
11. R. J. Hinkle, A. J. McNeil, Q. A. Thomas, M. N. Andrews, *J. Am. Chem. Soc.* **121**, 7437–7438 (1999).
12. Y. Apeloig, W. Franke, Z. Rappoport, H. Schwarz, D. Stahl, *J. Am. Chem. Soc.* **103**, 2770–2780 (1981).
13. J. A. Pople, W. A. Lathan, W. J. Hehre, *J. Am. Chem. Soc.* **93**, 808–815 (1971).
14. R. Pellicciari *et al.*, *J. Am. Chem. Soc.* **118**, 1–12 (1996).
15. P. A. Byrne, S. Kobayashi, E. U. Würthwein, J. Ammer, H. Mayr, *J. Am. Chem. Soc.* **139**, 1499–1511 (2017).
16. A. J. Walkinshaw, W. Xu, M. G. Suero, M. J. Gaunt, *J. Am. Chem. Soc.* **135**, 12532–12535 (2013).
17. W. S. Johnson *et al.*, *J. Am. Chem. Soc.* **103**, 88–98 (1981).
18. M. Hanack, *Acc. Chem. Res.* **3**, 209–216 (1970).
19. P. J. Stang, Z. Rappaport, M. Hanack, L. R. Subramanian, *Vinyl Cations* (Academic Press, 1979).
20. B. Shao, A. L. Bagdasarian, S. Popov, H. M. Nelson, *Science* **355**, 1403–1407 (2017).
21. E. Lamparter, M. Hanack, *Eur. J. Inorg. Chem.* **105**, 3789–3793 (1972).
22. R. J. Hargrove, P. J. Stang, *Tetrahedron* **32**, 37–41 (1976).
23. G. A. Olah, H. Mayr, *J. Am. Chem. Soc.* **98**, 7333–7340 (1976).
24. M. I. Kanishchev, A. A. Shegolev, W. A. Smit, R. Caple, M. J. Kelner, *J. Am. Chem. Soc.* **101**, 5660–5671 (1979).
25. S. Fornarini, M. Speranza, *J. Phys. Chem.* **91**, 2154–2160 (1987).
26. U. Biermann, R. Koch, J. O. Metzger, *Angew. Chem. Int. Ed.* **45**, 3076–3079 (2006).
27. S. E. Cleary, M. J. Hensinger, M. Brewer, *Chem. Sci.* **8**, 6810–6814 (2017).
28. S. Körbe, P. J. Schreiber, J. Michl, *Chem. Rev.* **106**, 5208–5249 (2006).
29. F. A. Carey, H. S. Tremper, *J. Org. Chem.* **36**, 758–761 (1971).
30. D. H. Ess *et al.*, *Angew. Chem. Int. Ed.* **47**, 7592–7601 (2008).
31. P. Yu, A. Patel, K. N. Houk, *J. Am. Chem. Soc.* **137**, 13518–13523 (2015).
32. Y. J. Hong, D. J. Tantillo, *Nat. Chem.* **1**, 384–389 (2009).
33. P. R. Schleyer *et al.*, *J. Am. Chem. Soc.* **93**, 1513–1516 (1971).
34. See supplementary materials.
35. T. Okuyama, T. Takino, T. Sueda, M. Ochiai, *J. Am. Chem. Soc.* **117**, 3360–3367 (1995).
36. L. A. Curtiss, J. A. Pople, *J. Chem. Phys.* **88**, 7405–7409 (1988).
37. J. Weber, M. Yoshimine, A. D. McLean, *J. Chem. Phys.* **64**, 4159–4164 (1976).
38. B. T. Psciuk, V. A. Benderskii, H. B. Schlegel, *Theor. Chem. Acc.* **118**, 75–80 (2007).
39. G. A. Olah *et al.*, *J. Am. Chem. Soc.* **116**, 3187–3191 (1994).
40. J. M. Coxon, A. J. Thorpe, *J. Org. Chem.* **65**, 8421–8429 (2000).
41. P. J. Stang, A. Anderson, *J. Am. Chem. Soc.* **100**, 1520–1525 (1978).

ACKNOWLEDGMENTS

H.M.N. thanks M. Jung for reagents and advice. **Funding:** Support was generously provided by the David and Lucile Packard Foundation (to H.M.N.), the Alfred P. Sloan Foundation (to H.M.N.), the National Science Foundation (CHE-1361104 to K.N.H.), and the National Natural Science Foundation of China (grant nos. 11504130, 51673164, and 21501169 to L.Z.). A.L.B. thanks the Christopher S. Foote Fellowship for funding. The authors thank the UCLA Molecular Instrumentation Center for NMR instrumentation, x-ray crystallography, and the Mass Spectrometry facility at the University of California, Irvine. **Author contributions:** S.P., B.S., and A.L.B. designed and conducted experiments. T.R.B., L.Z., and Z.Y. designed and conducted computations. H.M.N., K.N.H., S.P., B.S., A.L.B., and T.R.B. prepared the manuscript. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** Crystallographic data are available free of charge from the Cambridge Crystallographic Data Centre under CCDC 1838441. Additional experimental procedures and characterization data are provided in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/381/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S44
Tables S1 and S2
References (*42*–*66*)
Movies S1 to S5

9 March 2018; accepted 5 June 2018
10.1126/science.aat5440

REPORT

DEVICE TECHNOLOGY

Dirac-source field-effect transistors as energy-efficient, high-performance electronic switches

Chenguang Qiu[1], Fei Liu[2], Lin Xu[1], Bing Deng[3], Mengmeng Xiao[1], Jia Si[1], Li Lin[3], Zhiyong Zhang[1]*, Jian Wang[2], Hong Guo[4], Hailin Peng[3], Lian-Mao Peng[1]*

An efficient way to reduce the power consumption of electronic devices is to lower the supply voltage, but this voltage is restricted by the thermionic limit of subthreshold swing (SS), 60 millivolts per decade, in field-effect transistors (FETs). We show that a graphene Dirac source (DS) with a much narrower electron density distribution around the Fermi level than that of conventional FETs can lower SS. A DS-FET with a carbon nanotube channel provided an average SS of 40 millivolts per decade over four decades of current at room temperature and high device current I_{60} of up to 40 microamperes per micrometer at 60 millivolts per decade. When compared with state-of-the-art silicon 14-nanometer node FETs, a similar on-state current I_{on} is realized but at a much lower supply voltage of 0.5 volts (versus 0.7 volts for silicon) and a much steeper SS below 35 millivolts per decade in the off-state.

The need for extended time of operation for devices powered by batteries has changed the electronic industry (*1*–*4*). Minimal power consumption may be accomplished by reducing the supply voltage V_{DD} and by designing transistors with steeper subthreshold swing (SS) so that they could be switched from off-state to on-state at faster rates (*5*–*7*). However, conventional Si complementary metal oxide semiconductor (CMOS) field-effect transistors (FETs) are constrained to a 60 mV/decade minimum constraint on SS at room temperature (*2*) because in a typical n-type FET, electrons in the source are prevented from flowing to the drain by a potential barrier (Fig. 1A) that can be tuned by the gate voltage. The electrons in the source have an energy distribution (usually a thermal Boltzmann distribution) that spreads to the values exceeding the potential barrier created by the gate, which sets a 60 mV/decade limit on SS.

Scaling V_{DD} < 0.5 V while maintaining a high on-state current I_{on} and low off-state current I_{off} will require transistors with a different operation mechanism. These devices include tunnel transistors (T-FETs) (*6*, *7*), impact ionization FETs (i-FETs) (*8*), and positive feedback and negative gate capacitance FETs (NC-FETs) (*9*–*11*). T-FETs have emerged as the most promising alternative, with the potential to overcome the thermally limited SS of 60 mV/decade by using a supply voltage below 0.5 V, thereby offering substantial power dissipation savings. However, all fabricated T-FETs have severely limited on-state current (typically lower than 10 μA/μm), and the steepest switching slope SS of sub–60 mV/decade is usually not sustained across the entire subthreshold region (*12*–*16*). Other kinds of sub–60 mV/decade transistors also suffer from vital drawbacks such as large hysteresis (*10*) and limited speed, stability, or voltage scalability (*8*, *17*).

We now consider introducing a "Dirac source" (DS) for the conventional FET, which is a "cold electron source" without a long thermal tail above the potential barrier in the channel. If the density of states (DOS) of electrons in the source is a decreasing function of energy, the density of electrons will decrease superexponentially with increasing energy (Fig. 1B and blue curve in Fig. 1C), resulting in a much more localized electron distribution around the Fermi level E_F. This DS still permits a large thermionic current in the on-state, as in a conventional FET (see Fig. 1D).

We chose graphene as the DS material. Graphene has a linear band dispersion, and when suitably doped it may yield a superexponentially decreasing electron density with increasing energy toward the Dirac point (Fig. 1B). In addition to providing a sharply localized carrier distribution around E_F, the DS should also have a clean interface with the channel and suitably aligned E_F and band edges. For a graphene source, semiconducting single-walled carbon nanotubes (SWCNTs) can form a highly transparent interface for the injection of carriers when the DS-FET is in its on-state, but they create a large barrier to block the leakage current when in its off-state. Given their extremely high mobility for both electrons and holes and their ultrathin body, CNTs have been used for building conventional FETs that show better performance than state-of-the-art Si MOSFETs when scaled down to the sub–10 nm technology node (*18*–*24*). When combined with graphene, a CNT FET could be scaled down to 5 nm without suffering noticeable short-channel effect, and a near–thermionic limit SS value of 60 mV/decade was demonstrated (*24*).

Our DS-FET device (Fig. 2A) consisted of three components: (i) a semiconducting SWCNT channel, (ii) a suitably doped "cold" graphene DS (GDS), and (iii) a highly efficient Y_2O_3/Pd top gate stack to provide electrostatics control of the channel (see fig. S1). The device must set E_F in the n-branch of the GDS for p-type FETs or in the p-branch for n-type FETs, and the Schottky barrier between the graphene and CNT must be small or even zero when the device is switched to on-state (blue region in fig. S1B), because a large Schottky barrier will dominate the carrier injection process (*25*).

We used a control gate (CG) to tune the Fermi level (E_{FS}) of the GDS (Fig. 2, A and B). In Fig. 2A, the DS FET was designed to be a p-type FET because the CNT channel contacts a Pd (p-type) drain. In Fig. 2C, we show a transport model based on a calculated band-edge profile using full quantum transport simulations (detailed in fig. S2). Beneath the control gate, the graphene is n-doped. When biased with a drain voltage, electrons are injected from the n-doped graphene source to the drain, passing through an n-p junction in graphene via tunneling before reaching a CNT bulk barrier, the height and width of which can be tuned by a gate on top of the CNT. In graphene, the n-p tunneling is Klein tunneling and is essentially transparent (*26*, *27*). Thus, the device current was mainly modulated by the top gate via varying the height and width of the CNT bulk barrier.

In the off-state, the CNT bulk barrier width was >100 nm and the tunneling current through this barrier was negligible; the dominating contribution to the off-state current originated from the thermal emission current over the CNT bulk barrier (Fig. 2C). During the switching-off process, the barrier height for hole injection increased and simultaneously the DOS of the n-doped graphene source over the top of the barrier decreased toward the Dirac point (see Fig. 2C and fig. S3). Thus, the thermally activated hole density over the barrier in the graphene source decreased with decreasing energy superexponentially, leading to a slope steeper than 60 mV/decade in the DS-FET (Fig. 2C). To verify the dominating operation mechanism for sub–60 mV/decade SS in DS-FETs, we simulated the device operation theoretically using both analytical and numerical methods (see supplementary materials and fig. S1). A ballistic FET model with an n-doped graphene and Pd-contacted CNT was used to fit

[1]Key Laboratory for the Physics and Chemistry of Nanodevices and Department of Electronics, Peking University, Beijing 100871, China. [2]Department of Physics, University of Hong Kong, Hong Kong, China. [3]Center for Nanochemistry, Beijing Science and Engineering Center for Nanocarbons, Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China. [4]Department of Physics, McGill University, Montreal, Quebec H3A 2T8, Canada.
*Corresponding author. Email: zyzhang@pku.edu.cn (Z.Z.); lmpeng@pku.edu.cn (L.-M.P.)

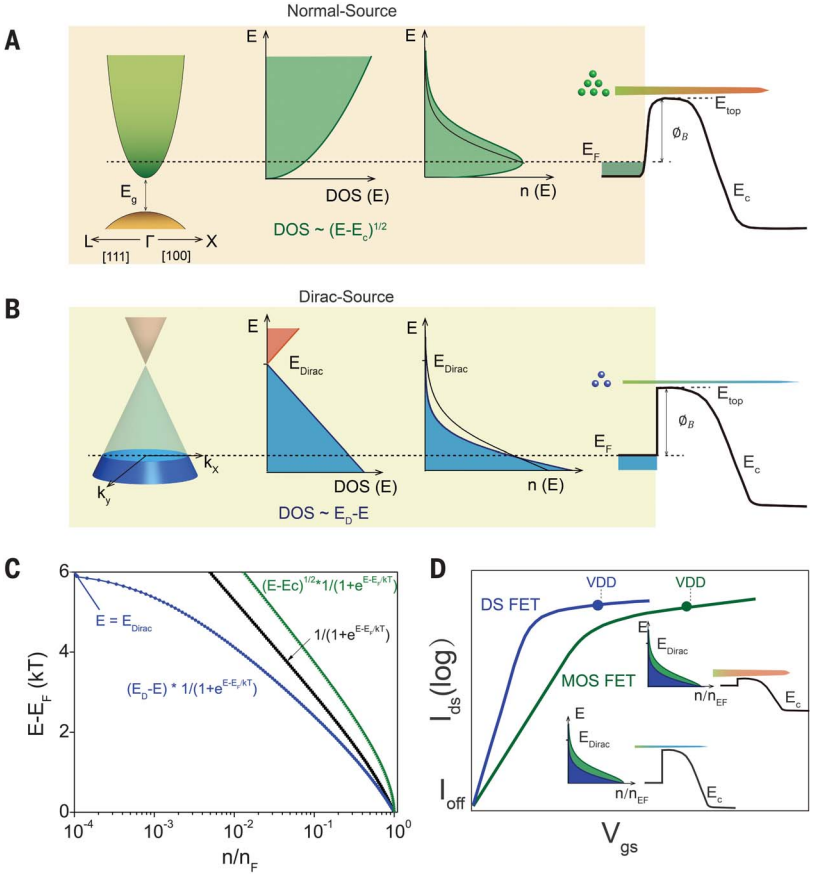


Fig. 1. Electron density distribution and characteristics of normal and Dirac sources. (**A** and **B**) Schematic band structure, density of states, electron density, and thermionic electron emission over a potential barrier in the conduction channel for a normal source with a subexponential decaying of electron density over the potential barrier (A) and a Dirac source with a superexponential decaying and thus a more narrowly localized electron density around the Fermi level (B). The solid line in $n(E)$ represents a Boltzmann distribution with exponential decaying tail toward higher energy; $n(E)$ represents density of electrons. (**C**) Comparison of electron density distributions between normal sources, based on 3D semiconductor (green curve) and on 2D material (black curve), and a Dirac source based on graphene; the blue curve shows increasingly more localized electron distribution around the Fermi surface for graphene. (**D**) Schematic transfer characteristics of a conventional FET with a normal source and thus 60 mV/decade–limited subthreshold swing (SS), and a DS-FET with a Dirac source and much steeper SS. Insets are electron density distributions for a normal source (green) and Dirac source (blue), and thermionic electron emission over the potential barrier at low bias (off-state, lower inset) and high bias (on-state, upper inset).

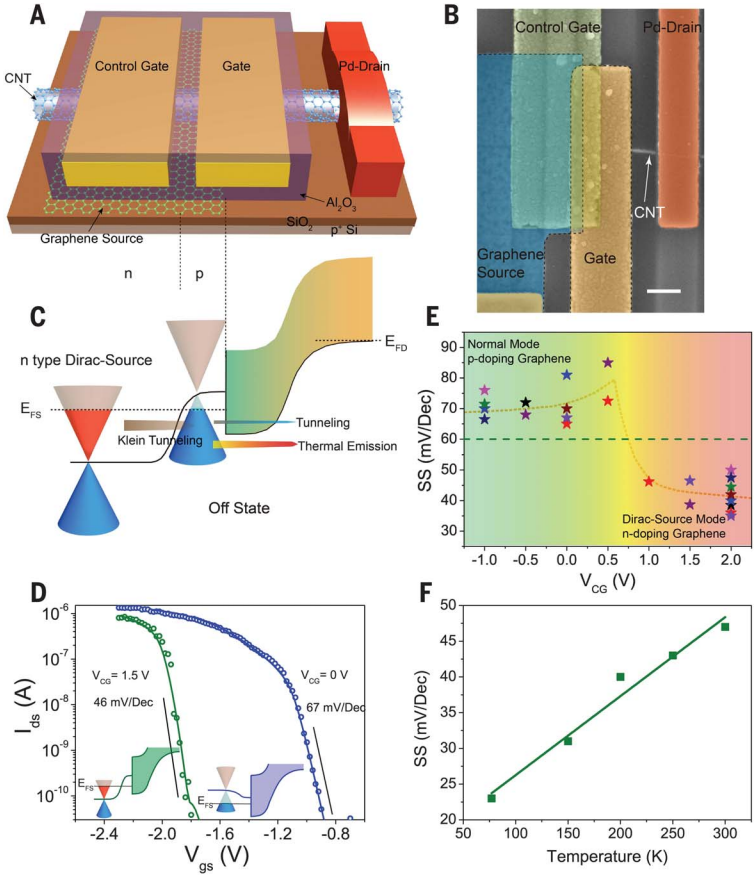


Fig. 2. Structure and performance of a DS-FET with a control gate. (**A**) Schematic diagram showing a DS-FET with a control gate in addition to the normal gate. (**B**) Top-view SEM image of a DS-FET. Scale bar, 200 nm. The devices are fabricated on doped Si substrate covered with a 300-nm layer of SiO_2. (**C**) Schematic diagrams illustrating the off-state of the DS-FET. (**D**) Transfer characteristics of a typical DS-FET at different values of V_{CG}. Circles, experimental results; curves, simulated results. Green and blue denote results obtained at V_{CG} = 1.5 V and 0 V, respectively. Insets are schematic band edge profiles for fitted data situations. (**E**) Subthreshold swing phase diagram as a function of the control gate voltage, showing the condition for the appearance of sub–60 mV/decade behavior. The orange dashed line is a guiding curve of SS-V_{CG} for DS-FETs according to simulation results. Each colored star represents one DS-FET at the corresponding V_{CG}. (**F**) Temperature-dependent SS of a typical DS-FET measured at temperatures between 77 and 300 K. V_{CG} was set at 2 V to keep the device in Dirac-source mode. SS varied by more than 100% from 77 to 300 K. In all measurements, the substrate was biased with −20 V to keep the ungated region near the drain open.

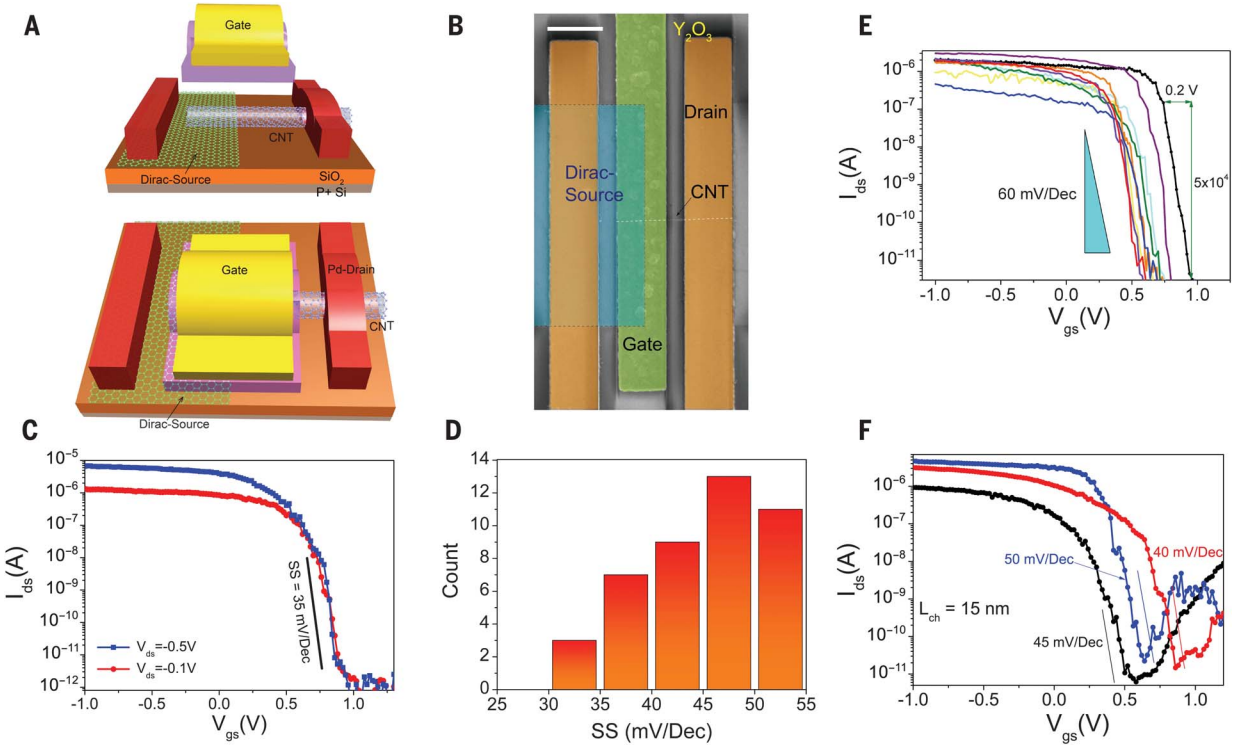


Fig. 3. Structure and performance of DS-FETs with chemically doped Dirac sources. (**A**) Schematic diagrams illustrating the structure of a DS-FET, with a chemically n-doped graphene source and Pd-contacted drain. (**B**) Top-view SEM image showing a DS-FET. Scale bar, 500 nm. (**C**) Transfer characteristics of a typical DS-FET with V_{ds} = −0.1 V and −0.5 V. The substrate was biased with −20 V. (**D**) Statistical distribution of SS for 43 DS-FETs with chemically doped graphene as the Dirac source. (**E**) Transfer characteristics for some representative DS-FETs with V_{ds} = −0.1 V and average SS ranging from 40 to 50 mV/decade. (**F**) Transfer characteristics of three extremely scaled DS-FETs with channel length of about 15 nm and at bias of −0.1 V.

the measured transfer characteristics (Fig. 2D), showing excellent agreement between experimental and theoretical results.

The two transfer characteristics shown in Fig. 2D correspond to two operation modes of the CNT FET with graphene as the source (for more transfer characteristics, see fig. S4). When the control gate was set at zero (V_{CG} = 0 V), the graphene source was usually p-doped; the carrier DOS increased with increasing V_{CG} toward the off-state (Fig. 2D, right inset). The device worked as a conventional FET with a SS of 67 mV/decade. However, applying a large positive V_{CG} to the control gate tuned the GDS to its n-branch (Fig. 2D, left inset). The carrier DOS in the GDS decreased with increasing V_{CG} toward off-state. The transistor now had a SS of 46 mV/decade at room temperature (Fig. 2D, green curve).

Extensive measurements were then carried out on the dependence of SS on V_{CG} for multiple DS-FETs. Figure 2E defines the region in which the CNT FET works as a DS-FET, yielding <60 mV/decade, and the main trend of experimental results is captured very well by our model simulations (fig. S1B). When in DS-FET mode, the transfer characteristics of a typical FET were measured at various temperatures T between 77 and 300 K (fig. S5A), and the extracted SS (Fig. 2F) showed a linear dependence on T, as predicted by theoretical simulations (see fig. S5B). The linear SS-T relation further confirmed that the main mechanism for electron injection over the potential barrier at subthreshold region is thermal emission rather than band-to-band tunneling (BTBT), because for BTBT SS shows much weaker T dependence (*2*, *28*). Tunneling assisted by phonons and mid-gap trap states may occur in T-FETs, resulting in T-dependent SS (*29*). In these T-FETs, SS varies weakly with T (within a few percent from 100 to 300 K) and/or becomes much greater than 60 mV/decade at room temperature. Trap states may act as stepping stones in these devices that lead to band-to-band tunneling in off-state and degrade SS (*29*, *30*).

In real applications, FETs with additional control gates are undesirable because this structure introduces additional supply burden and is also not scalable. An alternative compact DS-FET structure (Fig. 3A) has a chemically doped GDS. The Pd film was patterned as the drain contact to define the polarity of the FET as p-type (*18*). The graphene was intentionally doped during the chemical vapor deposition growth process (*31*), and the Dirac point voltage of the as-grown graphene was distributed around −50 V (see fig. S6A) for the group of samples used in this work, thereby confirming that the graphene was effectively n-doped. Part of the graphene source and CNT channel are beneath the top gate, and an ultrathin Y_2O_3 is used as the gate dielectric with an equivalent oxide thickness of 1.5 nm (*32*). More than 50 such DS-FETs were fabricated; the SEM image of a typical DS-FET is shown in Fig. 3B. Transfer characteristics of the DS-FET (Fig. 3C) demonstrate the sub–60 mV/decade subthreshold behavior of this device. At the subthreshold voltage, the transistor presents a SS of 35 mV/decade at room temperature at both low bias (V_{ds} = −0.1 V) and high bias (V_{ds} = −0.5 V). Moreover, the sub–60 mV SS behavior was measured for both sweeping directions of V_{gs} and showed negligible hysteresis (see fig. S7). The devices achieved an average SS < 60 mV/decade over four decades of current (see Fig. 3E and fig. S7) and can thus meet the requirement set by the International Technology Roadmap for Semiconductors (ITRS) (*4*).

In addition, the DS-FET based on a single CNT can provide high on-state current of up to 6.5 μA at V_{DS} = −0.5 V, which is orders of magnitude higher than that observed in a tunnel FET based on a single CNT (*28*, *33*). The DS-FET also exhibits much suppressed ambipolar characteristics relative to conventional top-gated CNT FETs (*34*), presenting a large current on/off ratio of >10^6 at V_{ds} = −0.5 V. More than 40 DS-FETs have been successfully fabricated with sub–60 mV/decade SS, and the statistical distribution of SS shown in Fig. 3D demonstrates that most devices (32 devices) exhibited SS < 50 mV/decade at room temperature. The transfer characteristics

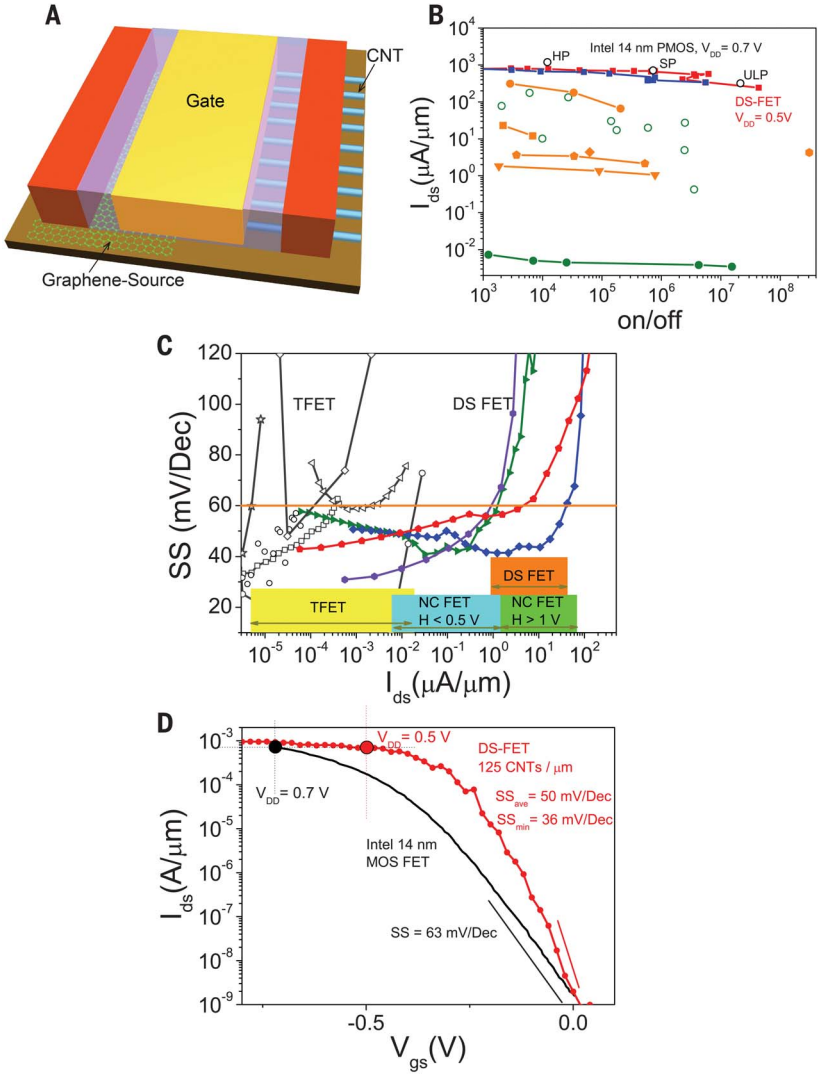


Fig. 4. Comparison of DS-FETs, state-of-the-art T-FETs, NC-FETs, and Intel 14-nm MOSFETs. (**A**) Schematic diagram illustrating the structure of a DS-FET with CNT array as the channel. (**B**) On-state current I_{ds} versus on/off current ratio for DS-FETs, state-of-the-art T-FETs, NC-FETs, and Intel 14-nm MOSFETs. Red and blue squares denote two typical DS-FETs; open black circles represent Intel 14-nm HP (high performance), SP (standard performance), and ULP (ultralow-power) MOSFETs (*35*); open green circles, III-V NW T-FET (*13*); solid green circles, MoS_2-Ge T-FET (*36*); solid orange circles, squares, pentagons, hexagon, diamond, and triangles, typical NC-FETs with V_{DD} < 1 V (*38*–*43*). (**C**) Subthreshold swing versus I_{ds}. I_{60} values in the yellow region are for T-FETs (*16*); those in the cyan region are for NC-FETs with hysteresis (H) smaller than 0.5 V; and those in the green region are for NC-FETs with H larger than 1 V (*38*–*45*). The red, green, blue, and purple curves with solid symbols at the upper right represent four different DS-FETs. The black curves with different open symbols at the left represent different T-FETs (*16*). (**D**) Comparison between a 400-nm DS-FET (125 CNTs/μm) and a commercial Intel 14-nm Si MOSFET. The Si MOSFET (black) was powered by 0.7 V (*35*); the DS-FET (red) was powered by a low bias of 0.5 V.

of nine representative devices are plotted in Fig. 3E.

Because miniaturization is an important issue for large-scale integrated circuits (ICs), we fabricated DS-FETs with ultrashort channel lengths of <15 nm (see figs. S8 and S9 for the relevant structure and fabrication process of the extremely scaled DS-FETs); Fig. 3F shows measurement results for three such DS-FETs with SS < 60 mV/decade. The short-channel DS-FETs showed very high on-state current of >10 μA per CNT (fig. S10), which is similar to that achieved in the best reported high-performance p-type conventional CNT FETs (*18*).

The DS-FET transistor can achieve higher performance and lower dissipation than the most advanced Si CMOS FETs at lower supply voltage. Pure semiconducting CNT arrays with high density (e.g., 125 tubes/μm) have been considered as a good choice for channel material to construct scalable FETs with performance exceeding that of the best Si CMOS FETs (*20*). We projected the performance of DS-FETs based on a CNT array (Fig. 4A) by normalizing the experimental data on single CNTs and assuming a CNT array density of 125 tubes/μm. The key challenge for realizing high-density CNT array–based DS-FETs results mainly from the nonuniformity of CNTs in the array and their contacts to graphene. Identical CNTs with the same chirality and highly uniform contacts to graphene are needed to retain the threshold voltage of every CNT in the channel within a required narrow range. Ideally, a high-performance FET should have high on-state current I_{on} and large current on/off ratio at small supply voltage. Figure 4B compares the on-state current I_{ds} and on/off ratio of our DS-FETs and that of state-of-the-art Si MOSFETs, the best T-FETs, and NC-FETs (with V_{DD} < 1 V) (*35*–*46*). For the same on/off ratio, two representative DS-FETs (blue and red lines) at V_{DD} = 0.5 V show I_{ds} values similar to that of the best Si p-type MOSFET (Intel 14-nm technology node) (*35*) driven by V_{DD} = 0.7 V, which is 10 to 100 times that of the most advanced III-V nanowire-based heterojunction T-FETs (*13*, *36*) and much greater than that of the best reported NC-FETs (*38*–*43*).

An important figure of merit—I_{60}, the current when the subthreshold swing SS equals 60 mV/decade—accounts for both switching and device on-state characteristics (*37*). Higher values of I_{60} correspond to higher on-state current and lower average SS. To benchmark the performance of sub–60 mV/decade devices by using I_{60}, we plot SS-I_{ds} relations of the best published T-FETs, NC-FETs, and our CNT-based DS-FETs in Fig. 4C. For a sub–60 mV/decade device to have a performance comparable to that of the conventional FET, the desired target value of I_{60} for sub–60 mV/decade devices is in the range 1 to 10 μA/μm at V_{DD} below 0.5 V. However, all the reported T-FETs have I_{60} values that are typically two orders of magnitude lower than the target (*16*); in particular, the most advanced InAs/Si nanowire (*46*) T-FET has an I_{60} of 0.02 μA/μm at V_{ds} = 1 V. As a comparison, CNT-based DS-FETs typically have I_{60} values in the range 1 to 40 μA/μm at V_{dd} = −0.5 V, which meets the I_{60} challenge for sub–60 mV/decade devices.

Additional challenges for T-FETs include optimization of the device architecture for high I_{on} combined with an average SS lower than 60 mV/decade over at least four decades. The representative transfer characteristics shown in Fig. 3E amply demonstrate that the sub–60 mV/decade region of CNT-based DS-FETs covers more than four decades of current, meeting the target of ITRS for sub–60 mV/decade transistors. Although some NC-FETs show high I_{60} values in the range 1 to 40 μA/μm (Fig. 4C), large hysteresis (larger than 1 V) remains the main challenge for their applications in ICs. For these reported NC-FETs with small hysteresis (<0.5 V), I_{60} is usually lower than 1 μA/μm (*38*–*45*). As a result, high performance and small hysteresis have not been achieved by the same NC-FET.

Figure 4D shows a direct comparison between the CNT array–based DS-FET (with a gate length of 400 nm) and state-of-the-art Si MOSFET (Intel 14-nm technology) (*35*). The 400-nm CNT DS-FET exhibited an I_{on} value similar to that of the 14-nm Si MOSFET, but with much steeper SS (with an average value of about 50 mV/decade over four decades of current) and much lower supply voltage (0.5 V versus 0.7 V). The power dissipation of modern ICs depends strongly on supply voltage (which is proportional to V_{DD}^3). The lowering in V_{DD} from 0.7 V to <0.5 V with DS-FETs will reduce power dissipation better than state-of-the-art Si-based computer chips while maintaining similar high performance.

Although we only demonstrated p-type DS FETs, according to the symmetric energy band (*19*) we could realize an n-type DS-FET based on CNT by using p-doped graphene as the source and Sc as the drain. In principle, it is possible to realize complementary DS-FETs following procedures similar to those outlined here. CNT provides an excellent conduction channel for DS-FET, but the key to the realization of sub–60 mV/decade SS in DS-FETs is not the CNT channel. In principle, GDS may also be combined with other semiconductor channel materials [e.g., semiconducting nanowires, two-dimensional (2D) semiconductors, or even conventional bulk semiconductors] to simultaneously achieve sub–60 mV/decade SS and high I_{on}. The DS-FET may thus be used as a general device or building block for future ICs with sub–0.5 V power supply.

REFERENCES AND NOTES

1. L. Chang *et al*., *Proc. IEEE* **98**, 215–236 (2010).
2. A. M. Ionescu, H. Riel, *Nature* **479**, 329–337 (2011).
3. C. M. Hu *et al*., paper presented at the 2010 International Electron Devices Meeting; doi:10.1109/IEDM.2010.5703372.
4. International Technology Roadmap for Semiconductors (2013 edition); www.itrs2.net/2013-itrs.html.
5. M. M. Waldrop, *Nature* **530**, 144–147 (2016).
6. T. Skotnicki, J. A. Hutchby, T. King, H.-S. P. Wong, F. Boeuf, *IEEE Circuits Devices* **21**, 16–26 (2005).
7. A. C. Seabaugh, Q. Zhang, *Proc. IEEE* **98**, 2095–2110 (2010).
8. K. Gopalakrishnan, P. B. Griffin, J. D. Plummer, *IEEE Trans. Electron Dev.* **52**, 69–76 (2005).
9. S. Salahuddin, S. Datta, *Nano Lett.* **8**, 405–410 (2008).
10. A. I. Khan *et al*., paper presented at the 2011 International Electron Devices Meeting; doi:10.1109/IEDM.2011.6131532.
11. E. Gnani, S. Reggiani, A. Gnudi, G. Baccarani, *Solid-State Electron.* **65–66**, 108–113 (2011).
12. K. Jeon *et al*., paper presented at the 2010 Symposium on VLSI Technology; doi: 10.1109/VLSIT.2010.5556195.
13. H. Riel, L.-E. Wernersson, M. Hong, J. A. del Alamo, *MRS Bull.* **39**, 668–677 (2014).
14. S. H. Kim *et al*., *IEEE Electron Device Lett.* **31**, 1107–1109 (2010).
15. F. Mayer *et al*., paper presented at the 2008 IEEE International Electron Devices Meeting; doi:10.1109/IEDM.2008.4796641.
16. H. Lu, A. Seabaugh, *IEEE J. Electron Devices Soc.* **2**, 44–49 (2014).
17. A. Padilla *et al*., paper presented at the 2008 IEEE International Electron Devices Meeting; doi:10.1109/IEDM.2008.4796643.
18. A. Javey, J. Guo, Q. Wang, M. Lundstrom, H. Dai, *Nature* **424**, 654–657 (2003).
19. Z. Y. Zhang *et al*., *Nano Lett.* **7**, 3603–3607 (2007).
20. G. S. Tulevski *et al*., *ACS Nano* **8**, 8730–8745 (2014).
21. Q. Cao, J. Tersoff, D. B. Farmer, Y. Zhu, S.-J. Han, *Science* **356**, 1369–1372 (2017).
22. G. J. Brady *et al*., *Sci. Adv.* **2**, e1601240 (2016).
23. A. D. Franklin *et al*., *Nano Lett.* **12**, 758–762 (2012).
24. C. Qiu *et al*., *Science* **355**, 271–276 (2017).
25. L. Britnell *et al*., *Science* **335**, 947–950 (2012).
26. S. Das Sarma, S. Adam, E. H. Hwang, E. Rossi, *Rev. Mod. Phys.* **83**, 407–470 (2011).
27. N. Stander, B. Huard, D. Goldhaber-Gordon, *Phys. Rev. Lett.* **102**, 026807 (2009).
28. J. Appenzeller, Y.-M. Lin, J. Knoch, P. Avouris, *Phys. Rev. Lett.* **93**, 196805 (2004).
29. S. Mookerjea, D. Mohata, T. Mayer, V. Narayanan, S. Datta, *IEEE Electron Device Lett.* **31**, 564–566 (2010).
30. D. Esseni, M. G. Pala, *IEEE Trans. Electron Dev.* **60**, 2802–2807 (2013).
31. L. Lin *et al*., *Nano Lett.* **16**, 4094–4101 (2016).
32. H. Xu *et al*., *ACS Nano* **5**, 2340–2347 (2011).
33. G. Zhang *et al*., paper presented at the 2006 International Electron Devices Meeting; doi:10.1109/IEDM.2006.346804.
34. C. Qiu *et al*., *ACS Nano* **9**, 969–977 (2015).
35. S. Natarajan *et al*., paper presented at the 2014 IEEE International Electron Devices Meeting; doi:10.1109/IEDM.2014.7046976.
36. D. Sarkar *et al*., *Nature* **526**, 91–95 (2015).
37. W. G. Vandenberghe *et al*., *Appl. Phys. Lett.* **102**, 013510 (2013).
38. A. I. Khan *et al*., *IEEE Electron Device Lett.* **37**, 111–114 (2016).
39. F. A. McGuire *et al*., *Nano Lett.* **17**, 4801–4806 (2017).
40. E. Ko, H. Lee, Y. Goh, S. Jeon, C. Shin, *IEEE J. Electron Devices Soc.* **5**, 306–309 (2017).
41. J. Zhou *et al*., *IEEE Electron Device Lett.* **38**, 1157–1160 (2017).
42. M. H. Lee *et al*., paper presented at the 2015 IEEE International Electron Devices Meeting; doi:10.1109/IEDM.2015.7409759.
43. M. Si *et al*., *Nat. Nanotechnol.* **13**, 24–28 (2018).
44. K.-S. Li *et al*., paper presented at the 2015 IEEE International Electron Devices Meeting; doi:10.1109/IEDM.2015.7409760.
45. J. Jo, C. Shin, *IEEE Electron Device Lett.* **37**, 245–248 (2016).
46. K. Tomioka *et al*., paper presented at the 2012 Symposium on VLSI Technology; doi:10.1109/VLSIT.2012.6242454.

ACKNOWLEDGMENTS

Funding: Supported by National Science Foundation of China grant 61621061, National Key Research & Development Program grants 2016YFA0201901 and 2016YFA0201902, and Beijing Municipal Science and Technology Commission grants D171100006617002 1-2. Also supported by University Grant Council contract AoE/P-04/08 of the Government of HKSAR (J.W., F.L., and H.G.) and NSERC of Canada (H.G.). We thank CalcuQuebec and Compute Canada for computation facilities. **Author contributions:** L.-M.P. and Z.Z. proposed and supervised the project; Z.Z., C.Q., and L.-M.P. designed the experiment; C.Q. performed the device fabrication and characterization; M.X. grew the nanotubes; B.D., L.L., and H.P. grew the graphene; F.L., C.Q., L.X., Z.Z., J.W., and H.G. built the model and simulated the devices; Z.Z., C.Q., and L.-M.P. analyzed the data and co-wrote the manuscript; and all authors discussed the results and commented on the manuscript. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/387/suppl/DC1
Materials and Methods
Figs. S1 to S10
References (*47–61*)

9 September 2017; resubmitted 4 April 2018
Accepted 25 May 2018
Published online 14 June 2018
10.1126/science.aap9195

NEUROSCIENCE

The hippocampal engram maps experience but not place

Kazumasa Z. Tanaka[1]*, Hongshen He[1,2], Anupratap Tomar[1]†, Kazue Niisato[1], Arthur J. Y. Huang[1], Thomas J. McHugh[1,2]*

Episodic memories are encoded by a sparse population of hippocampal neurons. In mice, optogenetic manipulation of this memory engram established that these neurons are indispensable and inducing for memory recall. However, little is known about their in vivo activity or precise role in memory. We found that during memory encoding, only a fraction of CA1 place cells function as engram neurons, distinguished by firing repetitive bursts paced at the theta frequency. During memory recall, these neurons remained highly context specific, yet demonstrated preferential remapping of their place fields. These data demonstrate a dissociation of precise spatial coding and contextual indexing by distinct hippocampal ensembles and suggest that the hippocampal engram serves as an index of memory content.

Numerous theories have attempted to link the physiology of the hippocampus with its role in episodic memory (*1*, *2*); however, experimental tests of these ideas remain scarce. One prevailing model suggests that the hippocampal memory trace contains rich information about the animal's current location within the cognitive domain, providing a spatial framework on which events and items can be anchored and related [cognitive map theory (*3*, *4*)]. This model is supported by data demonstrating that synaptic plasticity stabilizes these hippocampal maps and that animals can reliably recall these representations, even at remote time points (*5*, *6*). An alternative, although not mutually exclusive, hypothesis, the memory index theory, asserts that the hippocampal memory trace is primarily an index that provides rapid and efficient access to the content of an episodic memory stored in the neocortex (*7*, *8*). This theory is agnostic to the information content in the hippocampus, emphasizing its role in reactivating downstream cortical modules, with plasticity serving to establish a link between the hippocampal index and the neocortical activity pattern.

Although place cell physiology supports the cognitive map theory (*9*), behavioral studies using contextual fear conditioning [CFC (*10*)] may be better interpreted via the memory index theory. In CFC, the contextual representation is index-like because its formation requires conjunctive exposure to cues and its rapid recall does not require physical exploration of the environment (*11*). Moreover, the expression of activity-induced genes, such as c-*Fos*, is rapidly and robustly induced in a specific neuronal ensemble after presentation of novel configurations of stimuli and is thought to define the neuronal substrates of the contextual representation (*12*, *13*). Recent studies (*14*) further strengthen this view, demonstrating that inhibition of hippocampal engrams can block memory recall in contextual tasks, whereas activation can drive context-triggered behavior (*15*, *16*). The link between this contextual memory representation and precise locations or routes represented by place cells during exploration remains unknown.

We therefore conducted tetrode recordings from CA1 pyramidal cells in freely moving c-Fos-tTA (tetracycline-transactivator) transgenic mice infused with virus expressing channelrhodopsin2 and enhanced yellow fluorescent protein under the control of the Tet-responsive element (AAV-TRE-ChR2-EFYP). Doxycycline was removed from the diet of the mice, triggering labeling of c-Fos–expressing (positive) cells with ChR2, and place cell activity was recorded as mice explored a novel context (A; encoding context; Fig. 1A). The next day (12 to 14 hours later; fig. S1), the animals were reexposed to context A (recall) to examine the stability of the spatial map. CA1 was stimulated with pulses of blue light (10 mW, 0.5 Hz, 15 ms) to identify the subset of cells that expressed ChR2 as a result of c-Fos expression during the first session (Fig. 1B and fig. S2). In a separate cohort of mice, we verified that this OptID protocol did not affect excitability or spatial coding properties of the labeled neurons (fig. S3). Across the seven animals used in this study, 19.59 ± 2.65% of putative pyramidal cells exhibited light-induced spikes and thus were identified as c-Fos positive (Fig. 1C). This fraction is consistent with previous reports (*17*). Finally, the same mice explored a distinct environment (B) to examine context-specific activity (*18*).

We first examined the spatial firing of c-Fos–positive neurons as the animals explored the novel context A. The majority (79.31%) of these cells had place fields in the labeled context (Fig. 1D; see methods for place field criteria). However, a large fraction (75.53%) of the active place cells were not labeled with ChR2 (Fig. 1E). Although physiologically similar, with indistinguishable peak firing rates (Fig. 1F; pos: 6.99 ± 1.06 Hz; neg: 6.06 ± 0.53 Hz, $W = 640$, $p = 0.20$), c-Fos–positive place cells exhibited significantly higher mean firing rates (Fig. 1G; pos: 1.43 ± 0.22 Hz; neg: 0.77 ± 0.11 Hz, $W = 357$, $p = 5.4 \times 10^{-5}$). Accordingly, positive cells had larger place fields (Fig. 1H; pos: 28.00 ± 3.50 cm^2; neg: 17.97 ± 1.79 cm^2, $W = 517$, $p = 0.0085$) and on average, their spikes carried lower spatial information (Fig. 1I; pos: 0.69 ± 0.81 bits/spike; neg: 1.21 ± 0.83 bits/spike, $W = 1191$, $p = 0.001$).

We next analyzed the temporal structures of spike activity during exploration (A, encoding). Interspike interval (ISI) analysis revealed that spikes from engram cells were more likely to occur in bursts (3- to 15-ms ISI) (Fig. 2A). These neurons had significantly higher burst rates (Fig. 2B; pos: 16.80 ± 3.04 bursts/min; neg: 8.87 ± 1.44 bursts/min, $W = 1243$, $p = 0.00018$) and shorter average interburst intervals (IBIs) (Fig. 2C; IBI: pos: 6.55 ± 1.30 s; neg: 16.27 ± 2.15 s, $W = 415$, $p = 0.00042$). Burst events in c-Fos–positive neurons were preferentially spaced at ~125 ms, an interval corresponding to the 8-Hz theta rhythm in the hippocampus; thus, bursts from these neurons were significantly more theta modulated than bursts from c-Fos–negative cells (Fig. 2, D to F; theta modulation index, $W = 418$, $p = 0.0009$; theta power in burst spectrum, $W = 1151$, $p = 4.8 \times 10^{-5}$). We defined theta-burst events (TBEs) as bursts repeated at IBIs of 83 to 167 ms (6 to 12 Hz) and compared this property across the groups. During exploration of the novel context A, c-Fos–positive cells showed higher rates of TBEs than c-Fos–negative place cells (Fig. 2, G and H; pos: 5.44 ± 1.41 bursts/min; neg: 2.58 ± 0.53 bursts/min; $W = 417.5$, $p = 0.00046$) and longer repetitions of bursts (Fig. 2, I and J; pos: 2.68 ± 0.095 bursts/event; neg: 2.56 ± 0.052 bursts/event; $W = 652.5$, $p = 0.20$; Fig. 2I, max TBE length: pos: 7.13 ± 0.78 bursts; neg: 5.28 ± 0.34 bursts, $W = 542$, $p = 0.022$). These burst features were no longer prominent in ChR2-labeled neurons when the animals explored context B (fig. S5), suggesting their involvement in contextual encoding.

We then examined the temporal modulation of place cell spiking by the oscillatory population activity in the local field potential (LFP) (A, encoding). CA1 neurons can be entrained by the theta (6 to 12 Hz) rhythm, as well as by slow (30 to 50 Hz) and fast (55 to 85 Hz) gamma oscillations, which correlate with CA3 and entorhinal cortical inputs to CA1, respectively (*19*). Spikes from c-Fos–positive and –negative cells were similarly theta entrained and preferred the ascending phase of the oscillation (Fig. 3, A and B; mean preferred phase, pos: 229.9 ± 1.73°; neg: 199.2 ± 1.85°, $W = 0.42$, $p = 0.81$). However, TBE spikes in positive neurons preferentially occurred during the descending phase, whereas those in negative neurons remained locked to the ascending phase (Fig. 3, C to E; mean

[1]Laboratory for Circuit and Behavioral Physiology, RIKEN Center for Brain Science, 2-1 Hirosawa, Wakoshi, Saitama, Japan. [2]Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, Tokyo, Japan.
*Corresponding author. Email: kazumasa.tanaka@riken.jp (K.Z.T.); thomas.mchugh@riken.jp (T.J.M.)
†Present address: School of Physiology, Pharmacology and Neuroscience, University of Bristol, Bristol, UK.

preferred phase: pos: 170.7 ± 1.86°; neg: 325.7 ± 1.54°, $W = 62.66$, $p = 2.5 \times 10^{-14}$). Next, to address any differences in coupling of positive and negative neurons to CA1 inputs, we examined spikes during gamma events. Significantly more spikes from positive neurons occurred during fast gamma events, and a larger fraction of these neurons were phase locked to fast gamma oscillations compared to the c-Fos–negative population (Fig. 3F; percentage of spikes: pos: 2.62 ± 0.17%; neg: 2.19 ± 0.11%, $W = 1067$, $p = 0.0072$; percentage of phase-locked cells: pos: 26.09%; neg: 14.08%, chi-squared test, $p = 0.0014$). There was no significant difference in spiking or phase locking between the groups during slow gamma (Fig. 3G; percentage of spikes: pos: 5.81 ± 0.31%; neg: 5.53 ± 0.23%; $W = 888$, $p = 0.53$; percentage of phase-locked cells: pos: 26.09%; neg: 29.58%, chi-squared test, $p = 0.62$). These data suggest that although both populations of neurons are similarly tuned to CA3-driven excitation, c-Fos–positive neurons may be more responsive to input from the entorhinal cortex, resulting in the temporal shift in their distinctive theta-paced bursting.

Finally, we examined hippocampal activity during memory recall and assessed the stability of the spatial representation of engram cells when animals revisited the labeling context (A). Contrary to expectations of a spatial memory trace, many of these cells shifted their firing locations during the second visit (Fig. 4A and fig. S6). In positive neurons, the average correlation of the firing-rate maps from the encoding and recall sessions was close to zero, significantly lower than that from negative place cells (Fig. 4B; pos: 0.079 ± 0.070; neg: 0.31 ± 0.052, $W = 560$, $p = 0.010$). However, correlations of the mean firing rate, independent of position, were similar across both populations of cells (Fig. 4C; pos: 0.36 ± 0.053; neg: 0.36 ± 0.029, $W = 412$, $p = 0.81$), suggesting that c-Fos–positive neurons represent contextual information in a different manner. Therefore, we examined their activity in a distinct context. When animals explored context B, many c-Fos–positive cells remained silent (Fig. 4A and fig. S6; 43.5% of positive place cells had a peak rate of <1 Hz in context B; peak

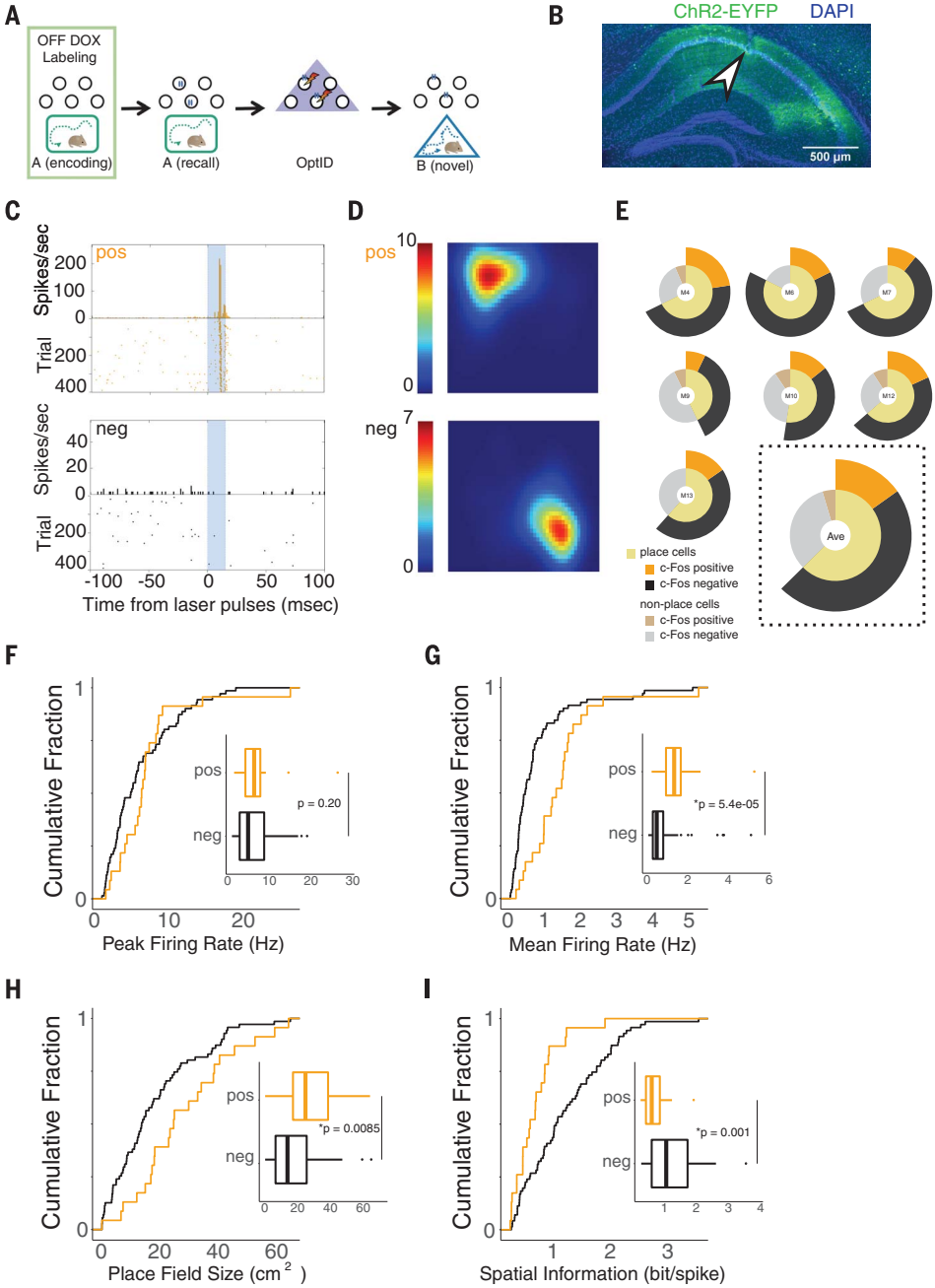


Fig. 1. c-Fos induction in a fraction of place cells during exploration of a novel context. (**A**) A schematic of the experimental protocol. Mice were exposed to a novel context (A, encoding) when doxycycline was removed from the diet (OFF DOX), followed by reexposure to the same context (A, recall). Twenty-four hours after A (encoding), labeled cells were identified through light-induced spikes. Finally, mice explored a distinct context B (novel). (**B**) A representative image of ChR2-EYFP (green) expression in the dorsal CA1 of the hippocampus. The white arrowhead indicates a tetrode location. DAPI, 4′,6-diamidino-2-phenylindole. (**C**) Peristimulus time histogram of representative single units classified as c-Fos–labeled (pos) or not labeled (neg). Blue area represent light on epoch (15 ms). (**D**) Firing-rate maps of representative place cells showing location-specific firing during exploration in context A (encoding). (**E**) Pie plots describing percentages of each cell type in recorded animals ($N = 7$), and their average (dashed square). (**F** to **I**) Cumulative density plots with inset box plots comparing peak firing rates (F), mean firing rates (G), place field size (H), or spatial information (I) between c-Fos–positive (orange) and –negative (black) place cells. Box plots show median, first, and third quantiles, and minimum and maximum values within 1.5× the interquantile range (IQR) from each quantile.

firing rate in context A: 6.99 ± 1.06 Hz; context B: 3.42 ± 0.76 Hz, W = 392, p = 0.005). Consequently, there were fewer c-Fos–positive place cells, and the spatial information of these neurons was smaller than in context A (Fig. 4D, Friedman rank sum test, Friedman chi-squared = 6, p = 0.014; two-way analysis of variance, significant interaction, $F_{(1,89)}$ = 4.9, p = 0.029; Fig. 4E, spatial information: context A: 0.89 ± 0.21 bit/s; context B: 0.32 ± 0.071 bit/s, W = 396, p = 0.0033). By contrast, the average spatial information carried by c-Fos–negative place cells was unchanged (Fig. 4F; context A: 0.51 ± 0.064 bit/s; context B: 0.37 ± 0.044 bit/s, W = 2643, p = 0.30), although these cells typically shifted the location of their spatial firing ("remapped") in the new context (Fig. 4G; spatial correlation, A/A: 0.31 ± 0.052; A/B: 0.013 ± 0.052, W = 982, p = 0.00029). These data are consistent with the hypothesis that engram cells do not necessarily represent reliable spatial information about the external world, but rather through their net activity serve as an index to episodic information stored elsewhere in the brain (*7*, *8*). We therefore sought to discriminate

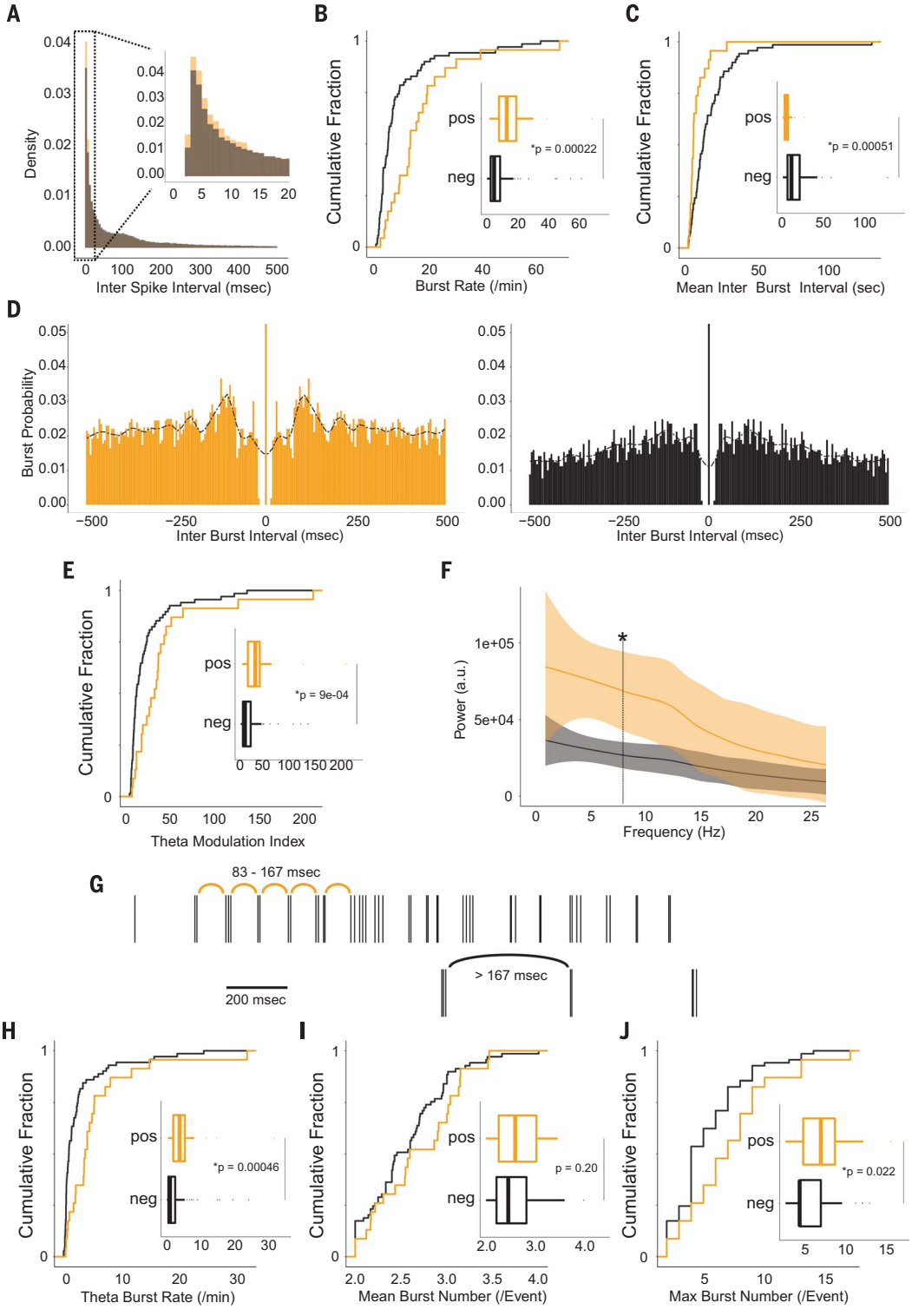


Fig. 2. Theta-paced burst activity in c-Fos–positive place cells. (**A**) Histograms showing densities of interspike intervals from all c-Fos–positive (orange) or –negative (black) place cells. (**B** and **C**) Cumulative density plots and box plots comparing burst rates (B) and mean interburst intervals (C) of the two cell types. (**D**) Autocorrelogram of burst activity from an example positive (orange) and negative (black) place cell. Local polynomial regression lines are plotted over the histograms. (**E**) Theta modulation index of burst autocorrelograms. (**F**) Power spectrograms of burst trains. Vertical line denotes 8 Hz. (**G**) Example raster plots showing theta-paced burst events (TBEs; IBI = 83 to 167 ms) and repetitive bursts outside of the criteria (IBI < 83 ms or IBI > 167 ms). (**H** to **J**) Cumulative density plots and box plots comparing TBE rates (H), mean burst number per TBE (I), and maximum burst number within a single TBE (J) of the two cell types. Box plots show median, first, and third quantiles, and minimum and maximum values within 1.5× the IQR from each quantile.

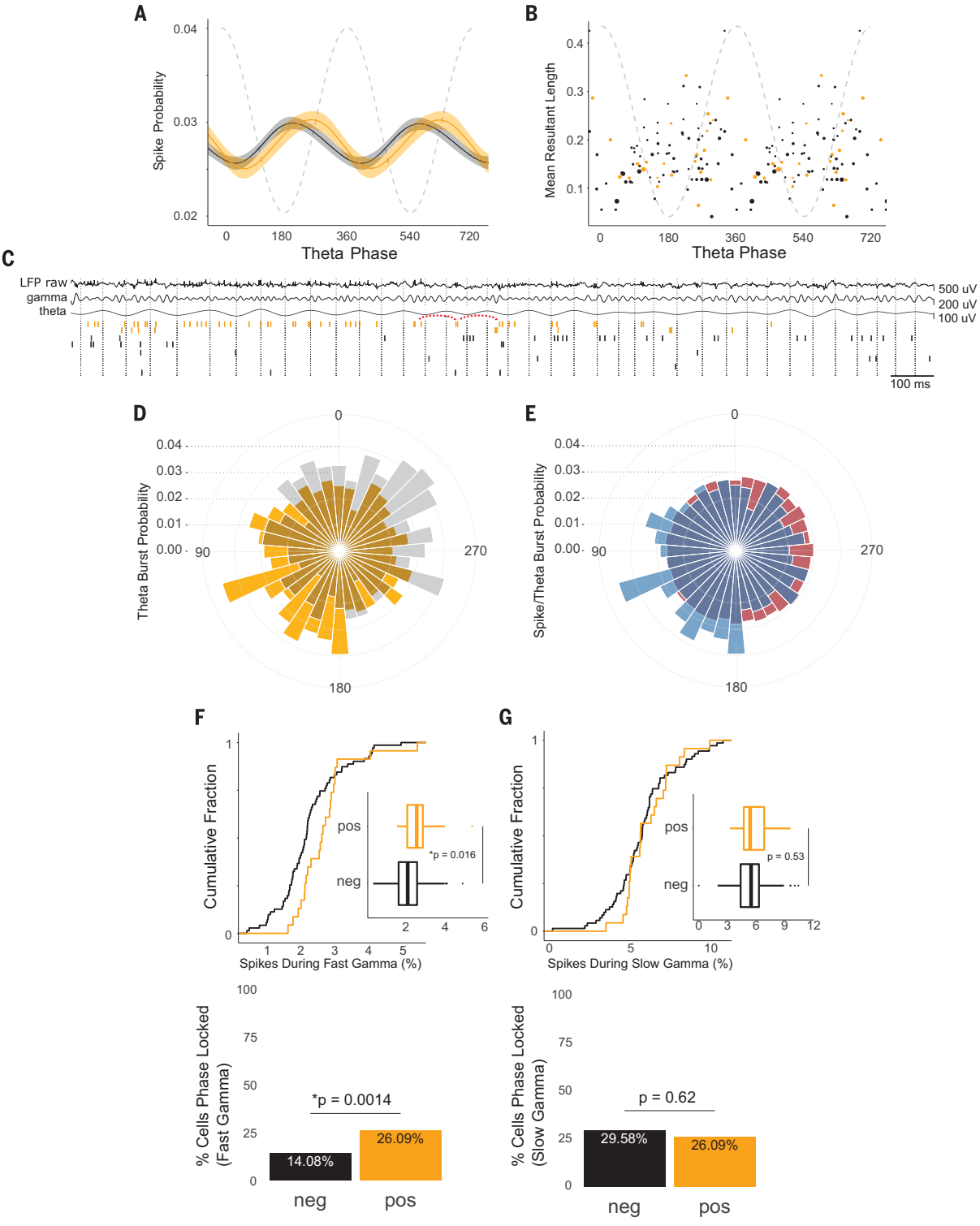


Fig. 3. LFP modulation of spikes from c-Fos–positive and –negative place cells. (**A**) Spike probability of the two types of cells over phases of theta oscillation (6 to 12 Hz) [positive (orange) or negative (black); only significantly phase-locked cells were used]. Envelopes represent confidence intervals; the dashed line represents theta. (**B**) Mean resultant length of c-Fos–positive (orange) or –negative (black) cells plotted as a function of preferred theta phase. Dot size represents mean firing rate of that cell. Only significantly phase-locked cells are plotted (positive: 22 cells; negative: 68 cells). (**C**) Example spike raster plots with LFP traces [raw, theta filtered (6 to 12 Hz), gamma filtered (30 to 85 Hz)]. Red arcs represent a TBE occurring at the descending phase of theta. (**D**) A rose plot showing TBE spike probability across theta phases [positive (orange) or negative (black), 10° bins]. (**E**) A rose plot showing spike probability of c-Fos–positive place cells over theta phases (TBE spikes are in blue, all spikes in red). (**F**) Percentage of spikes occurring during fast gamma events (top), and percentage of cells phase locked to fast gamma oscillations [bottom; positive (orange) or negative (black)]. (**G**) Percentage of spikes occurring during slow gamma events (top), and percentage of cells phase locked to slow gamma oscillations [bottom; positive (orange) or negative (black)].

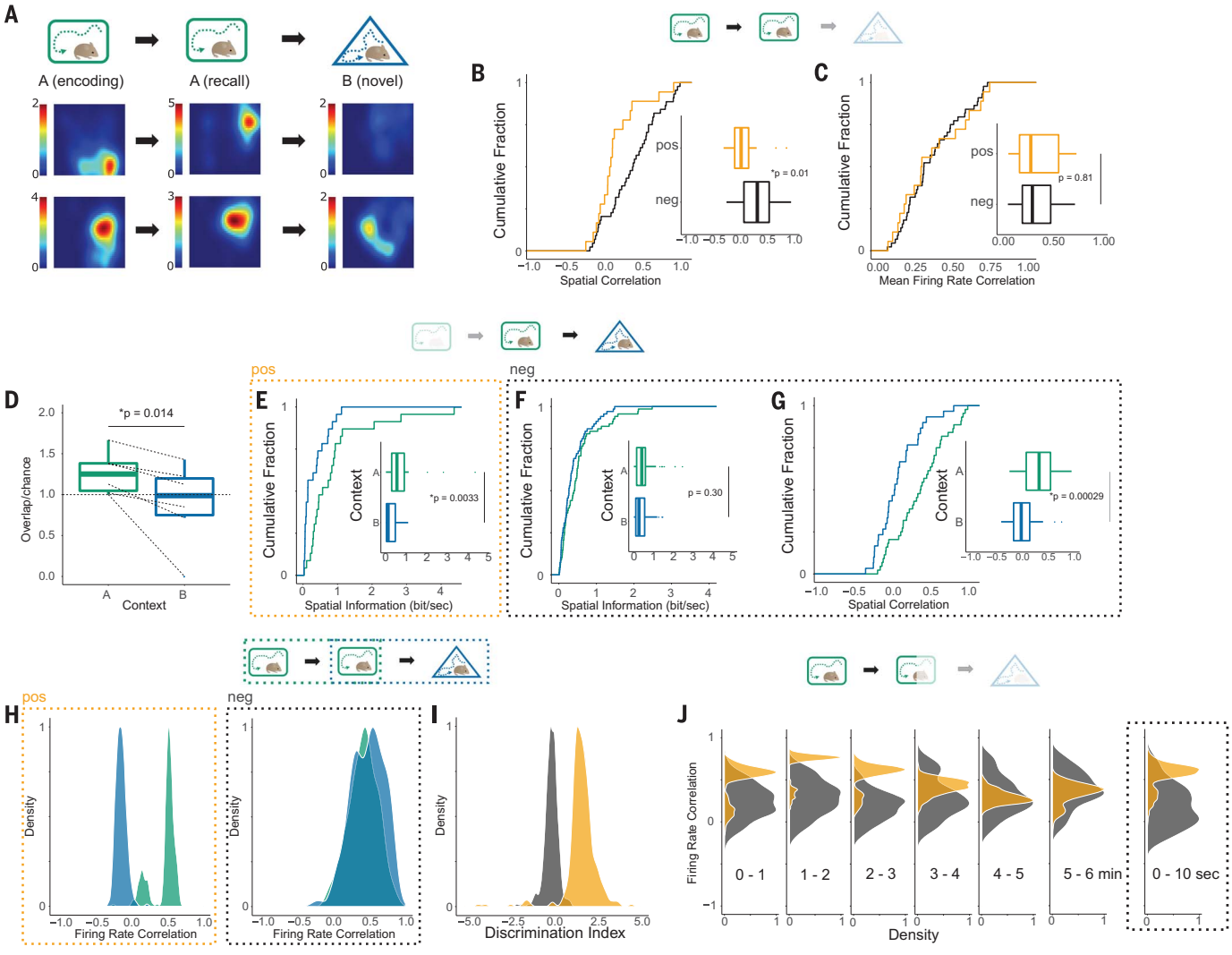


Fig. 4. Different responses of c-Fos–positive and –negative place cells to the encoding context and a distinct novel context. (**A**) Firing-rate maps of c-Fos–positive (top) or –negative (bottom) place cell activity during encoding (left), reexposure to the encoding context (middle), or exploration of a distinct novel context (right). (**B**) Spatial correlation between place maps in A (encoding) and A (recall) [positive (orange) or negative (black)]. (**C**) Mean firing-rate correlation between place maps in A (encoding) and A (recall) [positive (orange) or negative (black)]. (**D**) An animal-by-animal box plot showing observed overlap between c-Fos–positive cells and place cells normalized by those expected by chance, comparing context A (encoding, green) versus B (novel, blue) sessions. Spatial information of (**E**) c-Fos–positive or (**F**) c-Fos–negative place cells when animals explore context A (encoding; green) or B (novel; blue). (**G**) Spatial correlations of c-Fos–negative place cells in A (encoding; green) and B (novel; blue). (**H**) Distribution of ensemble firing-rate correlations of c-Fos–positive or –negative place cells between context A (encoding) – A (recall) (green) versus A (encoding) – B (novel) (blue). (**I**) Distribution of discrimination indexes obtained by c-Fos–positive (orange) or –negative (black) ensemble firing-rate correlations. (**J**) Instantaneous ensemble firing-rate correlations of the initial 6 min of the context A (recall) session binned by 1 min (left), or that of the first 10 s (right). c-Fos–positive place cells are in orange, c-Fos–negative cells in black. For (H) to (J), values are plotted as probability densities (scaled). Box plots show median, first, and third quantiles, and minimum and maximum values within 1.5× the IQR from each quantile.

contexts only from firing-rate correlations of a subpopulation of neurons. We randomly subsampled c-Fos–positive neurons and an equivalent number of c-Fos–negative place cells and calculated the correlation of the mean firing rates of each population between context A (encoding) and A (recall) or between context A (encoding) and B (novel), repeating the procedure 1000 times (see methods). As a population, the firing rate of c-Fos–positive place cells reliably discriminated the two contexts, whereas the firing rate of the ensemble of negative cells did not (Fig. 4H; probability of finding a higher correlation in A/B than A/A, $p = 0.002$ for positive ensemble, $p = 0.57$ for negative ensemble). This was also evident in the discrimination index (Fig. 4I; probability of finding a greater index in the negative than the positive ensemble, $p = 0.049$). In contrast to the negative ensemble, context discrimination from firing-rate correlations of c-Fos–positive cells was prominent early in the recall session, demonstrating high correlation to the average rate of the encoding session during the first few minutes, or even the first 10 s (Fig. 4J and fig. S7; probability of finding a higher correlation in the negative than the positive ensemble, $p = 0.055$, 0.038, 0.011, 0.230, 0.906, and 0.754 for 1- to 6-min bins, $p = 0.046$ for the first 10-s bin), suggesting the rapid reactivation of contextual representation by the subset of engram neurons. These data demonstrate that engram cells are place cells with lower spatial stability and accuracy, but with activity that can reliably and quickly reflect contextual identity.

When animals explore a novel context, an engram is formed by a fraction of the place cells,

suggesting that the plasticity involved in engram formation is distinct from that for formation of place fields. During encoding in the labeled context, engram neurons demonstrate repetitive theta-frequency bursts. In vitro, this pattern of action potentials is known to produce a form of long-term potentiation mediated by brain-derived neurotrophic factor (BDNF) secretion and dependent on c-Fos expression (*20*, *21*). c-Fos expression and theta bursts have been observed in CA1 when animals perform a non–spatial memory task (*22*, *23*), suggesting that this activity may be a signature of hippocampal engrams indexing broader types of memories.

Our study does not support the idea that the hippocampal engram simply encodes spatial memory because these neurons do not maintain their firing locations in the same environment across time. Instead, we propose that these cells may serve as an index for contextual memory through a shift of firing rates. Compared to other place cells, during recall these c-Fos–positive neurons rapidly spike at rates that are highly correlated to their activity during the encoding period and are preferentially silent in a distinct context. This feature may explain why optogenetic stimulation of the hippocampal engram can influence contextual memory recall (fig. S8) (*24*), despite lacking temporally meaningful pattern of spikes. Inhibition of the CA1 engram reduces downstream cortical reactivation (*25*), consistent with a role in indexing. Our data suggest that specific ensembles of hippocampal neurons simultaneously support distinct domains of episodic memories—spatially reliable c-Fos–negative neurons for configurational coding and spatially unstable engrams for contextual indexing.

REFERENCES AND NOTES

1. M. L. Shapiro, H. Eichenbaum, *Hippocampus* **9**, 365–384 (1999).
2. G. Buzsáki, E. I. Moser, *Nat. Neurosci.* **16**, 130–138 (2013).
3. J. O'Keefe, L. Nadel, *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).
4. O. Jensen, J. E. Lisman, *J. Neurophysiol.* **83**, 2602–2609 (2000).
5. C. Kentros *et al.*, *Science* **280**, 2121–2126 (1998).
6. T. J. McHugh, K. I. Blum, J. Z. Tsien, S. Tonegawa, M. A. Wilson, *Cell* **87**, 1339–1349 (1996).
7. T. J. Teyler, P. DiScenna, *Behav. Neurosci.* **100**, 147–154 (1986).
8. T. J. Teyler, J. W. Rudy, *Hippocampus* **17**, 1158–1169 (2007).
9. R. M. Grieves, K. J. Jeffery, *Behav. Processes* **135**, 113–131 (2017).
10. J. J. Kim, M. S. Fanselow, *Science* **256**, 675–677 (1992).
11. J. W. Rudy, R. C. O'Reilly, *Cogn. Affect. Behav. Neurosci.* **1**, 66–82 (2001).
12. J. F. Guzowski, B. L. McNaughton, C. A. Barnes, P. F. Worley, *Nat. Neurosci.* **2**, 1120–1124 (1999).
13. M. VanElzakker, R. D. Fevurly, T. Breindel, R. L. Spencer, *Learn. Mem.* **15**, 899–908 (2008).
14. L. G. Reijmers, B. L. Perkins, N. Matsuo, M. Mayford, *Science* **317**, 1230–1233 (2007).
15. S. Tonegawa, X. Liu, S. Ramirez, R. Redondo, *Neuron* **87**, 918–931 (2015).
16. S. A. Josselyn, S. Köhler, P. W. Frankland, *Nat. Rev. Neurosci.* **16**, 521–534 (2015).
17. K. K. Tayler, K. Z. Tanaka, L. G. Reijmers, B. J. Wiltgen, *Curr. Biol.* **23**, 99–106 (2013).
18. Local field potential and single-unit spiking were recorded during every session, as well as during pre- and postexposure rest to ensure recording stability (fig. S4).
19. L. L. Colgin *et al.*, *Nature* **462**, 353–357 (2009).
20. E. Edelmann *et al.*, *Neuron* **86**, 1041–1054 (2015).
21. B. Kuzniewska *et al.*, *Mol. Cell. Biol.* **33**, 2149–2162 (2013).
22. E. Amin, J. M. Pearce, M. W. Brown, J. P. Aggleton, *Eur. J. Neurosci.* **24**, 2611–2621 (2006).
23. T. Otto, H. Eichenbaum, S. I. Wiener, C. G. Wible, *Hippocampus* **1**, 181–192 (1991).
24. T. J. Ryan, D. S. Roy, M. Pignatelli, A. Arons, S. Tonegawa, *Science* **348**, 1007–1013 (2015).
25. K. Z. Tanaka *et al.*, *Neuron* **84**, 347–354 (2014).

ACKNOWLEDGMENTS

We thank J. P. Johansen (RIKEN CBS) and B. J. Wiltgen (UC Davis) for helpful comments on the manuscript. We thank R. Boehringer for technical support on tetrode recording, D. Polygalov for sharing scripts, M. Fujisawa and T. Tokiwa for daily assistance, the RIKEN Advanced Manufacturing Team for their assistance in microdrive production, and all the members of the Lab for Circuit and Behavioral Physiology (McHugh Lab) for advice. **Funding:** This work was supported by JSPS (Japan Society for the Promotion of Science) Grant-in-Aid for Young Scientists (B) (17K13272) (K.Z.T.), Brain Science Foundation Research Grant (K.Z.T.), the RIKEN Special Postdoctoral Researchers Program (K.Z.T.), the Uehara Memorial Foundation Research Grant (K.Z.T.), Takeda Science Foundation Visionary Research Grant (K.Z.T.), Grant-in-Aid for Scientific Research on Innovative Areas from MEXT (the Ministry of Education, Culture, Sports, Science and Technology of Japan) (17H05591, 17H05986) (T.J.M.), and RIKEN BSI and CBS (T.J.M.). **Author contributions:** K.Z.T. conceived the study and mainly conducted experiments and collected data under guidance and supervision of T.J.M. K.Z.T., H.H., and T.J.M. analyzed data. A.T. and K.N. contributed to data collection. A.J.Y.H. produced all AAV vectors. K.Z.T. and T.J.M. wrote the paper. All authors discussed and commented on the manuscript. **Competing interests:** No competing interests declared. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials. Data are archived on the servers of the Laboratory for Circuit and Behavioral Physiology at the RIKEN Center for Brain Science and are accessible at http://cbp.brain.riken.jp/tanaka_et_al.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/392/suppl/DC1
Materials and Methods
Figs. S1 to S8
References (*26*–*33*)

9 March 2018; accepted 20 June 2018
10.1126/science.aat5397

EUSOCIALITY

Social regulation of insulin signaling and the evolution of eusociality in ants

Vikram Chandra[1]*†, Ingrid Fetter-Pruneda[1]*†, Peter R. Oxley[1,2], Amelia L. Ritger[1], Sean K. McKenzie[1,3], Romain Libbrecht[1,4], Daniel J. C. Kronauer[1]*

Queens and workers of eusocial Hymenoptera are considered homologous to the reproductive and brood care phases of an ancestral subsocial life cycle. However, the molecular mechanisms underlying the evolution of reproductive division of labor remain obscure. Using a brain transcriptomics screen, we identified a single gene, *insulin-like peptide 2* (*ilp2*), which is always up-regulated in ant reproductives, likely because they are better nourished than their nonreproductive nestmates. In clonal raider ants (*Ooceraea biroi*), larval signals inhibit adult reproduction by suppressing *ilp2*, thus producing a colony reproductive cycle reminiscent of ancestral subsociality. However, increasing ILP2 peptide levels overrides larval suppression, thereby breaking the colony cycle and inducing a stable division of labor. These findings suggest a simple model for the origin of ant eusociality via nutritionally determined reproductive asymmetries potentially amplified by larval signals.

Eusocial insects exhibit a reproductive division of labor in which queens lay eggs and workers perform other tasks (*1*). Eusociality in ants, and in many other Hymenoptera, likely evolved from a subsocial state in which a female wasp would lay an egg and then care for the resulting larva until pupation (*1*–*3*). Such brood care may have been induced by larval signals, and observations of extant subsocial wasps are consistent with this scenario (*2*–*4*). This temporal reproductive and behavioral plasticity was then modified into a fixed reproductive asymmetry between queens and workers in eusocial colonies (*2*, *5*). This raises three important mechanistic questions: (i) How are subsocial reproductive cycles regulated? (ii) How is the eusocial reproductive division of labor regulated—i.e., what allows queens to lay eggs but prevents workers from doing so? (iii) What is the evolutionary trajectory that gave rise to fixed eusocial division of labor from subsocial cycles? Here we suggest that, in ants, evolutionary innovations in insulin signaling may have played a crucial role in each case.

Eusociality evolved once in a common ancestor of ants and, with the exception of a few derived social parasites, all extant ants are eusocial (*6*) (Fig. 1). To identify conserved potential regulators of division of labor between reproduction and brood care in ants, we conducted an unbiased screen for differentially expressed genes between whole brains or heads of reproductives and nonreproductives across seven ant species, including four previously published datasets (Fig. 1 and tables S1 and S2) (*7*–*11*). We sampled a range of reproductive strategies, from species with morphologically distinct queens and workers to queenless species. Among all 5581 identified single-copy orthologs, we found only one such gene: *insulin-like peptide 2* (*ilp2*). *ilp2* was always significantly up-regulated in reproductives (Fig. 1). Thus, the differential expression of *ilp2* is likely conserved across ants. Consequently, the most recent common ancestor of ants likely had *ilp2* expression that was high in reproductives and low in nonreproductives.

Although our approach is conservative and probably misses genes, it has the advantage of eliminating false positives. When we relaxed the statistical stringency for classifying genes as differentially expressed, our screen still returned *ilp2* as the single candidate gene (fig. S1). Relaxing other inclusion criteria revealed additional genes that might be expected to vary with reproductive state. For example, a total of 24 genes were consistently differentially expressed in subsets of five of the seven studied species (fig. S2 and table S3). This list includes *insulin-like peptide 1* (*ilp1*), as well as other genes implicated in insulin signaling (fig. S3 and table S3). Non–single-copy orthologs were excluded from our screen. One example is *vitellogenin* (*vg*), a gene that has undergone repeated duplications in ants (*12*). The vitellogenin protein is a lipid carrier that provisions developing oocytes with yolk and constitutes a reliable indicator of female reproductive activity (*12*, *13*). Studies of bees and other insects have shown that vitellogenin interacts with insulin signaling (*14*–*16*). *vg* indeed showed consistently higher expression in reproductives in our screen, even though this difference was not statistically significant in two of the ponerines (fig. S3). These findings further bolster the conclusion that insulin signaling played a major role in the evolution of reproductive division of labor in ants.

Insulin regulates reproduction and food-seeking behavior across a wide range of organisms, making it a prime candidate for the regulation of subsocial cycles and eusocial division of labor (*17*). Most studied hymenopterans have two ILPs: ILP1 and ILP2 (fig. S4). Whereas ILP1 resembles insulin-like growth factor, ILP2 is similar to canonical insulin (fig. S5) (*11*). In other holometabolous insects, these ILPs regulate larval growth, adult metabolism, and reproduction (*17*–*19*). Moreover, caste determination in most ant species relies on nutritional asymmetries during development: Queen-destined larvae eat more than worker-destined larvae, which likely explains how queens acquire higher ILP2 levels (*20*). A study of *Diacamma* sp. found that the asymmetry in reproductive potential between ants was correlated with insulin receptor expression in the ovaries (*21*). This suggests a possible secondary mode of reproductive control downstream of ILPs that may augment the initial reproductive asymmetry reflected by differential *ilp2* expression in the brain.

ILPs have not been studied functionally in eusocial insects in the context of reproductive division of labor between adults. However, insulin signaling has been implicated in other contexts, such as caste development and nonreproductive division of labor (*18*, *22*–*24*). Current data from wasps and bees do not typically indicate that *ilp2* is differentially expressed between adult queens and workers, suggesting that this expression pattern may be ant specific (table S5). This apparent inconsistency may be explained by the fact that eusociality has evolved independently in ants, bees, and wasps (*1*). Therefore, though insulin signaling may have been co-opted repeatedly during social evolution, the details likely differ between independent lineages.

We used the clonal raider ant *Ooceraea biroi* to study ILP2 in ants. *O. biroi* has secondarily lost queens, resulting in a species in which workers reproduce synchronously and asexually (*13*, *25*). Colonies alternate between reproductive and brood care phases. This colony cycle is regulated by the periodical presence of larvae, which suppress reproduction and induce brood care behavior in adults, and is reminiscent of the subsocial cycle presumed to precede eusociality in ants. Despite this unusual biology, *O. biroi* is eusocial. Workers display cooperative brood care, colonies contain overlapping generations of adults, and reproductive asymmetry exists within colonies (*25*).

We found that antibody staining of ILP2 exclusively localized to the brain, primarily in a single medial cluster of ~15 cells in the pars intercerebralis (Fig. 2, A to C, and fig. S6). These insulin-producing cells (IPCs) coincide in location with those of other insects (*26*, *27*). Axons likely project to the corpora cardiaca, the only other brain region staining positive for ILP2 (figs. S6 to S8). We quantified ILP2 in the IPCs and found that its levels are higher in the brood

[1]Laboratory of Social Evolution and Behavior, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA. [2]Samuel J. Wood Library, Weill Cornell Medicine, 1300 York Avenue, New York, NY 10065, USA. [3]Department of Ecology and Evolution, University of Lausanne, Biophore Building, 1015 Lausanne, Switzerland. [4]Institute of Organismic and Molecular Evolution, Johannes Gutenberg University, Johannes-von-Müller-Weg 6, 55128 Mainz, Germany. ***Corresponding author. Email: vchandra@rockefeller.edu (V.C.); ifetter@rockefeller.edu (I.F.-P.); dkronauer@rockefeller.edu (D.J.C.K.)** †These authors contributed equally to this work.

care than in the reproductive phase (Fig. 2D and fig. S6). Peptide levels are thus anticorrelated with transcription. This pattern is known from *Drosophila melanogaster*, in which the rate of ILP secretion correlates with the rate of *ilp* transcription (*27*). This suggests that the mechanisms of *ilp* expression and ILP secretion are conserved in holometabolous insects.

Because larvae regulate the *O. biroi* colony cycle, we asked whether larval communication altered *ilp2* expression in adults. When larvae are removed from colonies in the brood care phase, *ilp2* expression levels in adult brains increase markedly within 12 hours (Fig. 2E) (*28*). This increase occurs under identical nutritional conditions. Conversely, when ants in the reproductive phase are given larvae, their *ilp2* levels decrease (Fig. 2E). *vgq*, the vitellogenin gene upregulated in ant queens, responds similarly, albeit more slowly, to these changes (fig. S9A), raising the possibility that ILP2 regulates reproduction at least partly by acting on *vgq*. Although this experiment is highly suggestive, the addition of larvae was always correlated with the removal of pupae, and changes in expression occurring after the 24-hour time point were confounded by nutritional differences. We therefore repeated this experiment without pupae and under nutritionally controlled conditions. We removed larvae from colonies in the brood care phase, waited until the ants in these colonies activated their ovaries, and then compared brain gene expression between these and control colonies. Again, the removal of larvae increased *ilp2* (Fig. 2F) and *vgq* (fig. S9B) expression. This finding suggests that social signals can mediate insulin signaling independently of internal nutritional state and that this is a key regulatory mechanism underlying the *O. biroi* colony cycle. Given the conserved association of caste and *ilp2* expression in all ants, social regulation of *ilp2* may also underlie the life cycle of the subsocial ancestor.

In *D. melanogaster*, insulin signaling is necessary and sufficient to regulate the terminal differentiation of germline stem cells into oocytes. Moreover, it promotes yolk uptake in developing oocytes and is crucial for ovary activation (*29*). It is therefore plausible that the differential expression of *ilp2* in ants has a causal role in regulating ovary activation and reproductive division of labor. We further hypothesized that if the regulation of *ilp2* were freed, at least partially, from larval control, this would yield ants whose physiology is less susceptible to reproductive suppression. Such a mechanism would allow the evolution of distinct reproductive and nonreproductive castes from an ancestral subsocial cycle. To test

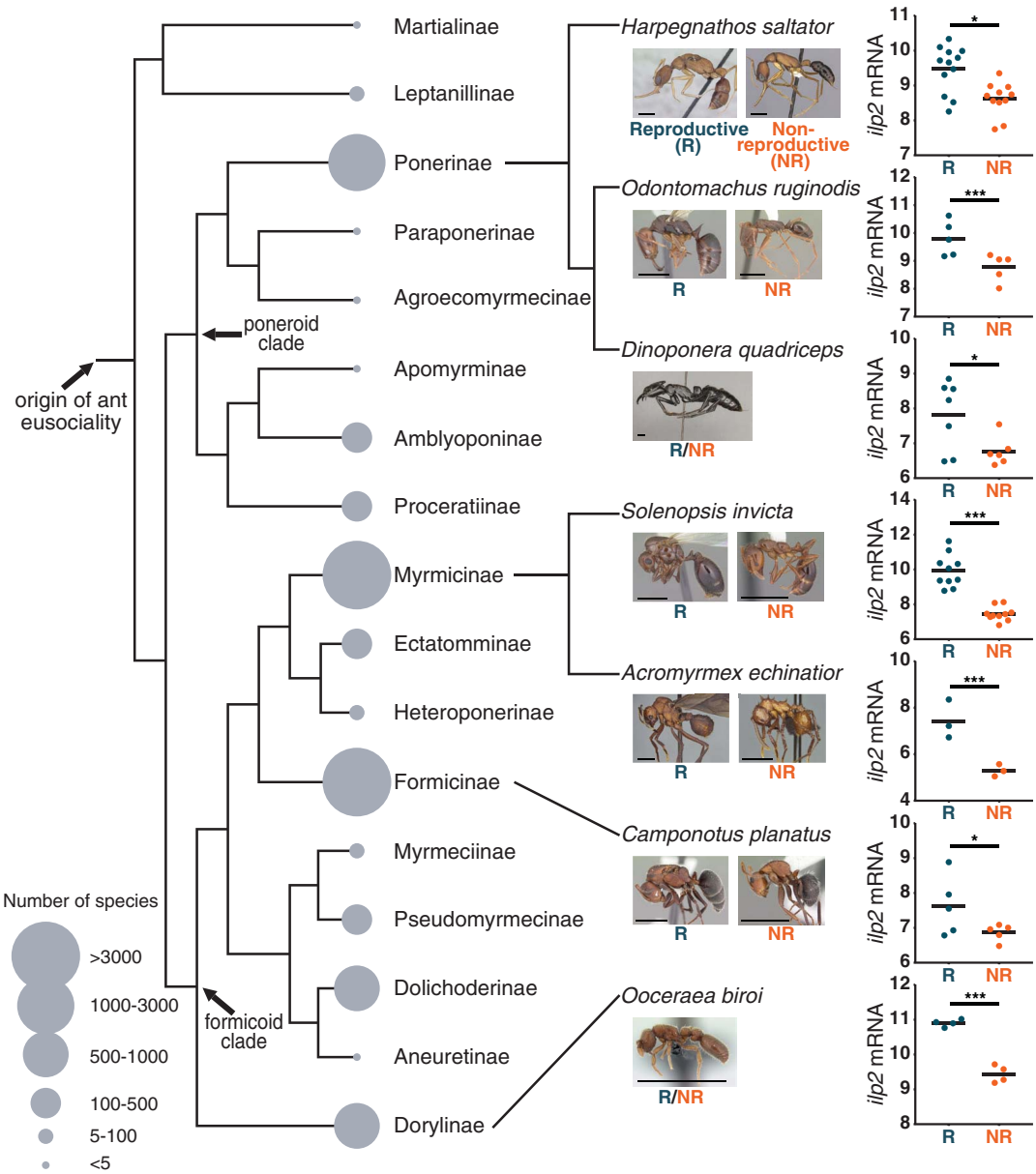


Fig. 1. Brain gene expression in seven ant species identifies one conserved differentially expressed gene. The figure shows the summary cladogram of the seven ant species used in this study in the context of the entire ant phylogeny with all subfamilies labeled. Five of the focal species have queens, whereas two (*Dinoponera quadriceps* and *O. biroi*) are queenless. Although *Harpegnathos saltator* is not queenless, the data compared reproductive and nonreproductive workers (table S1). The dot plots show variance-stabilized transformed read counts for *ilp2*. Blue and orange dots indicate reproductive (R) and nonreproductive (NR) ants, respectively. Horizontal bars indicate means, and asterisks indicate statistically significant differences between groups (Wald test: $*P < 0.05$; $***P < 0.001$). All photos except for those of *Acromyrmex echinatior* are from A. Nobile, S. Hartman, and E. Prado (www.antweb.org). Scale bars, 2 mm. The phylogeny is based on (*30*). Species numbers are from (*6*).

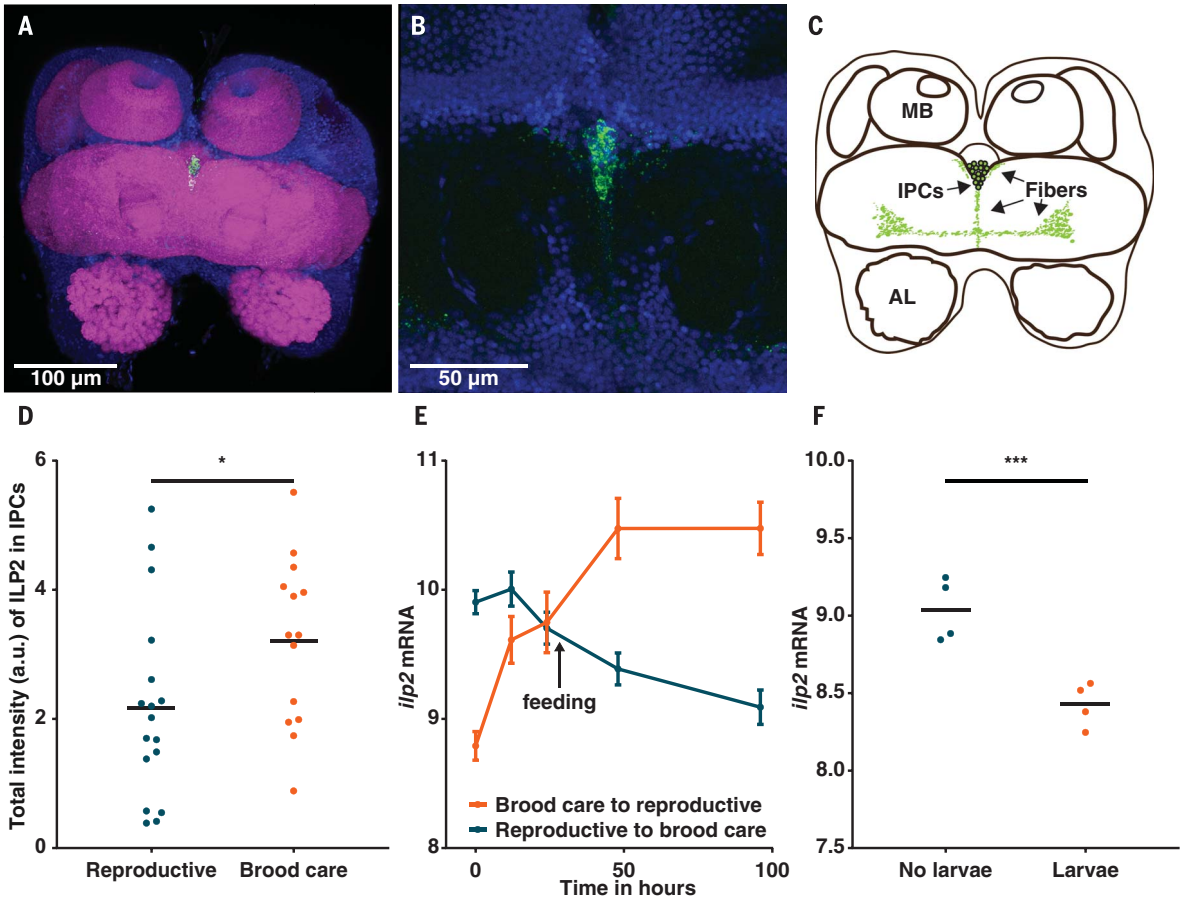


Fig. 2. Larvae regulate *ilp2* in adults. (**A** to **C**) Immunohistochemistry with antibody against ILP2 (anti-ILP2) on an *O. biroi* brain localizes ILP2 peptide to a single cluster of insulin-producing cells (IPCs) in the pars intercerebralis (body-axis dorsal view). Green, anti-ILP2; blue, DAPI (4′,6-diamidino-2-phenylindole); magenta, phalloidin. MB, mushroom body; AL, antennal lobe. (**D**) Total intensity of ILP2 in the IPCs is higher in the brood care phase than in the reproductive phase ($n \geq 14$ ants, *t* test; $^{*}P = 0.046$). a.u., arbitrary units. (**E**) RNA sequencing (RNA-seq) time course shows that the addition of larvae down-regulates *ilp2*, whereas the removal of larvae up-regulates *ilp2* [$n \geq 4$ biological replicates, time-transition interaction, likelihood ratio test with 5% false discovery rate (FDR) correction; $P < 10^{-15}$]. The arrow indicates when ants with larvae were fed (i.e., changes in expression beyond that time point are confounded by differences in nutrition). Error bars depict SEM. Data are from (*28*). (**F**) RNA-seq on ant brains shows that under nutritionally controlled conditions, *ilp2* is up-regulated 8 days after larvae are removed from *O. biroi* workers in the brood care phase ($n = 4$ biological replicates, Wald test with 5% FDR correction; $^{***}P < 10^{-6}$). Data are variance-stabilized transformed read counts. Horizontal bars indicate means.

this hypothesis, we injected synthetic *O. biroi* ILP2 mature peptide into workers in colonies with larvae. As a control, we injected the inactive B chain of this peptide (fig. S11A) (*19*). Injecting ILP2 mature peptide caused strong ovary activation despite the presence of larvae (Fig. 3, A to C, and fig. S10A). Higher doses of ILP2 caused ants to develop more eggs simultaneously (fig. S10, B and C), suggesting that quantitative differences in ILP2 levels vary the ants' positions along a spectrum of reproductive potential. To ensure that ILP2 does not have inhibitory effects during the opposite phase of the colony cycle, we injected ants in the reproductive phase with ILP2 and found no detectable effect on ovary state (fig. S11, B and C).

Finally, we hypothesized that as developmental nutritional asymmetries determine caste in most ants, this might be a general and natural mechanism that produces asymmetries in baseline adult ILP2 levels and consequently in reproductive potential. Whereas most *O. biroi* workers have two ovarioles, some individuals ("intercastes") have four or more (*25*) (fig. S12, A and B). We found that these differences can be determined by the amount of food a larva receives (fig. S13). Intercastes have longer and more active ovaries compared with those of regular workers in the brood care phase, suggesting intercastes are less sensitive to larval signals that suppress ovarian activity (Fig. 4A and fig. S12C). This finding is consistent with previous work showing that some intercastes fail to regress their ovaries during the brood care phase (*25*). Additionally, we found that the IPCs of intercastes contained more ILP2 than those of regular workers (Fig. 4, B and C). As we have shown above, ILP2 peptide levels are negatively correlated with *ilp2* expression, ovary state, and, by extension, circulating ILP2 levels in workers between the different phases of the cycle, probably owing to higher rates of peptide release during the reproductive phase (Fig. 2D). The phase-matched comparisons between different types of workers, on the other hand, show that intercastes consistently have higher ILP2 levels in their IPCs, and, given their more active ovaries and decreased sensitivity to larval signals (*25*), it is likely that they also have consistently higher levels of ILP2 in circulation.

How the ancestral subsocial cycle was regulated remains unknown. However, assuming that similar mechanisms underlie the *O. biroi* colony cycle, our findings suggest a plausible scenario for the evolution of ant sociality. First, during the transition from solitary to subsocial life, some signaling systems (probably including insulin signaling) in adults must have become responsive to larval signals. This shift allowed behavioral and physiological responses in adults to be appropriately modified for the nutritional requirements of the larvae. During the transition from subsocial to eusocial life, increased developmental variation may have caused some

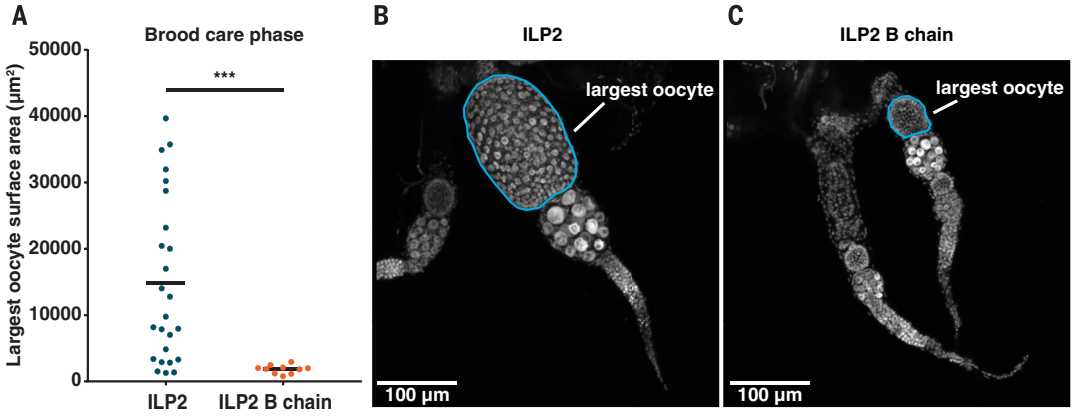


Fig. 3. ILP2 supplementation overrides larval suppression of adult reproduction. (**A**) Workers injected with 100 µM ILP2 in the brood care phase activate their ovaries relative to controls injected with 100 µM ILP2 B chain, despite being in contact with larvae [$n \geq 10$, Welch's t test with Bonferroni correction (related data in fig. S8); $^{***}P = 0.0005$]. (**B** and **C**) Confocal images of ovaries from ants injected with either 100 µM ILP2 (B) or 100 µM ILP2 B chain (C). Shown are the pairs of ovaries closest to the mean value from each treatment; the largest oocyte in each pair is circled in blue.

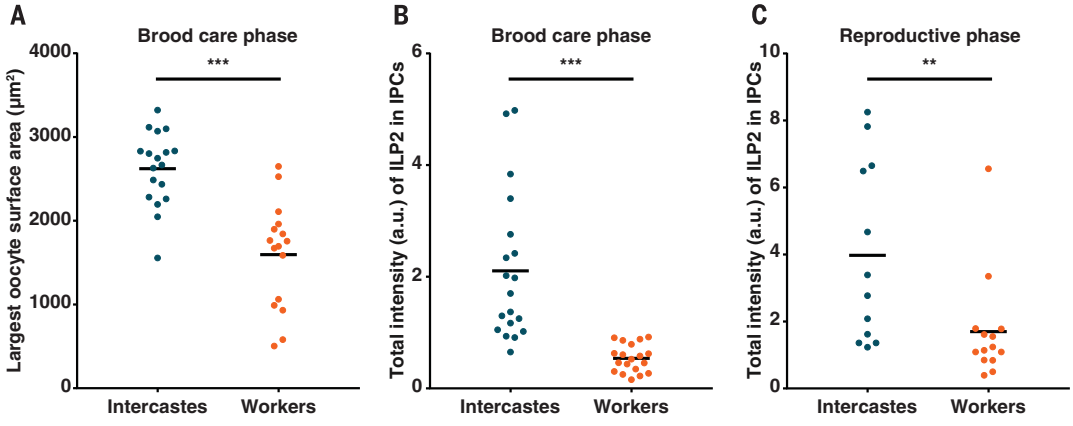


Fig. 4. Intercastes respond less to larvae and have more ILP2 than regular workers. (**A**) Intercastes have more active ovaries than age-matched regular workers in the brood care phase, despite both groups being in contact with larvae ($n \geq 16$, Welch's t test; $^{***}P < 0.0001$). (**B**) In the brood care phase ($n = 19$, Mann-Whitney U test; $^{***}P < 0.0001$) and (**C**) in the reproductive phase ($n \geq 12$, Mann-Whitney U test; $^{**}P = 0.0043$), intercastes have more ILP2 in their IPCs than age-matched regular workers. Horizontal bars indicate means on all dot plots.

adults to emerge from the pupa with low nutritional stores and low ILP2 levels. These subfertile individuals would have been more sensitive to larval signals that suppress reproduction and would consequently have foregone nest founding and ovary activation and instead assumed brood care roles. Other adults, meanwhile, would have emerged with high nutritional stores and high ILP2 levels. These adults would have had reduced sensitivity to larval signals and would have been more likely to reproduce despite the presence of larvae. This reproductive asymmetry could then have been enhanced or modified by natural selection to ultimately produce the obligately reproductive queens and sterile workers of advanced eusocial species (fig. S14). This scenario constitutes an explicit molecular version of Mary Jane West-Eberhard's model for the evolution of hymenopteran eusociality (*5*).

REFERENCES AND NOTES

1. E. O. Wilson, *The Insect Societies* (Belknap Press, 1971).
2. W. M. Wheeler, *Ants: Their Structure, Development and Behavior* (Columbia Univ. Press, 1910).
3. J. Hunt, *The Evolution of Social Wasps* (Oxford Univ. Press, 2007).
4. J. Field, *Behav. Ecol.* **16**, 770–778 (2005).
5. M. J. West-Eberhard, *in Animal Societies: Theories and Facts*, Y. Itô, J. L. Brown, J. Kikkawa, Eds. (Japan Scientific Societies Press, 1987), pp. 35–51.
6. P. S. Ward, *Annu. Rev. Ecol. Evol. Syst.* **45**, 23–43 (2014).
7. R. Libbrecht, P. R. Oxley, L. Keller, D. J. C. Kronauer, *Curr. Biol.* **26**, 391–395 (2016).
8. S. Patalano *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13970–13975 (2015).
9. Q. Li *et al.*, *Nat. Commun.* **5**, 4943 (2014).
10. J. Gospocic *et al.*, *Cell* **170**, 748–759.e12 (2017).
11. Materials and methods are available as supplementary materials.
12. M. Corona *et al.*, *PLOS Genet.* **9**, e1003730 (2013).
13. P. R. Oxley *et al.*, *Curr. Biol.* **24**, 451–458 (2014).
14. L. Badisco, P. Van Wielendaele, J. Vanden Broeck, *Front. Physiol.* **4**, 202 (2013).
15. M. Corona *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7128–7133 (2007).
16. K.-A. Nilsen *et al.*, *J. Exp. Biol.* **214**, 1488–1497 (2011).
17. A. L. Toth, G. E. Robinson, *Trends Genet.* **23**, 334–341 (2007).
18. Y. Wang, S. V. Azevedo, K. Hartfelder, G. V. Amdam, *J. Exp. Biol.* **216**, 4347–4357 (2013).
19. M. R. Brown *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5716–5721 (2008).
20. W. Trible, D. J. C. Kronauer, *J. Exp. Biol.* **220**, 53–62 (2017).
21. Y. Okada *et al.*, *J. Insect Physiol.* **56**, 288–295 (2010).
22. S. V. de Azevedo, K. Hartfelder, *J. Insect Physiol.* **54**, 1064–1071 (2008).
23. D. E. Wheeler, N. Buck, J. D. Evans, *Insect Mol. Biol.* **15**, 597–602 (2006).
24. S. A. Ament, M. Corona, H. S. Pollock, G. E. Robinson, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4226–4231 (2008).
25. S. Teseo, D. J. C. Kronauer, P. Jaisson, N. Châline, *Curr. Biol.* **23**, 328–332 (2013).
26. M. A. Riehle, Y. Fan, C. Cao, M. R. Brown, *Peptides* **27**, 2547–2560 (2006).
27. C. Géminard, E. J. Rulifson, P. Léopold, *Cell Metab.* **10**, 199–207 (2009).
28. R. Libbrecht, P. R. Oxley, D. J. C. Kronauer, bioRxiv 223255 [Preprint]. 22 November 2017. https://doi.org/10.1101/223255.
29. L. LaFever, D. Drummond-Barbosa, *Science* **309**, 1071–1073 (2005).
30. M. L. Borowiec *et al.*, bioRxiv 173393 [Preprint]. 8 August 2017. https://doi.org/10.1101/173393.
31. P. Oxley, V. Chandra, Social-evolution-and-behavior/insulin_signaling: Data and code: Social regulation of insulin signaling and the evolution of eusociality in ants, Version 1.1, Zenodo (2018); https://doi.org/10.5281/zenodo.1311222.

ACKNOWLEDGMENTS

We thank C. Zhao, H. Zebroski III, T. Tong, and the Rockefeller University Resource Centers for sequencing, peptide synthesis, and help with image analysis; M. Deyrup for assistance with ant collection; L. Olivos-Cisneros for dissecting brains; and

L. Vosshall, C. Bargmann, B. Chait, H. Yan, C. Desplan, and the entire Kronauer laboratory for helpful discussion. This is Clonal Raider Ant Project paper no. 9. **Funding:** This work was supported by grant 1DP2GM105454-01 from the NIH, a Searle Scholar Award, a Sinsheimer Scholar Award, a Hirschl/Weill-Caulier Trusts Award, a Klingenstein-Simons Award, a Pew Scholar Award, and an HHMI Faculty Scholar Award (D.J.C.K.); a Leon Levy Neuroscience Fellowship (P.R.O.); a Rockefeller University Women & Science Fellowship (I.F.-P.); and a Marie Curie International Outgoing Fellowship (PIOF-GA-2012-327992; R.L.). **Author contributions:** D.J.C.K., V.C., I.F.-P., and P.R.O. designed the study; S.K.M., I.F.-P., and V.C. performed fieldwork; P.R.O., V.C., R.L., S.K.M., I.F.-P., and A.L.R. performed genomic analyses; I.F.-P., V.C., A.L.R., R.L., and S.K.M. performed immunostains; V.C., A.L.R., I.F.-P., and P.R.O. performed pharmacological experiments; V.C., I.F.-P., and D.J.C.K. wrote the manuscript with feedback from all authors; and D.J.C.K. supervised the project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Raw sequence data are available through NCBI (BioProject PRJNA472392); scripts are available on GitHub (*31*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/398/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S14
Tables S1 to S6
References (*32*–*76*)

22 November 2017; resubmitted 1 April 2018
Accepted 30 May 2018
10.1126/science.aar5723

GUT INFLAMMATION

Lectins modulate the microbiota of social amoebae

Christopher Dinh[1,2], Timothy Farinholt[1], Shigenori Hirose[1], Olga Zhuchenko[1], Adam Kuspa[1,2]*

The social amoeba *Dictyostelium discoideum* maintains a microbiome during multicellular development; bacteria are carried in migrating slugs and as endosymbionts within amoebae and spores. Bacterial carriage and endosymbiosis are induced by the secreted lectin discoidin I that binds bacteria, protects them from extracellular killing, and alters their retention within amoebae. This altered handling of bacteria also occurs with bacteria coated by plant lectins and leads to DNA transfer from bacteria to amoebae. Thus, lectins alter the cellular response of *D. discoideum* to bacteria to establish the amoebae's microbiome. Mammalian cells can also maintain intracellular bacteria when presented with bacteria coated with lectins, so heterologous lectins may induce endosymbiosis in animals. Our results suggest that endogenous or environmental lectins may influence microbiome homeostasis across eukaryotic phylogeny.

Dictyostelium discoideum are amoebae that live in the soil and feed on bacteria; as food becomes scarce, they aggregate into a mound and form a migrating slug, which eventually develops into a fruiting body containing a spore-filled sorus on a cellular stalk (*1*). It had been thought that development occurred free of bacteria and that the final fruiting body was sterile (*2*). *D. discoideum* has an innate immune system composed of "sentinel" cells that are able to rid the slug of interstitial bacteria (*3*, *4*); however, it was recently reported that about one-third of wild isolates maintain symbiotic interactions with the bacteria in their soil environment and carry bacteria stably through cycles of growth and development (*5*). In a form of microbial farming, carried bacteria can seed a new food supply for germinating spores as the spores establish a new colony of amoebae (*5*). Additional interactions between *D. discoideum* and their bacterial associates have been documented, and there is evidence that they are controlled by specific signals (*6*–*8*). Our finding that amoebal lectins mediate bacterial carriage suggests that *D. discoideum* actively controls its microbiome and that amoeba-bacterium interactions may serve as a useful model for microbiome homeostasis.

Amoebae that carry bacteria are easily distinguished from noncarriers because migrating carrier slugs leave behind bacteria that form colonies visible to the naked eye (fig. S1A). To assay for carriage, we allowed amoebae to initiate development in the presence of food bacteria such as *Klebsiella pneumoniae* (fig. S1B) and scored individual fruiting bodies for the presence of viable bacteria (fig. S1C). By this measure, >90% of a carrier strain's sori contained bacteria (table S1) and bacterial endosymbionts in their spores (fig. S1D), whereas <10% of sori from noncarrier strains contained viable bacteria (table S1). It is unclear how bacterial carriage occurs, given the existence of an innate immune system, although significantly reduced sentinel cell populations in carrier strains have been reported (*9*). All of our carrier strains had sentinel cells (table S1), so we tested whether they cleared bacteria during extended slug migration. We found that carriers maintained their bacteria through >6 cm of slug migration, whereas no viable bacteria were recovered from noncarriers after slug migration, suggesting that sentinel cells do not influence bacterial carriage (table S1).

Because carrier strains maintain bacteria during development, we assessed their ability to kill bacteria as they transition from growth to development. The noncarrier laboratory strain AX4 killed extracellular bacteria within the plaque, whereas just outside the plaque, the bacteria were alive (Fig. 1A). Plaques of noncarrier strains were similar to those of AX4, whereas plaques of carriers had a mottled appearance, with live and dead bacteria segregated into distinct areas near the edges (Fig. 1B). Thus, extracellular bacterial killing within carrier plaques is uneven, and this might explain the presence of live bacteria as the amoebae begin to develop.

We tested extracellular killing of *K. pneumoniae* by amoebae in suspension and found that noncarrier strains killed bacteria efficiently, but carrier strains did not (Fig. 1C). We tested whether this difference in bacterial killing might be explained by the proteins that they secrete during growth by examining the secretomes harvested from amoebae consuming *K. pneumoniae*. We fractionated the secretomes by ion-exchange chromatography and followed antibacterial activity by using an end-point dilution assay that measures the killing of *K. pneumoniae* (fig. S2A). Antibacterial activity in noncarrier secretomes eluted from the column in a broad peak and was heat-labile (10 min, 65°C), but carriers produced very little, or undetectable, activity (Fig. 1D; fig. S2, B to D; and table S2). This prompted us to identify regulators of bacterial carriage by screening for proteins that are differentially secreted by carrier versus noncarrier amoebae. The most prominent differences between carriers and noncarriers were two proteins in carriers with apparent molecular weights of ~24 and ~30 kDa (fig. S3A). We used mass spectrometry to determine that the 30-kDa band contained a mixture of the well-studied discoidin I lectins discoidin A (DscA) and discoidin C (DscC) (table S3) and that the 24-kDa band contained a mixture of the calcium-dependent adhesion protein CadA and the related proteins Cad2 and Cad3 (table S4) (*10*–*13*).

Discoidin I forms dumbbell-shaped trimers with the subunits oriented in parallel, positioning the C-terminal H-type lectin domains at one end (*14*). Discoidin I binds to polysaccharides containing *N*-acetylgalactosamine (GalNAc) and galactosamine in β1–β3 linkages that are found in *D. discoideum* cell surface glycoproteins and bacterial carbohydrates (*15*, *16*). In laboratory strains, discoidin I is found inside amoebae, and it is secreted when they aggregate after the start of development (*17*, *18*). We surveyed the secretion of discoidin I in wild strains, using antibodies against DscA and DscC, and found that carriers appeared to secret higher levels of these proteins at the start of development than noncarriers (Fig. 2A). One carrier that we examined in detail secreted >80 times more discoidin I than noncarriers (fig. S3B). We examined the timing of discoidin I secretion from the end of vegetative growth through developmental aggregation and found that carriers secreted higher amounts of discoidin I at earlier times compared with noncarriers (e.g., Fig. 2B). These results suggest that discoidin I secretion before development is a physiological feature of carriers that may determine bacterial carriage. To explore this, we first measured bacterial binding and found that DscA binds quantitatively to *K. pneumoniae* bacteria, likely through the GalNAc moieties present in the lipopolysaccharide of many *K. pneumoniae* strains (Fig. 3A) (*14*, *19*). DscA displayed saturation binding to *K. pneumoniae* and *Escherichia coli* (Fig. 3B), and we estimated that DscA binds to *K. pneumoniae* with micromolar affinity (association constant $K_a = 1.2 \times 10^6\ M^{-1}$) through ~$1.4 \times 10^6$ binding sites on each bacterium (supplementary materials) (*20*). About 2.75×10^5 DscA proteins would be required to form a monolayer covering the surface of one *K. pneumoniae* bacterium, assuming that the lectin ends of the roughly cylindrical trimers were opposing an idealized smooth surface with the shape and volume of the bacterium (supplementary materials). Thus, additional binding sites within the glycocalyx of *K. pneumoniae* likely account for the fivefold higher DscA binding that we observed.

We tested whether discoidin I binding alters *K. pneumoniae*'s interaction with amoebae by examining bacterial killing, carriage, and endosymbiosis. First, we tested whether discoidin I protects bacteria from killing by *D. discoideum*

[1]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA. [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA.
*Corresponding author. Email: akuspa@bcm.edu

antibacterial proteins (Dabs). DscA-bound *K. pneumoniae* were protected from killing by Dabs in an overnight outgrowth assay (fig. S2F) and in a short-term viability assay (Fig. 3C). Notably, the protection in the outgrowth assay diminished when the amount of discoidin added was below that needed to achieve saturation binding to the bacteria (fig. S2F). However, DscA-coated *K. pneumoniae* were not resistant to killing by heat (65°C, 10 min) or antibiotic treatment (kanamycin at 50 μg/ml). If protection by discoidin I is a mechanism by which carrier strains spare bacteria from killing, carriers would have to secrete sufficient discoidin to protect some bacteria as the amoebae enter development. Thus, we reexamined carrier secretomes for cryptic Dabs that might have been masked by discoidin I. Removing discoidin I from carrier secretomes by passing them through polygalactose affinity columns revealed latent antibacterial activity (Fig. 1D, fig. S2E, and table S2). Our results indicate that discoidin I can counteract the Dabs secreted by carriers, at least in vitro, and this suggests that discoidin I is secreted by carrier strains at levels sufficient to spare some bacteria for carriage as amoebae enter development.

We tested whether discoidin I influences the bacterial carriage and endosymbiosis that define the carrier phenotype. We mixed *D. discoideum* amoebae with *K. pneumoniae* or with DscA-coated *K. pneumoniae* at the start of development and tested the resulting fruiting-body sori for live bacteria. A single treatment of *K. pneumoniae* with DscA resulted in bacterial carriage by the noncarrier strain QS4 and the noncarrier laboratory strain AX4 (table S1). We also observed live green fluorescent protein (GFP)–expressing *K. pneumoniae* within the *D. discoideum* spores, but only when the bacteria were pretreated with DscA (Fig. 4A). These results suggest that extracellular discoidin I is sufficient to account for bacterial carriage in *D. discoideum* by binding bacteria, protecting them from secreted antibacterial proteins, and inducing carriage and endosymbiosis.

Some of the internalized discoidin-coated bacteria escape intracellular killing and digestion by the phagolysosomal pathway of the amoebae and end up in the spores. Thus, discoidin I alters the handling of bacteria by amoebae so that live bacteria persist, a phenomenon that we have termed lectin-induced modified bacterial internalization, or LIMBI. We demonstrated the specificity of LIMBI for lectin-coated bacteria with mixing experiments using discoidin-coated bacteria and uncoated bacteria. We presented DscA-bound red fluorescent protein (RFP)–expressing *E. coli* and GFP-expressing *E. coli* together to *D. discoideum* and followed their uptake by the amoebae over several hours by

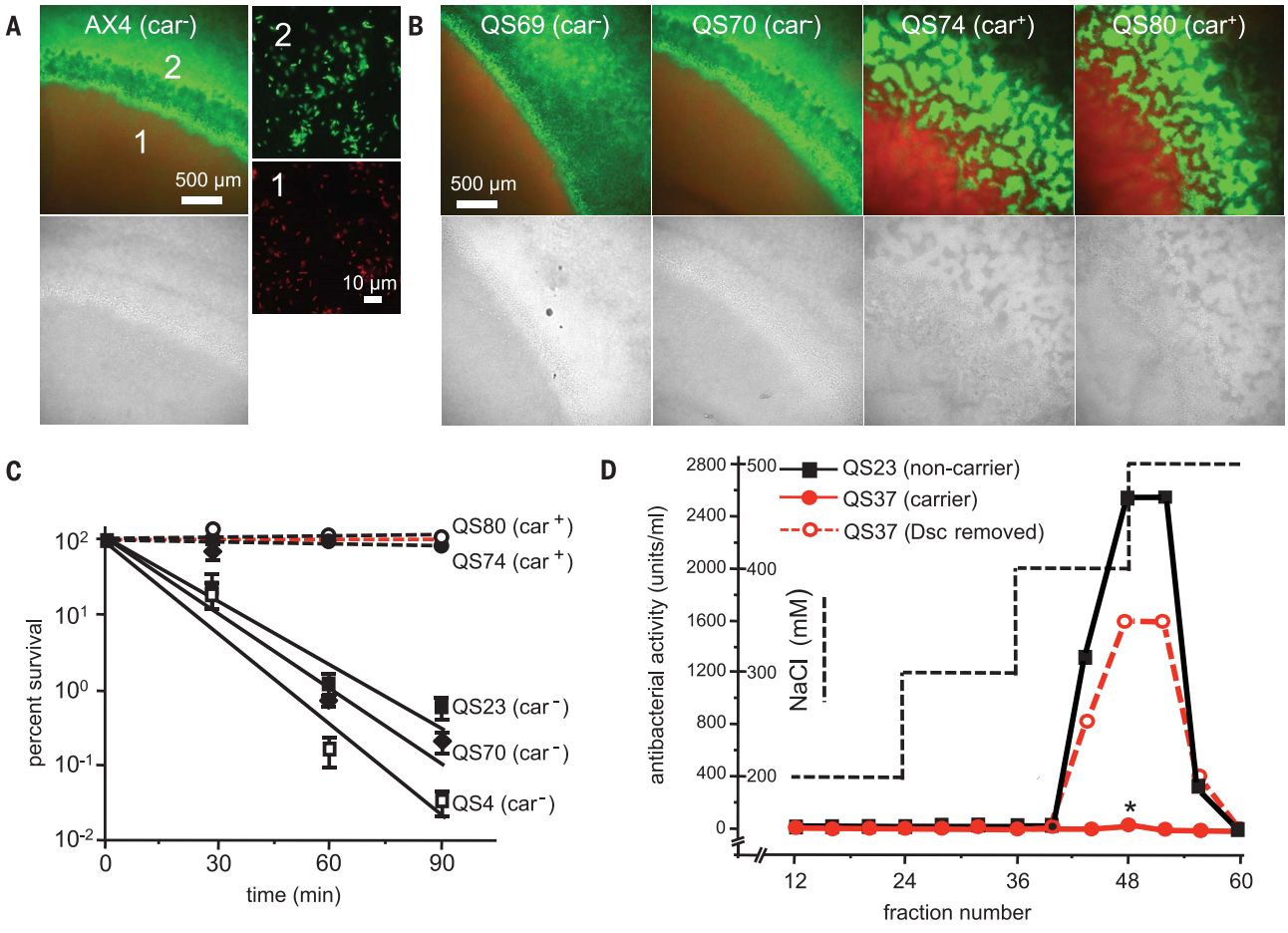


Fig. 1. Extracellular killing of bacteria by *D. discoideum* and secreted antibacterial proteins. (**A**) Plaques of *D. discoideum* strain AX4 growing on lawns of *K. pneumoniae* were stained with fluorescent dyes that distinguish live and dead bacteria, as visualized by fluorescence microscopy (upper left panel) and differential interference contrast (DIC) microscopy (lower left panel). Numbers (1 and 2) indicate the areas that were sampled to reveal dead (red) and live (green) bacteria (right panels). (**B**) Plaques of wild carrier (car$^+$) and noncarrier (car$^-$) strains, stained as in (A). (**C**) Carrier (circles) and noncarrier (squares and diamonds) amoebae were mixed with *K. pneumoniae* in buffer, and the number of viable bacteria were determined over time and normalized to 100% at the start of the assay. Values are means ± SEM for three independent experiments. (**D**) Secretomes from QS23 (noncarrier) and QS37 (carrier) were resolved by ion-exchange chromatography (DEAE-Sepharose) with a step-elution of increasing salt (NaCl, dashed line), and the fractions (1 ml) were assayed for antibacterial activity against *K. pneumoniae*. Trace activity was sometimes observed in carrier secretomes and elutes near the peak observed in noncarriers (asterisk). When discoidin I is removed from carrier secretomes, antibacterial activity becomes apparent (open circles).

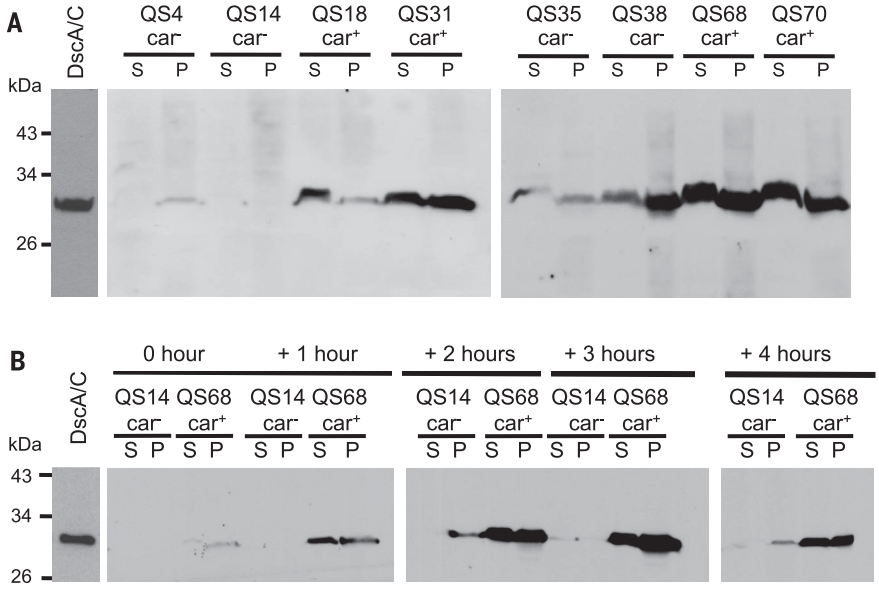


Fig. 2. Discoidin I secretion by wild carrier strains. Wild strains were harvested during late-stage growth on lawns of *K. pneumoniae* bacteria (when most of the bacteria had been consumed), shaken in Sor buffer for 1 hour before being separated into supernatant (S) and pellet (P) fractions, and resolved on SDS–polyacrylamide gel electrophoresis (SDS-PAGE) protein gels. (**A**) Western blots stained with antibodies against discoidin I, showing soluble and cell-associated discoidin I in carrier and noncarrier strains. The DscA/C control (left lane) was purified from AX4 (supplementary materials). (**B**) A representative time course of discoidin I production, from before bacteria are consumed to the start of the growth-to-development transition, for a noncarrier (QS14) and a carrier (QS68) strain. The 3-hour point is roughly equivalent to the time of harvest in (A).

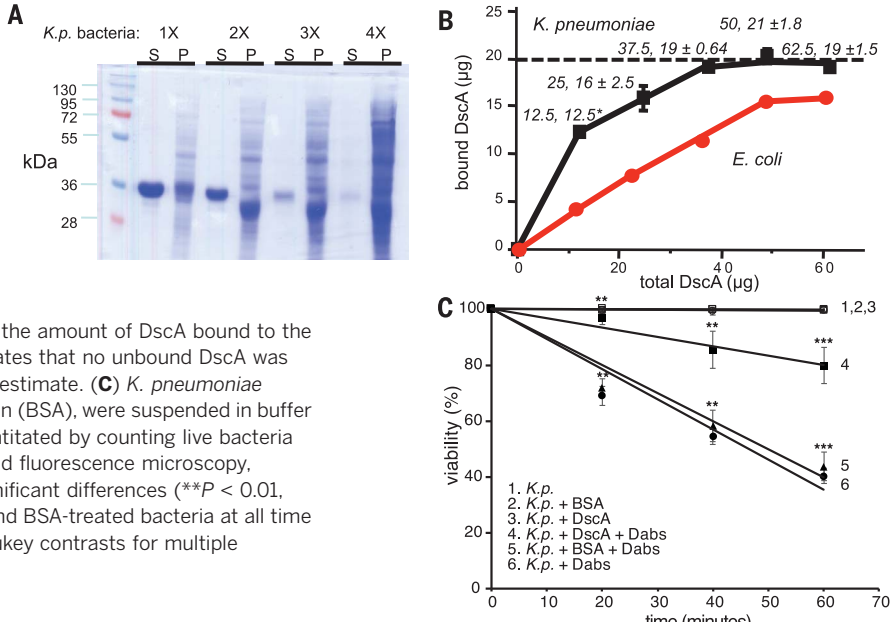


Fig. 3. Discoidin I binds and protects bacteria. (**A**) Increasing numbers of *K. pneumoniae* (*K.p.*) bacteria (1X = 2.3 × 10^7 bacteria) were incubated with purified DscA (300 µg/ml) in 100 µl buffer for 60 min at room temperature. Bacteria were separated from soluble DscA by centrifugation, and the supernatant (S) and pellet (P) fractions were resolved by SDS-PAGE and visualized by Coomassie blue staining. (**B**) DscA (12.5 to 62.5 µg) was incubated with 1.3 × 10^8 *K. pneumoniae* or *E. coli* in 100 µl for 60 min. Bacteria were removed by centrifugation, and the amount of unbound DscA was determined and used to calculate the amount of DscA bound to the bacteria (shown as means ± SEM). The asterisk indicates that no unbound DscA was detected. The dashed line shows the saturation point estimate. (**C**) *K. pneumoniae* bacteria, pretreated with DscA or bovine serum albumin (BSA), were suspended in buffer and challenged with Dabs. Bacterial viability was quantitated by counting live bacteria at various times, using a live-dead staining reagent and fluorescence microscopy, normalized to the zero-time samples. Statistically significant differences (**$P < 0.01$, ***$P < 0.001$) were observed between DscA-treated and BSA-treated bacteria at all time points (one-way pairwise analysis of variance using Tukey contrasts for multiple comparisons of means). Values are means ± SEM.

means of time-lapsed fluorescence imaging. Amoebae took up and digested uncoated bacteria, whereas some discoidin-coated bacteria remained intact and were scattered throughout the cytoplasm. The reciprocal experiment with discoidin-coated GFP-expressing *E. coli* resulted in retention of only those bacteria. The differential handling within the same amoeba of discoidin-coated and uncoated bacteria indicates that discoidin must be bound to the bacterium for it to persist after internalization. We quantified the persistence of bacteria after LIMBI by measuring the length of time that GFP-labeled bacteria remained intact within vegetative amoebae, using time-lapse imaging in an under-agar assay to create a pseudo–two-dimensional view (Fig. 4B) (*21*). By following bacteria from the time that they were internalized by amoebae to the time that they lost their structural integrity, we found that discoidin-coated bacteria persisted nearly seven times longer than uncoated bacteria (Fig. 4C).

LIMBI may provide an opportunity for bacterial DNA transfer into the host nucleus, and we examined this by inducing LIMBI of DscA-coated *E. coli* harboring a shuttle plasmid and selecting for drug-resistant amoebae. Each of the discoidin-coated *E. coli* strains introduced into *D. discoideum* by LIMBI led to genetic transformation of the amoebae at high frequencies (~2 to 20%), including those carrying plasmids with G418 (neo^r) or blasticidin (bs^r) resistance genes (table S5). To visualize LIMBI transformation, we used plasmids that express *D. discoideum* histone H2b fused to the RFP mCherry and observed that all of the transformed amoebae expressed the mCherry-tagged histone within their nuclei (Fig. 4D and table S5). Immunoblots of proteins from these cells confirmed that the fusion protein was expressed (fig. S4A). Somewhat unexpectedly, ~10% of the amoebae expressed RFP-H2b 24 hours after LIMBI without selection for drug resistance, suggesting that transient expression also occurs (table S5). These results indicate that DNA can be transferred from the bacteria into the nucleus of the amoebae, implying that LIMBI is functionally distinct from the phagolysosomal digestion pathway that amoebae use to process bacteria as food.

We tested the generality of LIMBI in eukaryotes by testing plant lectins as elicitors and mammalian cells as recipients. We identified several

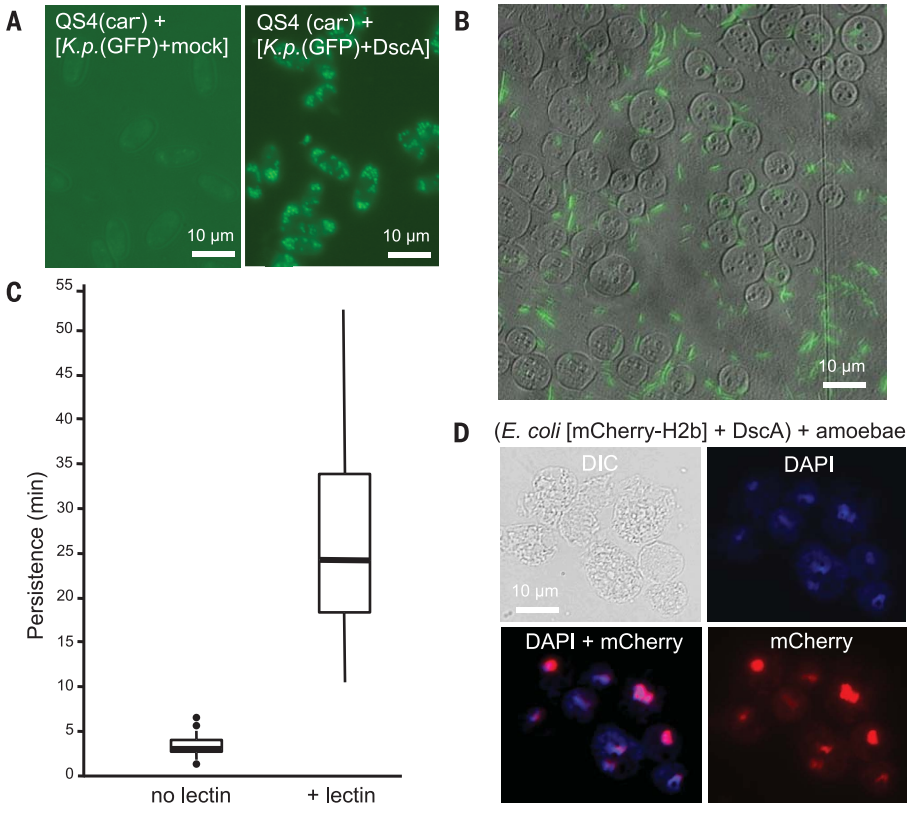


Fig. 4. Lectin-induced modified bacterial internalization (LIMBI) results in bacterial endosymbiosis, increased persistence of bacteria within amoebae, and genetic transformation. (**A**) *D. discoideum* amoebae (QS4) were mixed with GFP-expressing *K. pneumoniae* that had been pretreated with buffer (mock) or DscA and allowed to develop into fruiting bodies. Fluorescence microscopy of the resulting spores shows intact bacteria in spores after LIMBI with DscA. (**B**) An overlay image of DIC and fluorescence microscopy, showing vegetative amoebae and GFP-expressing and DscA-coated *E. coli* under agar. The persistence of the bacteria within amoebae was quantified by determining the number of minutes that each bacterium was observed to remain intact. (**C**) Box plots of the persistence time of discoidin-coated bacteria within amoebae after LIMBI. The edges of the boxes represent the 75th and 25th percentiles, thick lines represent medians, the whiskers are the minimums and maximums, and dots are outliers. The difference between LIMBI and the no-lectin control was statistically significant ($P < 0.001$, Mann-Whitney-Wilcoxon rank-sum test). (**D**) LIMBI-mediated gene transfer from *E. coli* into *D. discoideum*. Discoidin-coated *E. coli* harboring a plasmid engineered to express a red fluorescent protein fusion to *D. discoideum* histone H2b (mCherry-H2b) were mixed with AX4 amoebae, followed by 10 days of drug selection for the plasmid. Shown are a DIC microscopy image, fluorescence microscopy images of nuclear DNA stained with DAPI (4′,6-diamidino-2-phenylindole) or mCherry, and an overlay showing nuclear expression of mCherry-H2b.

plant-derived lectins that bound to *E. coli*, including *Wisteria floribunda* agglutinin, *Dolichos biflorus* agglutinin, soybean agglutinin (SBA), and concanavalin A (ConA) (*22*). We mixed lectin-coated bacteria with *D. discoideum*, using the same protocols developed with discoidin I, and found that each of the plant lectins appeared to be equally effective for LIMBI and LIMBI-mediated transformation (fig. S4, B and C, and table S6). We also introduced lectin-coated *E. coli* to mammalian cells in culture and found that DscA, ConA, and SBA were each equally effective for LIMBI and LIMBI-mediated transformation of mouse RAW264 macrophages (tables S5 and S6). Competition experiments with admixtures of GFP- and RFP-expressing bacteria showed that only lectin-coated bacteria persisted intact within RAW264 macrophages after overnight incubation (fig. S5). Our results show that LIMBI can occur in amoebae and in animal cells, using amoebal or plant lectins, suggesting that modified uptake and persistence of lectin-coated bacteria in eukaryotic cells is a general phenomena.

We have described physiological differences between *D. discoideum* carrier and noncarrier strains that appear to account for the salient features of bacterial carriage during social amoebae development. Because discoidin I is sufficient to induce carriage by noncarrier strains, its secretion by carrier strains provides a plausible mechanism for the transition of carrier amoebae from active bacterial killing and feeding to starvation and bacterial symbiosis. Prestarvation factor (PSF) is a ~70-kDa secreted protein that reports on the bacterial food supply at the growth-to-development transition (*23*, *24*). PSF is sequestered by bacteria when bacterium/amoeba ratios are high, but as the bacteria are depleted by amoebal feeding, the level of free PSF rises and stimulates *D. discoideum* development through autocrine signaling. PSF has never been identified, and because discoidin I binds bacteria and affects the kill-versus-carriage dynamics of amoebae, we are testing whether discoidin I is PSF. Discoidin I's H-type lectin domains are on the C termini of the trimer subunits, but the N-terminal discoidin domains have distinctive structural similarities to F-type lectins and have been proposed to bind fucose and sialic acid containing *N*- or *O*-glycans (*14*). If this is true, it is possible that the H-type lectin domains of discoidin I bind bacteria, leaving the discoidin domains free to interact with N- or O-linked glycans on the cell surface of *D. discoideum* amoebae to initiate LIMBI or engage in receptor-mediated signaling. If the uptake mechanism does involve amoebal glycoproteins as receptors, plant lectins might act as bacteria-bound ligands for those same receptors.

Lectinophagocytosis results in efficient bacterial uptake by eukaryotic cells through specific receptor recognition of lectins provided by the bacteria or the host cell (*25*, *26*). Plant lectins can also mediate macrophage lectinophagocytosis (*27*, *28*), and in some cases this is concomitant with the suppression of bacterial killing (*29*). It is well known that lectins target bacteria for destruction by the innate immune system of the host, but there are reports of lectins with functions unrelated to defense (*30*, *31*). If lectin protection of bacteria within hosts proves to be a general feature of multicellular eukaryotes, it would provide a new perspective on the regulation of microbiomes in well-studied systems, including humans. For example, LIMBI may promote the maintenance of intracellular bacteria in the mucosal epithelium or myofibroblasts of the colon and contribute to inflammatory disease processes.

REFERENCES AND NOTES

1. R. H. Kessin, *Dictyostelium: Evolution, Cell Biology, and the Development of Multicellularity* (Cambridge Univ. Press, 2001).
2. K. B. Raper, *J. Agric. Res.* **55**, 289–316 (1937).
3. G. Chen, O. Zhuchenko, A. Kuspa, *Science* **317**, 678–681 (2007).
4. X. Zhang, O. Zhuchenko, A. Kuspa, T. Soldati, *Nat. Commun.* **7**, 10938 (2016).
5. D. A. Brock, T. E. Douglas, D. C. Queller, J. E. Strassmann, *Nature* **469**, 393–396 (2011).
6. P. Stallforth *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14528–14533 (2013).
7. D. A. Brock, S. Read, A. Bozhchenko, D. C. Queller, J. E. Strassmann, *Nat. Commun.* **4**, 2385 (2013).
8. S. DiSalvo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5029–E5037 (2015).
9. D. A. Brock, W. E. Callison, J. E. Strassmann, D. C. Queller, *Proc. Biol. Sci.* **283**, 20152727 (2016).
10. S. H. Barondes, D. N. W. Cooper, W. R. Springer, in *Methods in Cell Biology*, J. A. Spudich, Ed. (Academic Press, 1987), vol. 28, pp. 387–409.
11. E. F. S. Wong, S. K. Brar, H. Sesaki, C. Yang, C. H. Siu, *J. Biol. Chem.* **271**, 16399–16408 (1996).

12. S. Basu, P. Fey, D. Jimenez-Morales, R. J. Dodson, R. L. Chisholm, *Genesis* **53**, 523–534 (2015).
13. L. Eichinger *et al*., *Nature* **435**, 43–57 (2005).
14. S. V. Mathieu, K. S. Aragão, A. Imberty, A. Varrot, *J. Mol. Biol.* **400**, 540–554 (2010).
15. D. N. Cooper, S. C. Lee, S. H. Barondes, *J. Biol. Chem.* **258**, 8745–8750 (1983).
16. D. N. W. Cooper, P. L. Haywood-Reid, W. R. Springer, S. H. Barondes, *Dev. Biol.* **114**, 416–425 (1986).
17. D. L. Simpson, S. D. Rosen, S. H. Barondes, *Biochemistry* **13**, 3487–3493 (1974).
18. W. A. Frazier, S. D. Rosen, R. W. Reitherman, S. H. Barondes, *J. Biol. Chem.* **250**, 7714–7721 (1975).
19. E. Vinogradov *et al*., *J. Biol. Chem.* **277**, 25070–25081 (2002).
20. R. W. Reitherman, S. D. Rosen, W. A. Frasier, S. H. Barondes, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 3541–3545 (1975).
21. S. Hirose, B. Santhanam, M. Katoh-Kurosawa, G. Shaulsky, A. Kuspa, *Development* **142**, 3561–3570 (2015).
22. I. Lagarda-Diaz, A. M. Guzman-Partida, L. Vazquez-Moreno, *Int. J. Mol. Sci.* **18**, 1242 (2017).
23. M. Clarke, S. C. Kayman, K. Riley, *Differentiation* **34**, 79–87 (1987).
24. M. Clarke, J. Yang, S. C. Kayman, *Dev. Genet.* **9**, 315–326 (1988).
25. I. Ofek, N. Sharon, *Infect. Immun.* **56**, 539–547 (1988).
26. A. Athamna *et al*., *Infect. Immun.* **59**, 1673–1682 (1991).
27. Z. Bar-Shavit, R. Goldman, *Exp. Cell Res.* **99**, 221–236 (1976).
28. R. Gallily, B. Vray, I. Stain, N. Sharon, *Immunology* **52**, 679–686 (1984).
29. R. Gallily, I. Stain, O. Zaady, *Immunol. Lett.* **13**, 151–158 (1986).
30. A. A. Gust, R. Willmann, Y. Desaki, H. M. Grabherr, T. Nürnberger, *Trends Plant Sci.* **17**, 495–502 (2012).
31. X. Pang *et al*., *Nat. Microbiol.* **1**, 16023 (2016).

ACKNOWLEDGMENTS

The authors thank W. F. Loomis, G. Shaulsky, and E. Ostrowski for providing insights during the course of this work. **Funding:** Dictyostelium Functional Genomics Program Project Grant from the National Institutes of Health (PO1 HD39691). **Author contributions:** C.D. and A.K. conceived and designed the experiments, C.D. directed and performed all of the experiments not done by others, T.F. and S.H. confirmed LIMBI and LIMBI-mediated transformation independently, T.F. performed the phagocytosis and short-term bacterial-killing assays, S.H. made the mCherry-H2b plasmid and performed the under-agar LIMBI imaging, and O.Z. quantified sentinel cells, performed the carriage assays of migrating slugs, and performed the amoebal bacterial-killing assays. A.K. wrote the manuscript with scientific and editorial input from C.D., O.Z., T.F., and S.H. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/402/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S6
References (*32*–*46*)

4 February 2018; resubmitted 20 April 2018
Accepted 21 June 2018
10.1126/science.aat2058

GUT INFLAMMATION

LKB1 deficiency in T cells promotes the development of gastrointestinal polyposis

M. C. Poffenberger[1,2], A. Metcalfe-Roach[1], E. Aguilar[1,2], J. Chen[1,2], B. E. Hsu[1,3], A. H. Wong[1,2], R. M. Johnson[1,4], B. Flynn[1,2], B. Samborska[1], E. H. Ma[1,2], S.-P. Gravel[1,5], L. Tonelli[1], L. Devorkin[6], P. Kim[6], A. Hall[1,7], S. Izreig[1,2], E. Loginicheva[8], N. Beauchemin[1,9], P. M. Siegel[1,3], M. N. Artyomov[8,10], J. J. Lum[6,11], G. Zogopoulos[1,7], J. Blagih[1,2], R. G. Jones[1,2,12]*

Germline mutations in *STK11*, which encodes the tumor suppressor liver kinase B1 (LKB1), promote Peutz–Jeghers syndrome (PJS), a cancer predisposition syndrome characterized by the development of gastrointestinal (GI) polyps. Here, we report that heterozygous deletion of *Stk11* in T cells (LT^{het} mice) is sufficient to promote GI polyposis. Polyps from LT^{het} mice, *Stk11*$^{+/-}$ mice, and human PJS patients display hallmarks of chronic inflammation, marked by inflammatory immune-cell infiltration, signal transducer and activator of transcription 3 (STAT3) activation, and increased expression of inflammatory factors associated with cancer progression [interleukin 6 (IL-6), IL-11, and CXCL2]. Targeting either T cells, IL-6, or STAT3 signaling reduced polyp growth in *Stk11*$^{+/-}$ animals. Our results identify LKB1-mediated inflammation as a tissue-extrinsic regulator of intestinal polyposis in PJS, suggesting possible therapeutic approaches by targeting deregulated inflammation in this disease.

Liver kinase B1 (LKB1) is a multifaceted serine/threonine kinase with roles in metabolism, cell polarity, cell size control, and proliferation (*1*). Somatic mutations in *STK11* are associated with several human cancers, including lung cancer (where it is frequently comutated with *KRAS*) and gynecological cancers (ovary, uterus, and cervix) (*2*). Heterozygous germline mutations in *STK11* predispose both humans and mice to the development of Peutz–Jeghers syndrome (PJS) (*3*–*7*), an autosomal dominant hereditary disease characterized by the development of benign gastrointestinal (GI) hamartomatous polyps (*8*). PJS patients also have a high cumulative cancer risk, with >90% chance of developing tumors by age 65 (*9*).

Much of our understanding of PJS development and pathogenesis has been derived from *Stk11*$^{+/-}$ mice, which develop GI polyps histologically similar to those found in PJS patients, with an average latency of 9 to 12 months (*5*–*7*). Deletion of *Stk11* in smooth-muscle cells promotes GI polyp formation with lower penetrance than that in whole-body *Stk11*$^{+/-}$ mice (~60% versus 85 to 100%) (*10*). We used a *Villin*-Cre transgene to delete *Stk11* in intestinal epithelial cells (IECs) of the small intestine and colon. We found no evidence of GI polyps in mice heterozygous (*Stk11*$^{fl/+}$;*Villin*-Cre) for *Stk11* in IECs, even those >52 weeks of age (table S1), which argues for additional nonepithelial functions for LKB1 that contribute to disease development in *Stk11*$^{+/-}$ mice and PJS patients.

LKB1 plays an essential role in regulating hematopoietic stem-cell metabolism and survival (*11*–*13*). Unexpectedly, we found that aged mice (>52 weeks) with heterozygous deletion of *Stk11* throughout the hematopoietic system (LH^{het}: *Stk11*$^{fl/+}$;*Vav*-iCre) developed GI polyps at the junction of the small intestine and stomach reminiscent of PJS polyps (Fig. 1A). Polyps observed in LH^{het} mice displayed an arborizing smooth-muscle core and were histologically indistinguishable from polyps observed in *Stk11*$^{+/-}$ mice (Fig. 1A). We next generated mice harboring heterozygous or homozygous *Stk11* mutations in either the T cell (LT^{het}: *Stk11*$^{fl/+}$;*Lck*-Cre and LT^{ko}: *Stk11*$^{fl/fl}$;*Lck*-Cre) or B cell (LB^{ko}: *Stk11*$^{fl/fl}$;*Cd19*-Cre) compartments. LT^{het} mice developed GI polyps at an incidence similar to that of LH^{het} mice, and homozygous deletion of *Stk11* in T cells (LT^{ko}) doubled the incidence of polyp formation (Fig. 1, B and C, and table S2). LB^{ko} mice did not develop disease (table S1). A subset of male LT^{het} and LT^{ko} mice also displayed liver lesions and increased rates of hepatocellular carcinoma (HCC) relative to those of control animals (table S3), a phenotype often observed in *Stk11*$^{+/-}$ mice (*14*). Genotyping and immunoblot analysis confirmed reduced LKB1 expression in T cells but not liver tissue of LT^{het} mice (fig. S1). Thus, T cell–specific loss of *Stk11* is sufficient to promote intestinal polyposis.

Many of LKB1's functions have been attributed to its stimulation of adenosine 5′-monophosphate (AMP)–dependent protein kinase (AMPK)–dependent stress responses (*2*). However, mice with T cell–specific (AT^{ko}: *Prkaa1*$^{fl/fl}$;*Cd4*-Cre) or hematopoietic cell–specific (AH^{ko}: *Prkaa1*$^{fl/fl}$;*Vav*-iCre) loss of AMPKα1 did not develop GI polyps (table S1). GI polyps from LT^{het} and LT^{ko} mice displayed increased mammalian target of rapamycin complex 1 (mTORC1) signaling (fig. S2A), which is characteristic of polyps from *Stk11*$^{+/-}$ mice (*15*). However, mice with conditional deletion of both LKB1 and mTOR in T cells (LT^{ko};T-Frap$^{-/-}$: *Stk11*$^{fl/fl}$;*Frap*$^{fl/fl}$;*Lck*-Cre^{+}) still developed GI polyps (fig. S2B). Thus, AMPK or mTOR pathway activity in T cells does not appear to contribute to PJS polyp development in these mouse models.

Histological analysis of PJS polyps revealed evidence of prominent immune cell infiltration in polyps from whole-body *Stk11*$^{+/-}$, LT^{het}, and LT^{ko} mice (Fig. 2A and fig. S3A). T cells (CD4^{+} and CD8^{+}), macrophages, and neutrophils were diffusely present throughout the polyps, with lymphocytes also organizing into concentrated centers in the smooth-muscle core of the polyps (Fig. 2A and fig. S3, A to C). Histological analysis of polyps from two independent, unrelated PJS patients revealed clustering of T cells in the smooth-muscle core and stroma of the polyps (Fig. 2B), similar to our observations in mouse samples. To assess the impact of lymphocytes on disease progression, we crossed *Stk11*$^{+/-}$ mice onto a *Rag2*-deficient background (which lack mature lymphocytes because of a block in development). Although *Stk11*$^{+/-}$;*Rag2*$^{-/-}$ mice still developed polyps, polyp size was significantly decreased in *Stk11*$^{+/-}$ mice lacking lymphocytes (Fig. 2, C and D).

We next assessed inflammatory cytokine expression in polyps and adjacent nonpolyp tissue from aged control and LT^{het} animals (>52 weeks of age) (Fig. 2E). Polyps from diseased LT^{het} mice consistently expressed increased levels of inflammatory cytokines—including *Il1b*, *Il6*, *Il11*, and *Tnfa*—compared with GI sections from control animals or LT^{het} mice that did not develop disease (Fig. 2E). Notable in this profile were high levels of *Il6* and *Il11*, cytokines that are associated with chronic gastric inflammation and GI tumor development (*16*–*18*). Similar changes in inflammatory cytokine gene expression were observed in polyps from *Stk11*$^{+/-}$ mice (Fig. 2F). The

[1]Goodman Cancer Research Centre, McGill University, Montreal, Quebec H3A 1A3, Canada. [2]Department of Physiology, McGill University, Montreal, Quebec H3G 1Y6, Canada. [3]Department of Medicine, McGill University, Montreal, Quebec H3G 2M1, Canada. [4]Genentech, 1 DNA Way South, San Francisco, CA 94080, USA. [5]Faculty of Pharmacy, University of Montreal, Montreal, Quebec H3C 3J7, Canada. [6]Trev and Joyce Deeley Research Centre, BC Cancer Agency, Victoria, British Columbia V8R 6V5, Canada. [7]Research Institute of the McGill University Health Centre, Montreal, Quebec H3H 2R9, Canada. [8]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA. [9]Department of Biochemistry, McGill University, Montreal, Quebec H3G 1Y6, Canada. [10]Center for Human Immunology and Immunotherapy Programs, Washington University at St. Louis, St. Louis, MO 63110, USA. [11]Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada. [12]Center for Cancer and Cell Biology, Van Andel Research Institute, Grand Rapids, MI 49503, USA.
*Corresponding author. Email: russell.jones@mcgill.ca, russell.jones@vai.org

expression of *Cxcl2* (MIP-2), a proinflammatory chemokine that directs immune-cell recruitment, was increased in both polyps and adjacent nonpolyp tissue from LT^{het} and $Stk11^{+/-}$ mice (Fig. 2, G and H). Profiling immune-cell subsets from GI tissue of control or polyp-bearing LT^{het} mice by means of flow cytometry revealed the increased presence of inflammatory monocytes ($CD11b^{+}Gr1^{mid}$) and neutrophils ($CD11b^{+}Gr1^{hi}$) in polyps of mice heterozygous for LKB1 in T cells (Fig. 2I).

To test whether inflammation could influence PJS polyp formation, we subjected young control and $Stk11^{+/-}$ mice to sublethal irradiation [4.5 grays (Gy)], which stimulates acute mucositis and inflammatory cytokine production in the GI tract (*19*). All $Stk11^{+/-}$ mice that received radiation developed polyps within 16 weeks of treatment, whereas age-matched nonirradiated $Stk11^{+/-}$ mice did not develop visible polyps during this time (Fig. 2, J and K). Thus, immune infiltration and inflammatory cytokines are hallmarks of PJS polyps, and damage-associated inflammation can initiate polyp formation in genetically susceptible $Stk11^{+/-}$ mice.

We next investigated how LKB1 loss in T cells could account for the proinflammatory nature of PJS polyps. RNA sequencing (RNA-seq) analysis of in vitro activated control and LT^{ko} $CD8^{+}$ T cells revealed enrichment in several pathways central to T cell activation (cytokines/chemokines, ribosome biogenesis, and phosphatidylinositol 3-kinase signaling) in LT^{ko} T cells (fig. S4). Among the pathways specifically enriched in activated LT^{ko} $CD8^{+}$ T cells were Janus kinase (JAK)–signal transducers and activators of transcription (STAT) signaling and cytokine–cytokine receptor interactions (figs. S5 and S6), the former scoring highly owing to increased cytokine mRNA expression in activated LT^{ko} T cells. The expression of several of these cytokines (*Ifng*, *Il9*, and *Il24*) in activated LT^{het} and LT^{ko} $CD8^{+}$ and $CD4^{+}$ T cells was confirmed with quantitative polymerase chain reaction (fig. S7, A and B). *Stk11*-deficient $CD4^{+}$ and $CD8^{+}$ T cells activated under nonpolarizing conditions displayed increased secretion of many of the inflammatory factors observed in PJS polyps, including CXCL2, CXCL9, and interleukin 6 (IL-6) (Fig. 3, A and B). The activation of naïve LT^{het} or LT^{ko} $CD4^{+}$ T cells led to the increased generation of interferon-γ (IFN-γ)– and IL-17A–producing $CD4^{+}$ T cells under T helper 1 (T_H1)– and T_H17-polarizing conditions, respectively (Fig. 3, C to E). LKB1-deficient T_H17 cells also displayed increased IL-6 and tumor necrosis factor–α (TNF-α) production (Fig. 3, E and F).

LKB1 deletion in T cells leads to deregulated metabolism and increased IFN-γ production by $CD4^{+}$ and $CD8^{+}$ T cells (*20*). We observed progressive deregulation of T cell homeostasis based on *Stk11* gene status, marked by an increased frequency of activated $CD4^{+}$ and $CD8^{+}$ T cells ($CD44^{hi}$) in LT^{het} and LT^{ko} mice (fig. S8, A and B). Increased proportions of activated IFN-γ– and IL-17A–producing T cells, as well as IL-6–producing $CD4^{+}$ T cells, were detected in the GI-draining mesenteric lymph nodes (mLN) of LT^{ko} mice (Fig. 3G and fig. S8C). Increased frequency of polyfunctional ($IFN\text{-}\gamma^{+}TNF\text{-}\alpha^{+}$) inflammatory T cells was observed in polyps isolated from LT^{het} and LT^{ko} animals (Fig. 3H). The expansion of inflammatory $CD4^{+}$ T_H cells in LT^{ko} mice was likely not due to defects in regulatory T cell (T_{reg} cell) populations as control, LT^{het}, and LT^{ko} mice displayed similar levels of $CD4^{+}Foxp3^{+}$ T_{reg} cells (fig. S9, A and B), and we observed no difference in suppressive capacity between control and LKB1-deficient $CD4^{+}$ T_{reg} cells (fig. S9, C and D).

We next tested whether *Stk11*-mutant T cells could modulate inflammation in surrounding tissues. Supernatants from *Stk11* mutant T cells stimulated a fourfold increase in *Il6* expression in nontransformed mouse embryonic fibroblasts (MEFs) compared with supernatants from wild-type T cells (Fig. 3I). $Stk11^{-/-}$ MEFs secreted IL-6 in the absence of inflammatory stimuli and produced significantly more CXCL2 and IL-6 when stimulated with lipopolysaccharide (LPS) (Fig. 3J and fig. S10A). Deregulated IL-6 production by $Stk1^{-/-}$ MEFs was not limited to LPS treatment because other Toll-like receptor (TLR) ligands strongly induced IL-6 production in LKB1-deficient MEFs (Fig. 3J). IL-6 production by LKB1-deficient MEFs was insensitive to mTORC1 inhibition by use of rapamycin (fig. S10B). Thus, in addition to aberrant immune-cell function, *Stk11* mutations in nonimmune tissues (epithelial and/or stromal cells) in PJS patients may confer increased responsiveness to inflammatory stimuli.

In inflammation-driven cancers such as gastric and colon cancer, tumor progression is driven by chronic STAT3 activation downstream of inflammatory cytokines such as IL-6 and IL-11 (*21*). Consistent with the abundance of STAT3-activating cytokines in the GI microenvironment, polyps from LT^{het}, LT^{ko}, and $Stk11^{+/-}$ mice displayed elevated phospho-STAT3 (Y705) staining compared with similar sections from LT^{wt} and $Stk11^{+/+}$ mice (Fig. 4A). A similar distribution of phospho-STAT3 (Y705) staining was observed in the stromal compartment of polyps from PJS patients (Fig. 4B). Both polyps and adjacent nonpolyp tissue from diseased mice exhibited increased STAT3 phosphorylation compared with that of nondiseased tissue from control mice (Fig. 4, C and D). Expression of several STAT3 target genes, including *Ccnd1* and *Socs3*, was increased in polyps (Fig. 4E and fig. S11A). Additionally, the STAT3 target gene *Tlr2*, which encodes a pattern-recognition receptor potentiating inflammatory responses to bacterial peptidoglycans, was significantly increased in the polyps of LT^{het} and $Stk11^{+/-}$ mice (Fig. 4F). Increased TLR2 expression and polymorphisms in the *TLR2* gene are associated with gastric cancer (*22*, *23*). Conditioned medium from activated LKB1-mutant (LT^{het} and LT^{ko}) T cells was

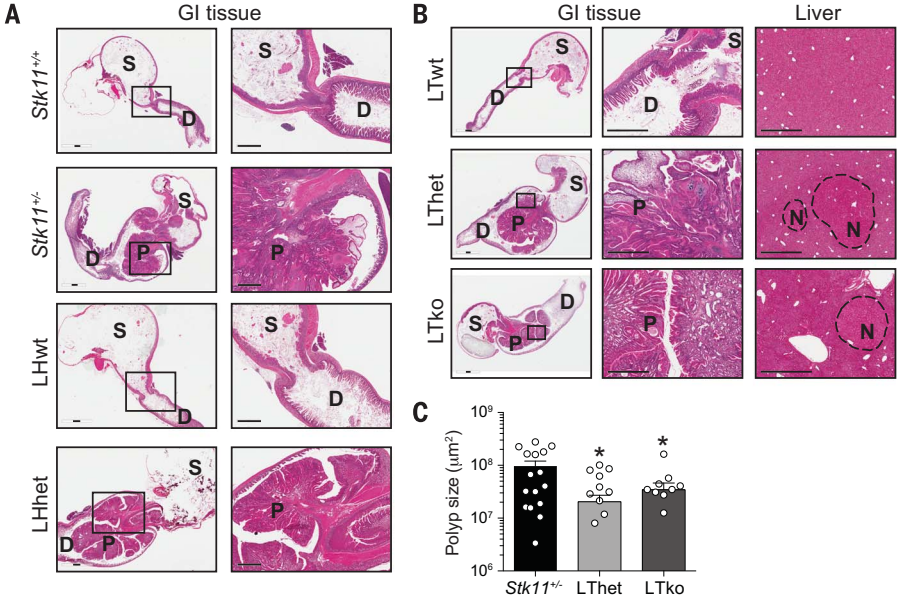


Fig. 1. T cell–specific deletion of LKB1 promotes intestinal polyposis. (**A**) Representative hematoxylin and eosin (H&E)–stained sections of stomachs (S), small intestines (D), and polyps (P) from $Stk11^{+/+}$ (n = 84), $Stk11^{+/-}$ (n = 64), LH^{wt} (n = 28), and LH^{het} (n = 38) mice aged >52 weeks. Images in the right column are a higher magnification of the boxed region indicated in the left column. Scale bars, 1 mm. (**B**) Representative H&E-stained sections of stomach (S), small intestine (D), polyps (P), liver, and liver nodules (N) from LT^{wt} (n = 102), LT^{het} (n = 64), and LT^{ko} (n = 50) mice aged >52 weeks. Images in the center column are a higher magnification of the boxed region in the left column. Scale bars, 1 mm. (**C**) Polyp size in $Stk11^{+/-}$ (n = 16), LT^{het} (n = 10), and LT^{ko} (n = 9) mouse strains aged >52 weeks. Data are represented as the mean ± SEM. *P < 0.05 compared with control samples ($Stk11^{+/-}$ mice), one-way analysis of variance (ANOVA), Dunnett's multiple comparisons test.

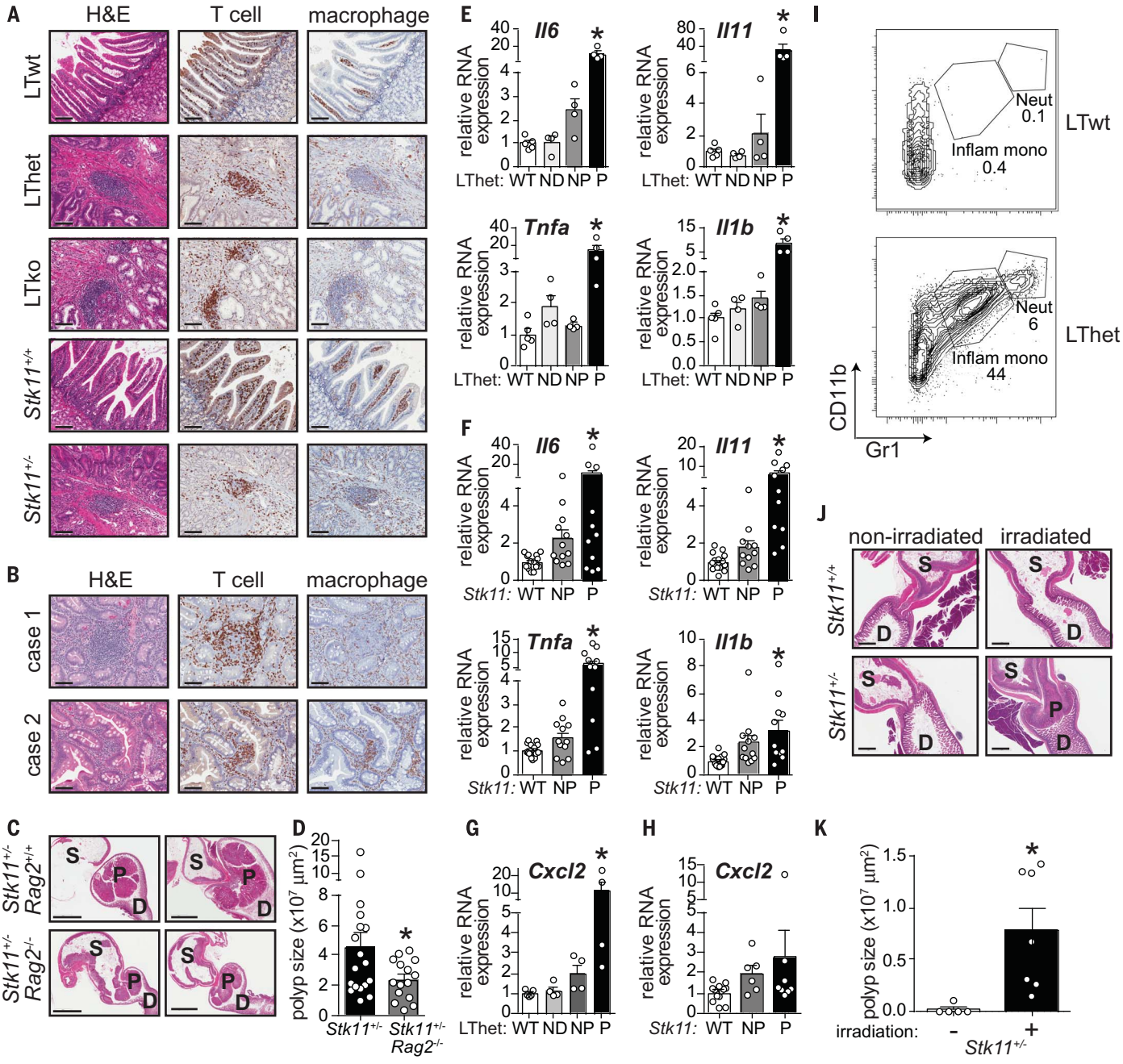


Fig. 2. Immune-cell infiltration and expression of proinflammatory genes are hallmarks of PJS polyps. (**A** and **B**) Representative H&E-, CD3- (T cell) and Iba-1– (macrophage) stained sections of stomach, small intestine, and polyps from (A) mice aged >52 weeks (n = 3 for each genotype) and (B) PJS patients (n = 2). Scale bars, 100 µm. (**C**) Representative H&E staining of GI sections from *Stk11*$^{+/-}$ (n = 18) or *Stk11*$^{+/-}$;*Rag2*$^{-/-}$ (n = 15) mice aged >52 weeks. Scale bars, 5 mm. (**D**) Polyp size for animals in (C). Data are represented as the mean ± SEM. *P < 0.05 compared with *Stk11*$^{+/-}$ mice, unpaired Student's t test. (**E**) Normalized relative mRNA levels of inflammatory cytokines (*Il1b*, *Il6*, *Il11*, and *Tnfa*) in GI tissue from LT^{wt} (WT) (n = 5) or nondiseased (ND) LT^{het} mice (n = 4), or polyps (P) (n = 4) and adjacent nonpolyp tissue (NP) (n = 4) from diseased LT^{het} mice aged >52 weeks. Gene expression relative to *Tbp* mRNA levels. Data are represented as the mean ± SEM. *P < 0.05 compared with LT^{wt} mice, one-way ANOVA, Dunnett's multiple comparisons test. (**F**) Normalized relative cytokine mRNA levels in GI tissue from *Stk11*$^{+/+}$ (WT) (n = 16) or polyps (P) (n = 11) and adjacent nonpolyp tissue (NP) (n = 12) from *Stk11*$^{+/-}$ mice aged >52 weeks. Gene expression relative to *Tbp* mRNA levels. Data are represented as the mean ± SEM. *P < 0.05 compared with *Stk11*$^{+/+}$ mice, one-way ANOVA, Dunnett's multiple comparisons test. (**G** and **H**) *Cxcl2* mRNA levels in LT^{het} (n = 4 or 5) and *Stk11*$^{+/-}$ (n = 6 to 13) mice aged >52 weeks as in (E) and (F). Gene expression relative to *Tbp* mRNA levels. Data are represented as the mean ± SEM. *P < 0.05 compared with LT^{wt} or *Stk11*$^{+/+}$ mice, one-way ANOVA, Dunnett's multiple comparisons test. (**I**) Representative flow cytometric plots for frequency of CD11b^{+} and Gr1^{+} neutrophils and inflammatory monocytes isolated from GI tissue from control (LT^{wt}) (n = 13) or diseased LT^{het} (n = 13) mice (CD45^{+} population). (**J**) Representative H&E staining of stomach (S), small intestine (D), and polyp (P) sections from *Stk11*$^{+/+}$ (n = 5) and *Stk11*$^{+/-}$ (n = 7) mice 16 weeks after irradiation (4.5 Gy). Scale bars, 1 mm. (**K**) Size of polyps in *Stk11*$^{+/-}$ mice 16 weeks after irradiation as in (J). Data are represented as the mean ± SEM and are representative of two separate experiments. *P < 0.05 compared with nonirradiated *Stk11*$^{+/-}$ mice, unpaired Student's t test.

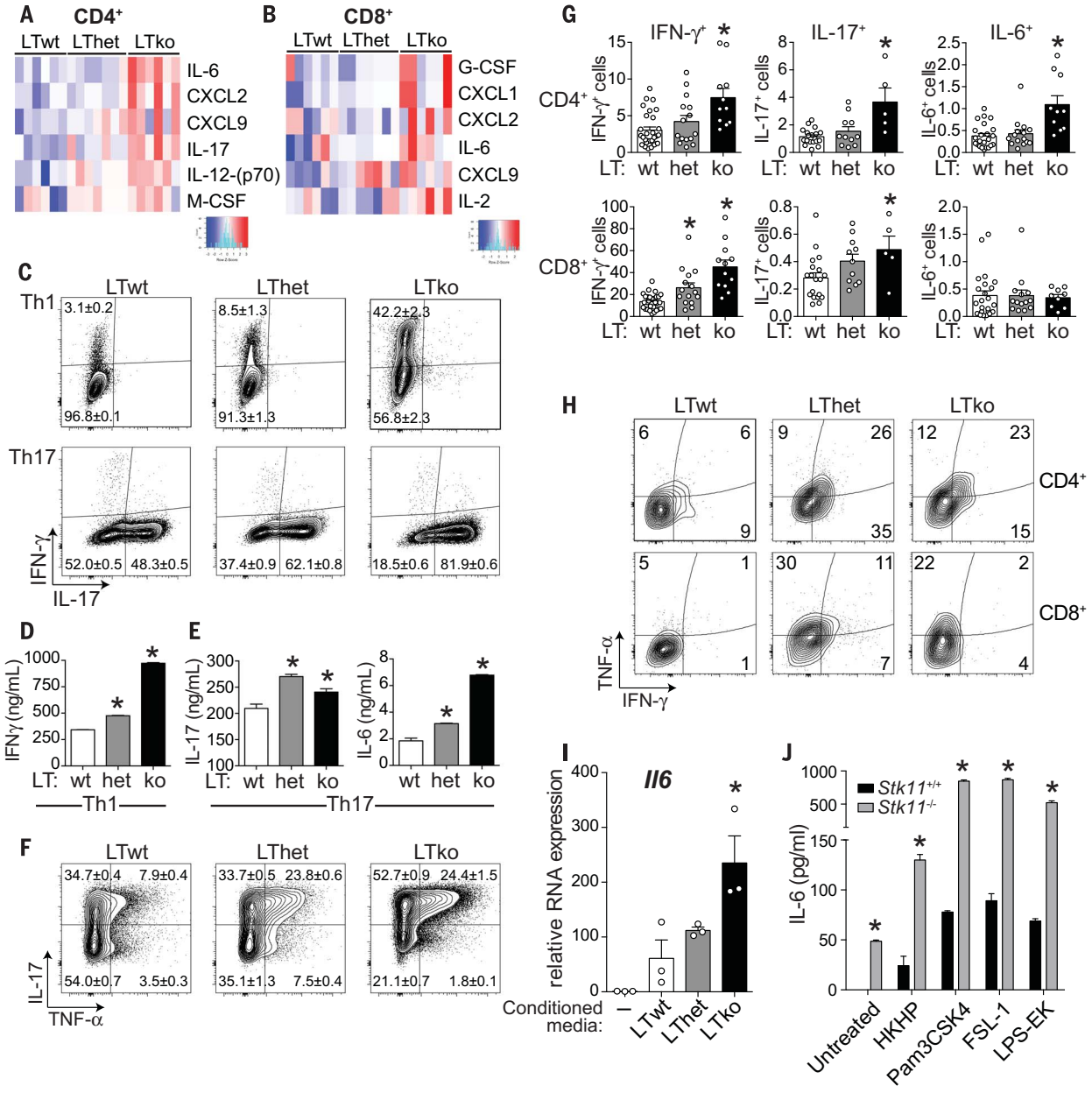


Fig. 3. Deregulated cytokine production and stimulation of inflammation by LKB1-deficient T cells. (**A** and **B**) Heatmap of cytokine production by activated unpolarized LT^{wt} (n = 6), LT^{het} (n = 7), and LT^{ko} (n = 6) (A) CD4$^+$ and (B) CD8$^+$ T cells as measured with MILLIPLEX assay. Shown are cytokines with significant (P < 0.05) up-regulation in LT^{ko} T cells. (**C**) Representative flow cytometry plots for IL-17 and IFN-γ production by LT^{wt}, LT^{het}, and LT^{ko} CD4$^+$ T cells cultured under T$_H$1- and T$_H$17-polarizing conditions. Values displayed represent average percent ± SEM for biological replicates (n = 2 or 3). Data are representative of three separate experiments. (**D** and **E**) Enzyme-linked immunosorbent assay for IFN-γ production by T$_H$1 cells (D) or IL-17 and IL-6 production by T$_H$17 cells (E) of the indicated genotypes generated as in (C). Data represent the mean ± SEM for technical triplicates. *P < 0.05 compared with LT^{wt} control samples, one-way ANOVA, Dunnett's multiple comparisons test. (**F**) Representative flow cytometric plots showing frequency of IL-17 and TNF-α production by LT^{wt}, LT^{het}, and LT^{ko} CD4$^+$ T cells cultured under T$_H$17-polarizing conditions as in (C). Values displayed represent average percent ± SEM for biological replicates (n = 2 or 3). (**G**) Number of IFN-γ–, IL-17–, and IL-6–producing CD4$^+$ and CD8$^+$ T cells in the mesenteric lymph nodes (mLN) of LTwt (n = 19 to 29), LT^{het} (n = 11 to 15), and LT^{ko} (n = 5 to 12) mice (>52 weeks of age). Cytokine production by CD4$^+$ and CD8$^+$ T cells was determined by means of intracellular staining after restimulation of T cells directly ex vivo. *P < 0.05 compared with LT^{wt} control samples, one-way ANOVA, Dunnett's multiple comparisons test. (**H**) Representative flow cytometric plots for frequency of TNF-α– and IFN-γ–producing CD4$^+$ and CD8$^+$ T cells isolated from healthy GI tissue from control mice (LT^{wt}) (n = 13) or polyp-bearing LT^{het} (n = 13) or LT^{ko} (n = 2) mice. (**I**) *Il6* mRNA levels in MEFs after culture (24 hours) with supernatants from activated LT^{wt}, LT^{het}, or LT^{ko} CD8$^+$ T cells (n = 3 mice for each genotype). Untreated MEFs cultured with regular medium are shown. Gene expression relative to *Tbp* mRNA levels. *P < 0.05 compared with control samples (LT^{wt}), one-way ANOVA, Dunnett's multiple comparisons test. (**J**) IL-6 production by *Stk11*$^{+/+}$ or *Stk11*$^{-/-}$ MEFs after a 6-hour stimulation with LPS from *Escherichia coli* (LPS-EK), heat-killed *Helicobacter pylori* (HKHP), Pam3CSK4, FSL-1, or sham treatment. Data are represented as the mean ± SEM for biological duplicates. Data are representative of two separate experiments. *P < 0.05 compared with *Stk11*$^{+/+}$ MEF samples, unpaired Student's t test.

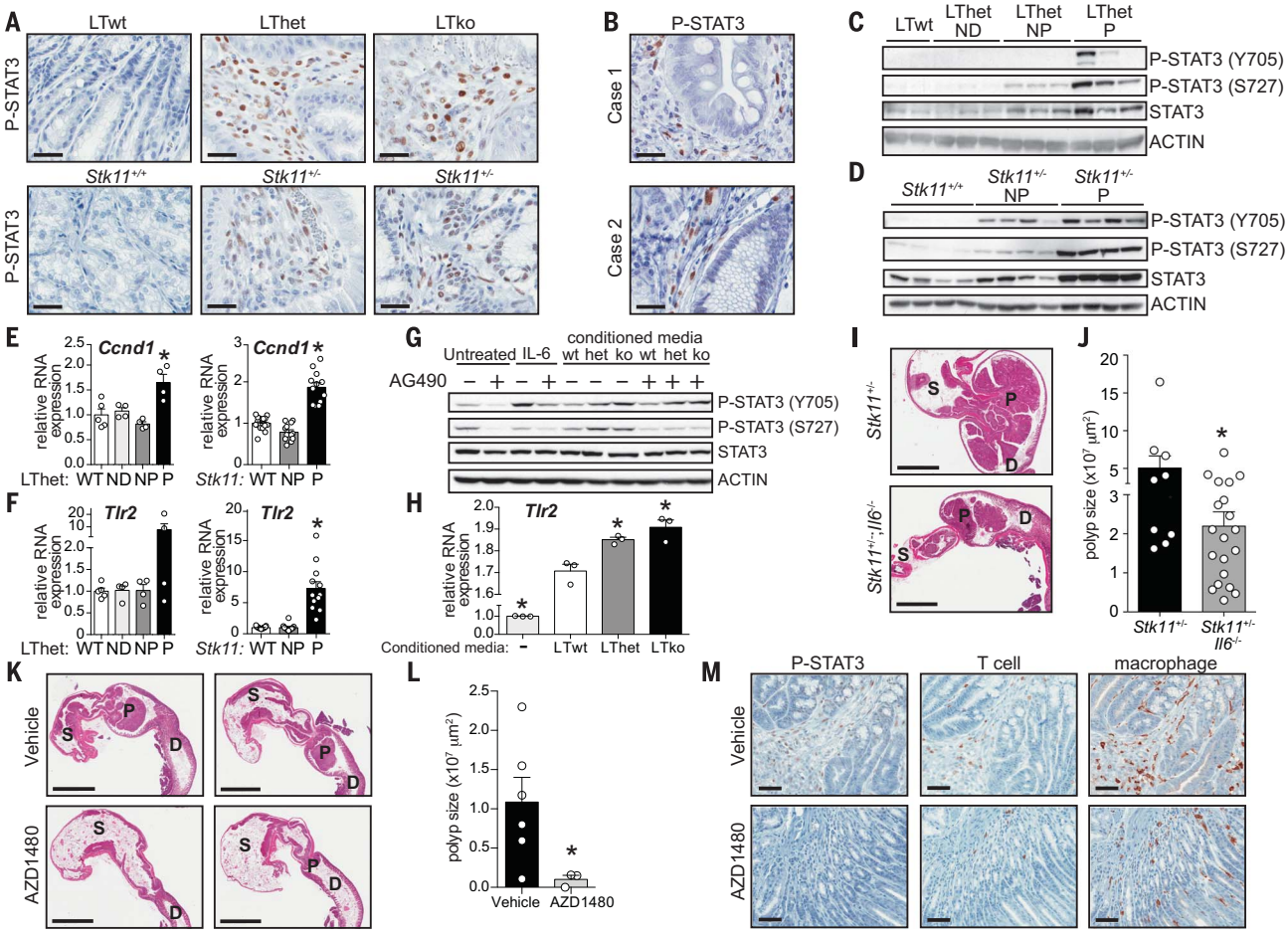


Fig. 4. LKB1-deficient T cells promote STAT3 activation and expression of STAT3-dependent growth-promoting and inflammatory genes in PJS polyps. (**A** and **B**) Representative immunohistochemistry (IHC) for P-STAT3 (Y705) staining in healthy tissue from control animals (LT^{wt} and $Stk11^{+/+}$) (n = 10 and 8, respectively) and polyps from (A) LT^{het} (n = 11), LT^{ko} (n = 12), and $Stk11^{+/-}$ (n = 8) mice aged >52 weeks or (B) PJS patients (n = 2). Scale bars, 30 µm. (**C** and **D**) Immunoblot for STAT3 activation (pY705 and pS727) in control (LT^{wt} and $Stk11^{+/+}$), nondiseased LT^{het} (ND), polyp (P), and adjacent healthy tissue (NP) from (C) LT^{het} and (D) $Stk11^{+/-}$ mice aged >52 weeks. (**E** and **F**) Relative mRNA expression for STAT3 target genes *Ccnd1* (E) and *Tlr2* (F) in (left) LT^{het} (n = 4 or 5) and (right) $Stk11^{+/-}$ mice (n = 11 to 16) aged >52 weeks. WT refers to paired control animals (LT^{wt} or $Stk11^{+/+}$) for each experiment. (**G**) Immunoblot for STAT3 activation in MEFs treated for 6 hours with IL-6 or supernatants from activated LT^{wt}, LT^{het}, or LT^{ko} $CD8^+$ T cells. Cells were treated with the JAK2 inhibitor AG490 for 1 hour before conditioned media treatment where indicated. Data are representative of three separate experiments. (**H**) *Tlr2* gene expression in MEFs after culture (24 hours) with supernatants from activated LT^{wt}, LT^{het}, or LT^{ko} T cells (n = 3 mice per genotype). Gene expression is relative to *Tbp* mRNA levels. Data represent the mean ± SEM. *P < 0.05 compared with control samples (LT^{wt}), one-way ANOVA, Dunnett's multiple comparisons test. (**I**) Representative H&E staining of GI sections from $Stk11^{+/-}$ (n = 9) or $Stk11^{+/-}$;$Il6^{-/-}$ (n = 20) mice aged >52 weeks. Scale bars, 5 mm. (**J**) Polyp size for animals in (I). Data represent the mean ± SEM. *P < 0.05 compared with control samples ($Stk11^{+/-}$ mice), unpaired Student's t test. (**K**) Representative H&E staining of GI sections from aged $Stk11^{+/-}$ mice treated orally with vehicle control (n = 6) or AZD1480 (n = 3) for 8 weeks. Scale bars, 5 mm. (**L**) Polyp size of vehicle- and AZD1480-treated $Stk11^{+/-}$ mice as in (K). Data represent the mean ± SEM. *P < 0.05 compared with control samples (vehicle-treated mice), Welch's t test. (**M**) Representative IHC for P-STAT3 (Y705), CD3 (T cell), and Iba-1 (macrophage) in GI sections from aged $Stk11^{fs-}$ mice treated with vehicle control (n = 3) or AZD1480 (n = 3) in (K). Scale bars, 50 µm.

sufficient to stimulate STAT3 phosphorylation (Fig. 4G) and promote increased expression of a subset of STAT3 target genes, including *Tlr2*, in nontransformed MEFs (Fig. 4H and fig. S11B).

Last, we assessed the contribution of IL-6–STAT3 signaling to PJS polyp growth. GI polyp size was significantly reduced in $Stk11^{+/-}$ mice lacking IL-6 expression (Fig. 4, I and J), similar to that observed in lymphocyte-deficient $Stk11^{+/-}$ mice (Fig. 2, C and D). Polyps from $Stk11^{+/-}$;$Il6^{-/-}$ mice still retained phospho-STAT3 (Y705) staining (fig. S11C), suggesting that other STAT3-activating factors such as IL-11 may be active in these tissues. Last, we inhibited STAT3 signaling in $Stk11^{+/-}$ mice in vivo using the JAK2 inhibitor AZD1480 (*24*, *25*). AZD1480-treated $Stk11^{+/-}$ mice displayed significant reductions in polyp size compared with that of controls (Fig. 4, K and L). AZD1480 treatment reduced phospho-STAT3 levels and immune cell infiltration in GI tissues of $Stk11^{+/-}$ mice (Fig. 4M and fig. S12, A and B), whereas mTORC1 activity was unaffected (fig. S12A). Thus, the blockade of STAT3 signaling can affect polyp development in this PJS mouse model.

The development of hamartomatous polyps in PJS patients has been attributed to cell-intrinsic tumor suppressor functions for *Stk11* in epithelial and/or stromal tissue cells in the GI tract. Our findings establish immune-mediated inflammation as a hallmark of PJS disease and highlight a critical role for *Stk11* mutant T cells in PJS disease progression. Our data argue for a more complex role for *Stk11* mutations in PJS disease development, with deregulated inflammatory

responses by LKB1 mutant immune cells, in addition to epithelial and stromal tissues (*26*), reinforcing tumor inflammation and chronic STAT3 activation to drive polyp growth. Consistent with this, elevated IL-6 levels have been observed in LKB1-deficient tumors (*27*). Our data raise the possibility that inflammatory events in the GI tract (pathogen interactions with *Stk11*$^{+/-}$ immune cells) may trigger the inflammation we have found associated with PJS polyps in mice and humans to stimulate intestinal polyposis. Targeting chronic GI inflammation may present a novel approach to reducing disease incidence and polyp burden in PJS patients.

REFERENCES AND NOTES

1. D. R. Alessi, K. Sakamoto, J. R. Bayascas, *Annu. Rev. Biochem.* **75**, 137–163 (2006).
2. D. B. Shackelford, R. J. Shaw, *Nat. Rev. Cancer* **9**, 563–575 (2009).
3. A. Hemminki *et al*., *Nat. Genet.* **15**, 87–90 (1997).
4. D. E. Jenne *et al*., *Nat. Genet.* **18**, 38–43 (1998).
5. N. Bardeesy *et al*., *Nature* **419**, 162–167 (2002).
6. K. Jishage *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8903–8908 (2002).
7. H. Miyoshi *et al*., *Cancer Res.* **62**, 2261–2266 (2002).
8. H. Jeghers, V. A. McKusick, K. H. Katz, *N. Engl. J. Med.* **241**, 993–1005, illust passim (1949).
9. F. M. Giardiello *et al*., *Gastroenterology* **119**, 1447–1453 (2000).
10. P. Katajisto *et al*., *Nat. Genet.* **40**, 455–459 (2008).
11. B. Gan *et al*., *Nature* **468**, 701–704 (2010).
12. S. Gurumurthy *et al*., *Nature* **468**, 659–663 (2010).
13. D. Nakada, T. L. Saunders, S. J. Morrison, *Nature* **468**, 653–658 (2010).
14. M. Nakau *et al*., *Cancer Res.* **62**, 4549–4553 (2002).
15. D. B. Shackelford *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11137–11142 (2009).
16. S. Grivennikov *et al*., *Cancer Cell* **15**, 103–113 (2009).
17. T. L. Putoczki *et al*., *Cancer Cell* **24**, 257–271 (2013).
18. M. Ernst *et al*., *J. Clin. Invest.* **118**, 1727–1738 (2008).
19. Z. Y. Ong *et al*., *Radiat. Oncol.* **5**, 22 (2010).
20. N. J. MacIver *et al*., *J. Immunol.* **187**, 4187–4198 (2011).
21. K. Taniguchi, M. Karin, *Semin. Immunol.* **26**, 54–74 (2014).
22. T. Tahara *et al*., *Cancer Sci.* **98**, 1790–1794 (2007).
23. H. Tye *et al*., *Cancer Cell* **22**, 466–478 (2012).
24. M. Hedvat *et al*., *Cancer Cell* **16**, 487–497 (2009).
25. E. Stuart *et al*., *Mol. Cancer Ther.* **13**, 468–474 (2014).
26. S. Ollila *et al*., *J. Clin. Invest.* **128**, 402–414 (2018).
27. S. Koyama *et al*., *Cancer Res.* **76**, 999–1008 (2016).

ACKNOWLEDGMENTS

We thank W. Foulkes, C. Duerr, and L. Osborne as well as members of the Jones laboratory for critical reading of the manuscript. We thank M. Ernst, B. Jenkins, A. West, G. Jones, and S. Jones for their experimental advice. We acknowledge technical support from the Histology Facilities of the Goodman Cancer Research Centre (GCRC) and Deeley Cancer Centre (BC Cancer Agency), J.-M. Lapointe of the Comparative Medicine Animal Research Centre (McGill), the Flow Cytometry Facility at McGill University, and the Centre for Applied Genomics (Hospital for Sick Children). We thank AstraZeneca for access to AZD1480 through its Open Innovation program. **Funding:** This work was supported by grants from the Canadian Cancer Society (CCSRI; 702566 to R.G.J.) and the Canadian Institutes of Health Research (CIHR; MOP-93799 and PJT-156397 to R.G.J., MOP-86582 to N.B., and MOP-142351 to J.J.L.). Fellowship support was provided from the McGill Integrated Cancer Research Training Program (to B.F., J.C., and E.H.M.), the Defi Canderel (to M.C.P.), the Fonds de Recherche du Québec–Santé (FRQS; to M.C.P., E.H.M., G.Z., and J.B.), and the CIHR (to M.C.P. and R.G.J.). **Author contributions:** Conceptualization was done by M.C.P., J.B., and R.G.J.; investigation was performed by M.C.P., J.B., A.M.-R., E.A., J.C., B.E.H., A.H.W., B.F., B.S., E.H.M., S.-P.G., L.T., L.D., P.K., and E.L.; resources were provided by R.M.J., J.J.L., A.H., S.I., N.B., and M.N.A.; writing and visualization were done by M.C.P. and R.G.J.; supervision and project administration was performed by M.C.P., P.M.S., M.N.A., J.J.L., G.Z., and R.G.J.; and funding acquisition was done by R.G.J. **Competing interests:** The authors declare no competing interests associated with this work. **Data and materials availability:** RNA-seq datasets can be found at NCBI (GSE115150). AZD1480 is available from AstraZeneca under a materials transfer agreement with R.G.J. and McGill University.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/406/suppl/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S3
References (*28*–*35*)

7 April 2017; resubmitted 6 February 2018
Accepted 14 June 2018
10.1126/science.aan3975

MOLECULAR BIOLOGY

Mediator and RNA polymerase II clusters associate in transcription-dependent condensates

Won-Ki Cho[1]*, Jan-Hendrik Spille[1]*, Micca Hecht[1], Choongman Lee[1], Charles Li[2,3], Valentin Grube[1,4], Ibrahim I. Cisse[1]†

Models of gene control have emerged from genetic and biochemical studies, with limited consideration of the spatial organization and dynamics of key components in living cells. We used live-cell superresolution and light-sheet imaging to study the organization and dynamics of the Mediator coactivator and RNA polymerase II (Pol II) directly. Mediator and Pol II each form small transient and large stable clusters in living embryonic stem cells. Mediator and Pol II are colocalized in the stable clusters, which associate with chromatin, have properties of phase-separated condensates, and are sensitive to transcriptional inhibitors. We suggest that large clusters of Mediator, recruited by transcription factors at large or clustered enhancer elements, interact with large Pol II clusters in transcriptional condensates in vivo.

A conventional view of eukaryotic gene regulation is that transcription factors, bound to enhancer DNA elements, recruit coactivators such as the Mediator complex, which is thought to interact with RNA polymerase II (Pol II) at the promoter (*1*–*5*). This model is supported by a large body of molecular genetic and biochemical evidence, yet the direct interaction of Mediator and Pol II has not been observed and characterized in living cells (*6*). Using superresolution (*7*–*9*) and light-sheet imaging (*10*), we studied the organization and dynamics of endogenous Mediator and Pol II in live mouse embryonic stem cells (mESCs). We directly tested whether Pol II and Mediator interact in a manner consistent with condensate formation (*11*–*13*), quantitatively characterized their biophysical properties, and considered the implications of these observations for transcription regulation in living mammalian cells.

To visualize Mediator and Pol II in live cells, we generated mESC lines with endogenous Mediator and Pol II labeled with Dendra2, a green-to-red photoconvertible fluorescent protein (materials and methods and figs. S1 and S2). We performed live-cell superresolution imaging and found that Mediator forms clusters (Fig. 1A and fig. S3) with a range of dynamic temporal signatures. Mediator exists in a population of transient small (~100 nm) clusters (Fig. 1B) with an average lifetime of 11.1 ± 0.9 s (mean ± SEM from 36 cells) (Fig. 1G), comparable to that of transient Pol II clusters observed in this study (Fig. 1, D, E, and H) and previously in differentiated cell types (*14*, *15*). In addition, we observed that both Mediator and Pol II form a population of large (>300 nm) clusters (~14 per cell), each comprising ~200 to 400 molecules, that are temporally stable (lasting the full acquisition window of the live-cell superresolution imaging) (Fig. 1, C and F to H, and figs. S4 to S6).

We tested the extent to which these clusters depend on the stem cell state. The mESCs were subjected to a protocol (*16*) to differentiate them into epiblastlike cells (EpiLCs) within 24 h (materials and methods and fig. S7). Differentiation had no apparent effect on the population of transient clusters, consistent with previous observations that transient clusters persist in differentiated cell types (*14*, *15*). However, both the size and the number of stable clusters decreased along the course of differentiation (fig. S8), suggesting that these stable clusters are prone to change as cells differentiate.

We focused on the stable clusters of Mediator and Pol II and investigated whether they are colocalized. We generated mESCs with endogenous Mediator and Pol II tagged with JF646-HaloTag (*15*, *17*) and Dendra2, respectively (materials and methods and figs. S1 and S2). Direct imaging of both JF646-Mediator (Fig. 2A) and Dendra2–Pol II (Fig. 2B) showed bright spots of large accumulations in the nucleus, which corresponded to stable Pol II clusters according to subsequent superresolution imaging of Dendra2–Pol II in the same nuclei (Fig. 2C). The same observations were made with Dendra2-Mediator (fig. S9). Of 143 Mediator clusters imaged by dual-color light-sheet imaging (Fig. 2, D to F), 129 (90%) had a colocalizing Pol II cluster (Fig. 2, G and H; materials and methods; and fig. S9). We conclude that these Mediator and Pol II clusters colocalize in live mESCs.

Previous studies have shown that high densities of Mediator are located at enhancer clusters called super-enhancers (SEs) and that some are

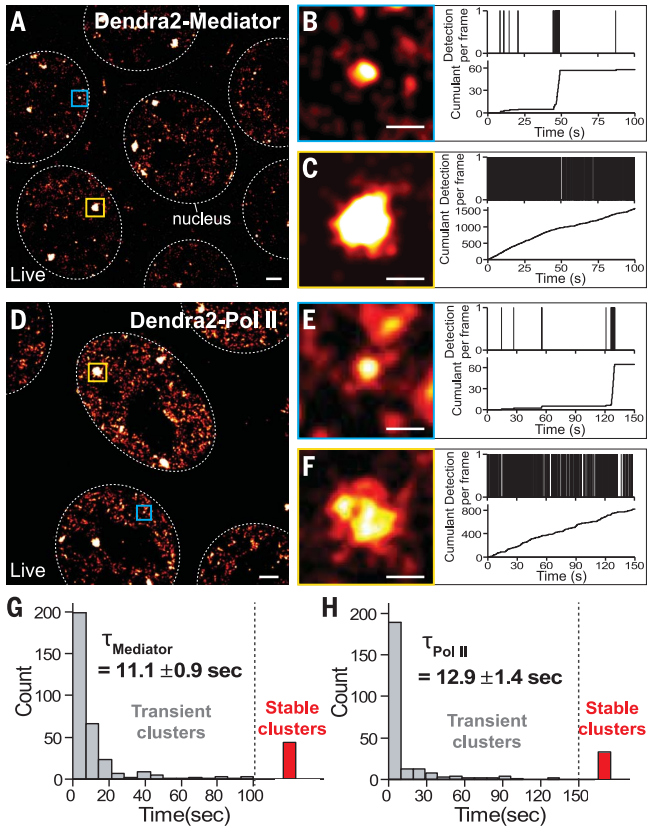


Fig. 1. Mediator and Pol II form transient and stable clusters in living mESCs. (**A**) A superresolution image of endogenous Mediator labeled with Dendra2 in living mESCs. (**B** and **C**) Representative superresolved images of transient and stable Mediator clusters and corresponding time-correlated photoactivation localization microscopy (tcPALM) traces. (B) and (C) correspond to areas boxed in blue and yellow, respectively, in (A). (**D**) Superresolution image of endogenous Pol II labeled with Dendra2 in living mESCs. (**E** and **F**) Representative superresolution images of transient and stable Pol II clusters and corresponding tcPALM traces. (E) and (F) correspond to areas boxed in blue and yellow, respectively, in (D). (**G** and **H**) Lifetime distributions of Mediator and Pol II clusters, respectively. Scale bars, 1 μm in (A) and (D) and 500 nm in (B), (C), (E), and (F).

[1]Department of Physics, MIT, Cambridge, MA 02139, USA. [2]Department of Biology, MIT, Cambridge, MA 02139, USA. [3]Whitehead Institute for Biomedical Research, Cambridge, MA 02139, USA. [4]Department of Physics, LMU Munich, Geschwister Scholl Platz 1, 80539 Munich, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: icisse@mit.edu

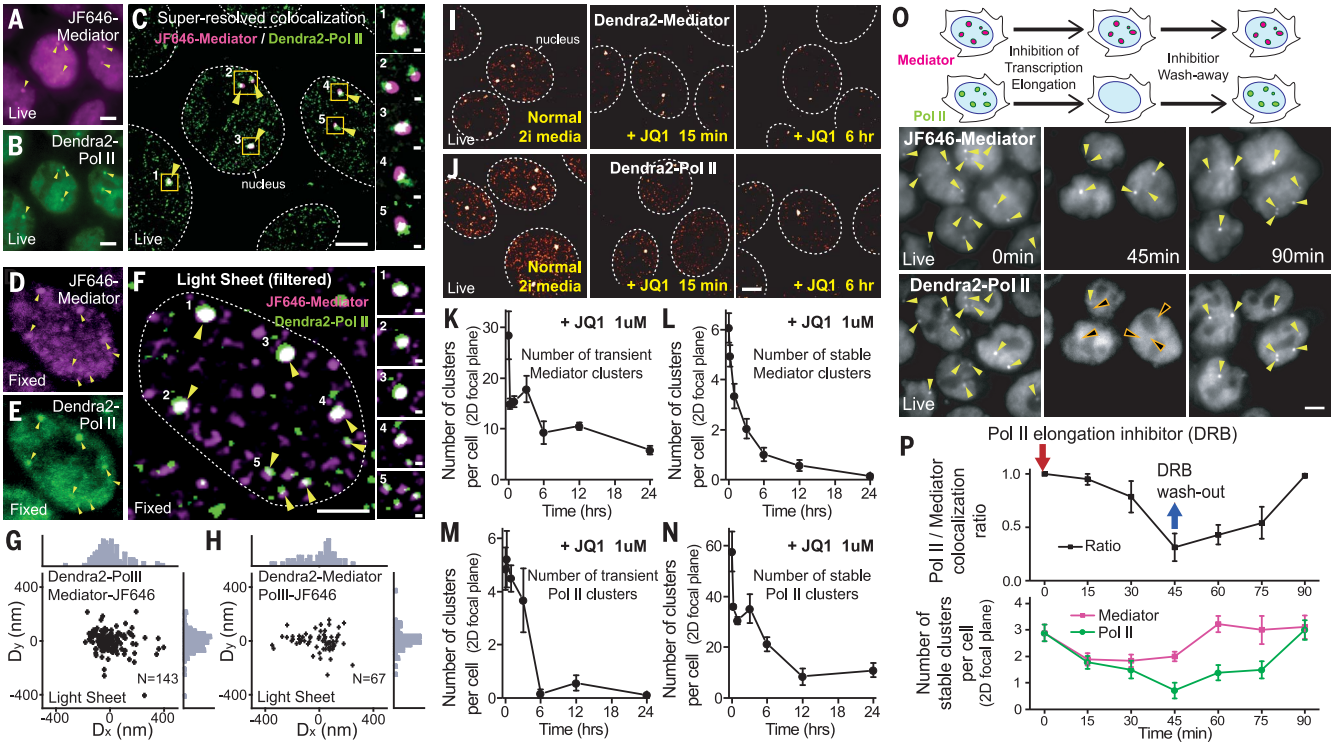


Fig. 2. Mediator and Pol II clusters colocalize in a transcription-dependent manner. Live-cell direct images of (**A**) JF646-Mediator and (**B**) Dendra2–Pol II. Yellow arrowheads indicate stable clusters. (**C**) Super-resolution image of Dendra2–Pol II overlaid with a background-subtracted JF646-Mediator image. Insets 1 to 5 show Mediator and Pol II colocalization in clusters. (**D**) JF646-Mediator and (**E**) Dendra2–Pol II maximum-intensity projections of a fixed cell imaged by lattice light-sheet microscopy. (**F**) Overlay of background-subtracted images. Yellow arrowheads indicate clusters identified in the Dendra2–Pol II channel. (**G**) Scatter plot of the distance *D* from a Dendra2–Pol II cluster to the nearest JF646-Mediator cluster (n = 143 clusters). Histograms outside the scatter plot show the distances along the *x* and *y* axes. (**H**) Same analysis for clusters identified in the Dendra2-Mediator channel (n = 67). (**I** and **J**) Superresolution images of Dendra2-Mediator and Dendra2–Pol II under normal conditions (left) and after 15 min (middle) or 6 hours (right) of incubation in 1 μM JQ1. (**K** and **M**) The number of transient Mediator and Pol II clusters per cell in a 2D focal plane as a function of time after JQ1 addition. (**L** and **N**) The number of stable Mediator and Pol II clusters per cell in a 2D focal plane. n = 17 to 25 cells and n = 14 to 24 cells at each JQ1 time point for Mediator and Pol II, respectively. (**O**) DRB treatment and washout experiments. DRB (100 μM) was added at 0 min and washed away after 45 min. Arrowheads indicate stable clusters identified in the JF646-Mediator channel. Black arrowheads in the middle panel (bottom) indicate Mediator clusters that did not colocalize with Pol II clusters. (**P**) Ratio (top) and absolute number (bottom) of clusters detected in the Pol II and Mediator channels per cell in a 2D focal plane. Nine to 15 cells were analyzed for each DRB incubation time point. The red arrow indicates the addition of DRB, and the blue arrow indicates DRB washout. Scale bars, 2 μm in overview images and 200 nm in insets.

disrupted by loss of the BET (bromodomain and extraterminal family) protein BRD4, which is a cofactor associated with Mediator (*18*, *19*). We found that treatment of mESCs with JQ1, a drug that causes loss of BRD4 from enhancer chromatin, dissolved transient and stable clusters of both Mediator and Pol II clusters (Fig. 2, I to N, and fig. S10).

After transcription initiation, Pol II transcribes a short distance (~100 base pairs), pauses, and is released to continue elongation when phosphorylated by CDK9 (*20*). We hypothesized that inhibition of CDK9 might selectively affect the Pol II stable clusters. We observed that upon incubation with DRB (5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole), Pol II stable clusters dissolved but Mediator stable clusters remained (Fig. 2O). Quantification of Mediator–Pol II colocalization revealed that incubation with DRB progressively decreased the fraction of Mediator stable clusters that colocalized with Pol II (Fig. 2P). This effect could be reversed when DRB was washed out; the colocalization fraction recovered completely. These results imply that the association between Mediator and Pol II clusters may be hierarchical, with upstream enhancer recruitment controlling both clusters but downstream transcription inhibition selectively affecting Pol II clusters.

We characterized the long-term dynamics of stable clusters by using lattice light-sheet imaging in live mESCs (movies S1 and S2). We observed that clusters can merge upon contact (Fig. 3, A to D, and movies S1 and S2). The time scale of coalescence was very rapid, comparable to our full volumetric acquisition frame rate (15-s time interval). The added-up intensity of the two precursor clusters was close to that of the newly merged cluster (Fig. 3E and fig. S11). These biophysical dynamics are reminiscent of those of biomolecular condensates in vivo (*21*).

In addition to coalescence, in vivo condensates had rapid turnover of the molecular components, as shown by fast recovery in fluorescence recovery after photobleaching (FRAP) assays, and were sensitive to a nonspecific aliphatic alcohol, 1,6-hexanediol (*21*). Our FRAP analyses of clusters revealed very rapid dynamics and turnover of their components: 60% of the Mediator and 90% of Pol II components were exchanged within ~10 s within clusters (Fig. 3, F to H). Moreover, the treatment of mESCs with 1,6-hexanediol resulted in the gradual dissolution of both Mediator and Pol II clusters (Fig. 3, I to K, and fig. S12). Together, these results suggest that the stable clusters are in vivo condensates of Mediator and Pol II.

We hypothesized that a phase separation model with induced condensation at the recruitment step of Mediator to enhancers would qualitatively account for the observations in this study (*22*). The model implies that the condensates are chromatin associated and colocalize with enhancer-controlled active genes. We therefore tested these two specific implications. We tracked the diffusion dynamics of Mediator clusters by computing their mean squared displacement as a function of time (n = 6 cells). On short time scales, the cluster motion was subdiffusive,

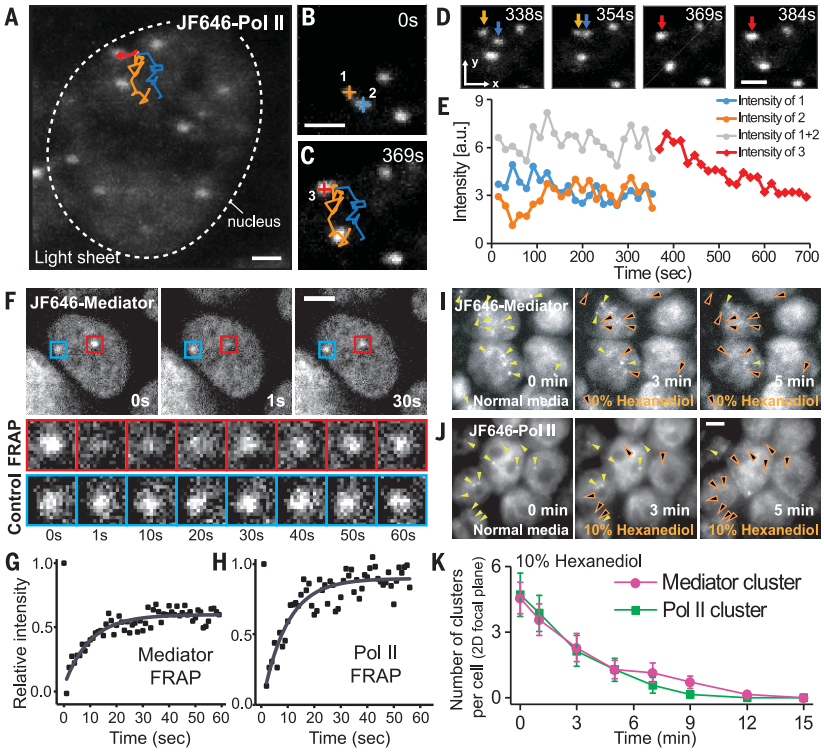


Fig. 3. Mediator and Pol II form condensates that coalesce, recover in FRAP, and are sensitive to hexanediol. (**A** to **E**) Cluster fusion. (A) Maximum-intensity projection of a live cell imaged by lattice light-sheet microscopy. Trajectories of two clusters are indicated. (B and C) Clusters observed at 0 s and fusing at 369 s. (D) Individual time points around the fusion event. Orange and blue arrows indicate the precursor clusters, and the red arrows indicate the fused cluster. (E) Time course of the cluster intensities. a.u., arbitrary units. (**F** to **H**) FRAP analysis of clusters. (F) (Top) Images of a JF646-Mediator cell before (0 s) (left), immediately after (1 s) (middle), and 30 s after (right) bleaching. The red box indicates the position of the cluster on which the FRAP beam was focused. The blue box indicates an unbleached control locus. (Bottom) Cropped images as a function of time for both loci. (G) The normalized recovery curve for Mediator (n = 9 cells) yielded a recovery fraction of 60% during the 60-s observation, with a half-recovery time of 10 s. (H) FRAP analysis of JF646–Pol II (n = 3 cells) yielded 90% recovery, with an identical half-recovery time of 10 s. (**I** and **J**) Treatment with 10% hexanediol (v/v) gradually dissolved clusters of JF646-Mediator (I) and JF646–Pol II (J). Maximum-intensity projections of epifluorescence z stacks are shown. Yellow arrowheads indicate clusters identified at 0 min. Black arrowheads indicate clusters that disappeared. (**K**) Average number of clusters per cell (single 2D focal plane) observed in direct imaging as a function of time after hexanediol addition (n = 14 cells for JF646-Mediator, and n = 14 cells for JF646–Pol II). Scale bars, 1 μm (A to D) and 5 μm (F and J).

with an exponent $\alpha = 0.40 \pm 0.12$ (best fit ± SEM) (fig. S13). This is the same exponent found in the subdiffusional behavior of chromatin loci in eukaryotic cells (*23*–*27*). We also observed the same diffusional parameters when tracking a chromatin locus labeled by dCas9-based chimeric array of guide RNA oligonucleotides (CARGO) in mESCs (fig. S13) (*23*). We concluded that clusters diffuse like chromatin-associated domains.

We hypothesized that clusters were in close physical proximity to actively transcribed genes that can be visualized by global run-on nascent RNA labeling with ethynyl uridine (EU) (fig. S14). The run-on results showed that 2 min after DRB washout, virtually all Mediator clusters observed were proximal or overlapping with nascent RNA accumulations, as imaged by Click labeling of EU in fixed cells (fig. S14). We also employed the MS2 endogenous RNA labeling system (*15*, *28*) (materials and methods and fig. S15) to investigate whether active transcription could be observed at *Esrrb*, one of the top SE-controlled genes in mESCs (*29*) (Fig. 4A). We observed bright foci consistent with nascent MS2-labeled gene loci and confirmed the gene loci by dual-color RNA fluorescence in situ hybridization (FISH) targeting the MS2 sequence and intronic regions of *Esrrb* (fig. S16). Intronic FISH on 125 *Esrrb* loci from 82 fixed cells showed that 93% of *Esrrb* loci had a stable Mediator cluster nearby (within 1 μm) but only ~22% of the loci colocalized with a stable Mediator cluster, suggesting that the Mediator-bound enhancer only occasionally colocalizes with the gene (fig. S17). The variability in colocalization may be explained by a dynamic "kissing" model, where a distal Mediator cluster colocalizes with the gene only at certain time points (Fig. 4A).

By dual-color three-dimensional (3D) live-cell imaging with lattice light-sheet microscopy, we found that some Mediator clusters were up to a micrometer away from the active *Esrrb* gene locus but in some instances directly colocalized with the gene (Fig. 4, B and C). In addition, we directly observed the dynamic interaction between Mediator clusters and the gene locus, supporting the dynamic kissing model (Fig. 4, D and E; fig. S18; and movie S3). Tracking of loci in all six cells indicated that colocalization below our resolution limit of 300 nm occurred at ~30% of the time points (Fig. 4F). However, even when they were not overlapping, the Mediator cluster and the gene loci moved as a pair through the nucleus (movie S3), consistent with two adjacent regions anchoring to the same underlying chromatin domain. We propose that Mediator clusters form at the *Esrrb* SE and then interact occasionally and transiently with the transcription apparatus at the *Esrrb* promoter.

We have found that Mediator and Pol II form large stable clusters in living cells and have shown that these clusters have properties expected for biomolecular condensates. The condensate properties were evident through coalescence, rapid recovery in FRAP analysis, and sensitivity to hexanediol. In a model of phase separation on the basis of scaffold-client relationships (*30*), it is possible that enhancer-associated Mediator forms a condensate and provides a "scaffold" for "client" RNA Pol II molecules. The model we propose whereby large Mediator clusters at

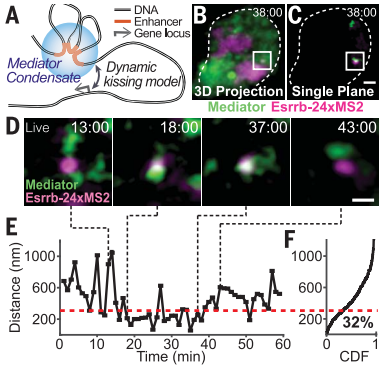


Fig. 4. Mediator clusters dynamically kiss actively transcribing SE-controlled genes. (**A**) Illustration of the working model describing cluster-kissing interaction with a gene locus. (**B**) Maximum-intensity projection of a cell imaged by using lattice light-sheet microscopy showing colocalization of a JF646-Mediator cluster with the actively transcribing *Esrrb* gene locus marked by MS2-tagged RNA (white box). (**C**) Single plane from the z stack after background subtraction. (**D**) Snapshot images of a Mediator cluster near the actively transcribing *Essrb* gene locus. Time points listed in (B) to (D) indicate minutes after the start of acquisition. Scale bars, 2 μm in (B) and (C) and 500 nm in (D). (**E**) Plot of the centroid-to-centroid distance from the gene locus to the nearest cluster as a function of time. (**F**) Cumulative distribution of distances from the *Esrrb* locus to the nearest Mediator cluster pooled from six cells (291 time points). The red dashed line in (E) and (F) indicates the colocalization threshold (300 nm). CDF, cumulative distribution function.

enhancers transiently kiss the transcription apparatus at promoters has a number of implications for gene control mechanisms. The presence of large Mediator clusters at some enhancers may allow Mediator condensates to contact the transcription apparatus at multiple gene promoters simultaneously. The large size of the Mediator clusters may also mean that the effective distance of the enhancer-promoter DNA elements can be in the same order as the size of the clusters (>300 nm), larger than the distance requirement for direct contact. We speculate that such clusters may help explain gaps of hundreds of nanometers that are found in previous studies measuring distances between functional enhancer-promoter DNA elements. Such cluster sizes also imply that some long-range interactions could go undetected in DNA interaction assays that depend on much closer physical proximity of enhancer and promoter DNA elements.

REFERENCES AND NOTES

1. P. J. Robinson *et al*., *Cell* **166**, 1411–1422.e16 (2016).
2. R. D. Kornberg, *Trends Biochem. Sci.* **30**, 235–239 (2005).
3. C. T. Ong, V. G. Corces, *Nat. Rev. Genet.* **12**, 283–293 (2011).
4. K. M. Lelli, M. Slattery, R. S. Mann, *Annu. Rev. Genet.* **46**, 43–68 (2012).
5. B. L. Allen, D. J. Taatjes, *Nat. Rev. Mol. Cell Biol.* **16**, 155–166 (2015).
6. M. Levine, C. Cattoglio, R. Tjian, *Cell* **157**, 13–25 (2014).
7. S. T. Hess, T. P. Girirajan, M. D. Mason, *Biophys. J.* **91**, 4258–4272 (2006).
8. M. J. Rust, M. Bates, X. Zhuang, *Nat. Methods* **3**, 793–796 (2006).
9. E. Betzig *et al*., *Science* **313**, 1642–1645 (2006).
10. B. C. Chen *et al*., *Science* **346**, 1257998 (2014).
11. D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, P. A. Sharp, *Cell* **169**, 13–23 (2017).
12. T. Fukaya, B. Lim, M. Levine, *Cell* **166**, 358–368 (2016).
13. Z. Liu *et al*., *eLife* **3**, e04236 (2014).
14. I. I. Cisse *et al*., *Science* **341**, 664–667 (2013).
15. W. K. Cho *et al*., *eLife* **5**, e13617 (2016).
16. C. Buecker *et al*., *Cell Stem Cell* **14**, 838–853 (2014).
17. J. B. Grimm *et al*., *Nat. Methods* **12**, 244–250 (2015).
18. P. Filippakopoulos *et al*., *Nature* **468**, 1067–1073 (2010).
19. J. Lovén *et al*., *Cell* **153**, 320–334 (2013).
20. N. F. Marshall, D. H. Price, *Mol. Cell. Biol.* **12**, 2078–2090 (1992).
21. Y. Shin, C. P. Brangwynne, *Science* **357**, eaaf4382 (2017).
22. B. R. Sabari *et al*., *Science* **361**, eaar3958 (2018).
23. B. Gu *et al*., *Science* **359**, 1050–1055 (2018).
24. J. R. Chubb, S. Boyle, P. Perry, W. A. Bickmore, *Curr. Biol.* **12**, 439–445 (2002).
25. J. S. Lucas, Y. Zhang, O. K. Dudko, C. Murre, *Cell* **158**, 339–352 (2014).
26. V. Dion, S. M. Gasser, *Cell* **152**, 1355–1364 (2013).
27. S. C. Weber, A. J. Spakowitz, J. A. Theriot, *Phys. Rev. Lett.* **104**, 238102 (2010).
28. E. Bertrand *et al*., *Mol. Cell* **2**, 437–445 (1998).
29. W. A. Whyte *et al*., *Cell* **153**, 307–319 (2013).
30. S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).

ACKNOWLEDGMENTS

We thank L. D. Lavis (HHMI, Janelia) and J. Grimm (HHMI, Janelia) for the gift of the JF646-Halo dyes and E. Calo (MIT) for the wild-type R1 cells and differentiation protocol. We thank J. Wysocka (Stanford) for the CARGO material. We thank R. Young (MIT) and members of the Young, Sharp, and Chakraborty groups (MIT) for helpful discussions and R. Young and J. Gore (MIT) for helpful comments on the manuscript. We acknowledge the students of the Cissé lab rotation in the 2017 Marine Biology Laboratory physiology course for participation in early aspects of Dendra2–Pol II characterization in mESCs and J. O. Andrews for assistance with the quantitative superresolution analysis software. The lattice light-sheet microscope was home built in the Cissé lab at MIT Physics under license from HHMI, Janelia Research Campus, and we thank E. Betzig (HHMI, Janelia) and W. Legant (HHMI, Janelia) for their critical support in the process. FRAP experiments were performed at the W. M. Keck Microscopy Facility at the Whitehead Institute. We thank L. Boyer for help with stem cell culture. **Funding:** This work was supported primarily by the NIH director's New Innovator award (DP2CA195769 to I.I.C.) and also by the Pew Charitable Trusts through the Pew Biomedical Scholars Program grant (to I.I.C.). I.I.C. is also supported by the NIH 4D Nucleome through NOFIC. J.-H.S. is supported by a postdoctoral fellowship from the German Research Foundation (DFG, SP1680/1-1). **Author contributions:** W.-K.C., J.-H.S., and I.I.C. conceived of and designed the study; W.-K.C. and J.-H.S. performed experiments and analyzed data with help from M.H., C.Le., and V.G.; M.H. cloned CRISPR repair templates and single-guide RNA plasmids and genotyped cell lines; C.Li conducted and analyzed the Western blot and chromatin immunoprecipitation sequencing (ChIP-seq) assays; W.-K.C., J.-H.S., and I.I.C. wrote the manuscript with input from all coauthors; and I.I.C. supervised all aspects of the project. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data and materials will be provided upon reasonable request to the corresponding author. ChIP-seq datasets generated in this study have been deposited in the Gene Expression Omnibus under accession number GSE115436. Other data described in the text are presented in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6400/412/suppl/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S10
References (*31*–*43*)
Movies S1 to S3

6 November 2017; resubmitted 17 April 2018
Accepted 11 June 2018
Published online 21 June 2018
10.1126/science.aar4199

NEW ENGLAND BioLabs® Inc.

be INSPIRED
drive DISCOVERY
stay GENUINE

Flexible Precision.

NEBNext Direct® Custom Ready Panels for NGS target enrichment

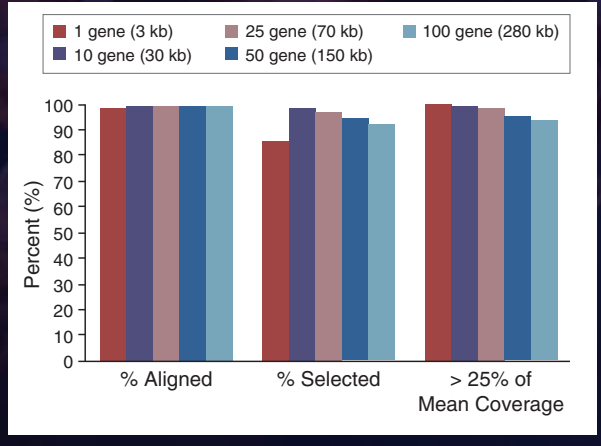
Employing the unique NEBNext Direct hybridization-based enrichment method, NEBNext Direct Custom Ready Panels allow rapid customization of targeted gene panels for Illumina® sequencing. Select from a list of genes for which baits have been carefully designed and optimized to give complete coverage of the full coding regions. High quality panels can be designed by you and rapidly delivered, from any combination of genes. NEBNext Direct Custom Ready Panels provide the content you want with the performance you need.

- Choose from a single gene to hundreds of genes
- Experience unmatched specificity and coverage uniformity
- Eliminate synthesis and optimization steps for faster turnaround
- Improve sensitivity with our Unique Molecule Index (UMI)
- Generate results in one day with our automation-friendly workflow

For more information visit

NEBNextDirect.com

NEBNext Direct Custom Ready Panels demonstrate optimum performance across a wide range of panel sizes



Key target enrichment metrics demonstrate consistent performance across a range of panel sizes. 100 ng of DNA was tested against panels of 1, 10, 25, 50 and 100 genes, and sequenced using Illumina® paired-end 150 bp sequencing. Larger panels included all genes present in smaller panels.

FORCEFORSCIENCE.ORG

STAND TOGETHER
Be a Force for Science

GET THE FACTS

Understand the science behind the issues that matter.

FOLLOW AAAS ADVOCACY

Champion public discussion and evidence-based policy.

TAKE ACTION

Learn ways you can become an advocate and stand up for science.

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

AAAS

Present your work to the world

Are you a representative of the upcoming generation of thought leaders in your field? Together we look forward to your application for the new Sartorius & *Science* Prize for Regenerative Medicine & Cell Therapy.

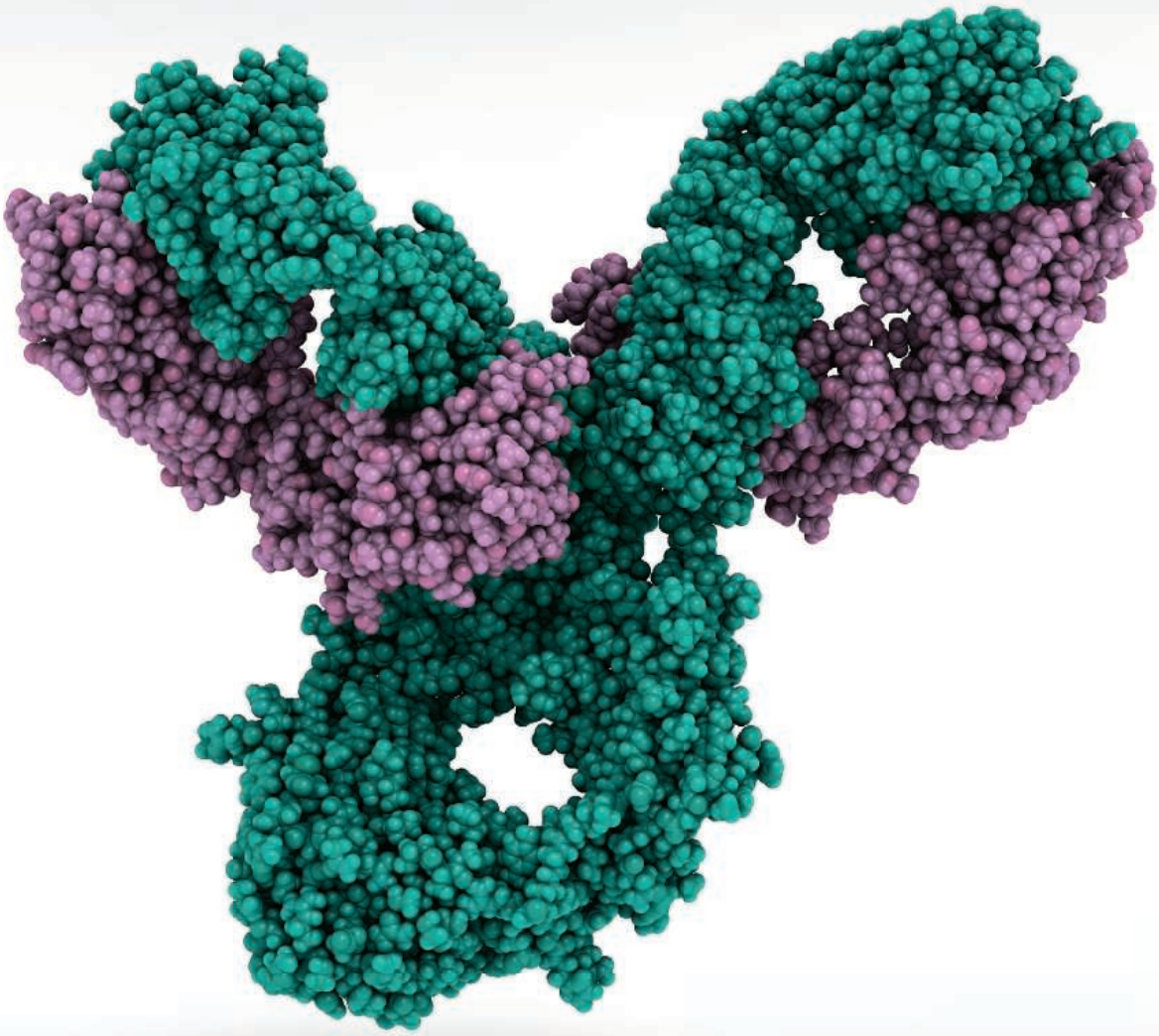
Apply now!

www.passionforscience.com/prize

The Sartorius & *Science* Prize for Regenerative Medicine & Cell Therapy

Awarded by sartorius Science

Publish your research in *Science Immunology*



Science Immunology publishes original, peer-reviewed, science-based research articles that report critical advances in all areas of immunological research, including important new tools and techniques.

For more information: **ScienceImmunology.org**

Science Immunology

AAAS

Clinical Analyzer

Thermo Fisher Scientific's Cascadion SM Clinical Analyzer brings together the ease of use of clinical analyzers with the selectivity and sensitivity of liquid chromatography-tandem mass spectrometry (LC-MS/MS). The fully automated analyzer is designed for use in a variety of settings, including hospital laboratories, and can provide results for a range of clinical tests. It is a turnkey solution that enables clinical labs to easily adopt the power and capabilities of LC-MS/MS as the gold standard in screening and diagnostic testing. The Cascadion system combines assays, software, accessories, consumables, and support/service in a standalone system designed to meet the regulatory requirements for routine and specialized clinical testing.

Thermo Fisher Scientific
For info: 800-955-6288
www.thermofisher.com

RNA Purification Kit

The FastGene Premium RNA kit is developed for rapid, efficient, and clean purification of RNA from tissues and cells for challenging applications such as next-generation sequencing. The kit's silica-based technology dispenses with phenols in favor of superior deoxyribonuclease treatment to ensure the near-complete elimination of genomic DNA. It is the only RNA purification kit available that uses three separate column technologies. This results in extracted RNA of the highest concentration and purity—free from contaminating nucleic acids, proteins, and other biomolecules.

Bulldog Bio
For info: 603-570-4248
www.bulldog-bio.com/fastgene_rna_premium.html

Laboratory Bottles

BrandTech Scientific is pleased to introduce VITgrip laboratory bottles from VITLAB. VITgrip bottles are optimized for storage and sample collection. They are built with an ergonomic slim waist design to aid in handling, double-sided graduations for improved grip and visibility, and tamper-evident screw caps for safe storage. Both the bottle thread and screw cap feature a leakproof design and are made of polypropylene, making them break-resistant and food-safe. The bottles are available in sizes from 125 mL to 2,000 mL.

BrandTech Scientific
For info: 888-522-2726
www.brandtech.com/product/vitgrip-lab-bottles

Purified Exosomes and Dexosomes

System Biosciences' purified exosomes from pooled human biofluid samples come from healthy donors and include exosomes isolated from serum, urine, cerebrospinal fluid, and saliva. Each lot of exosomes is carefully characterized for particle size and concentration by NanoSight analysis, and expression of specific exosome protein markers is validated by Western blot. These exosomes can be used for a variety of applications, including protein biomarker analysis, quantitative PCR for RNA biomarkers, high-throughput biomarker discovery (e.g., mass spectrometry), electron microscopy, and standardized controls for disease studies. Purified exosomes are provided in 1X phosphate buffered saline, and each unit contains 25 µg of exosomal protein. Each vial has enough material to run approximately five lanes in an SDS-PAGE gel (5 µg protein/lane).

System Biosciences
For info: 888-266-5066
www.systembio.com

Non-Ducted Cabinets

Asynt non-ducted filtration and fume cabinets employ extra-large capacity activated carbon filters for removal of fumes and/or high-efficiency particulate air (HEPA) filters for particulate removal. The filters are up to 30% bigger than that of rival systems and also significantly deeper, making fume removal more efficient and effective, thus delivering improved safety. Providing full compliance with Control of Substances Hazardous to Health (COSHH) regulations, the airflow in the cabinets is constantly monitored and the user alerted by audible and visual alarms should the airflow fall below optimum levels. Asynt will warrant units to be free from defects in materials or workmanship for a period of five years from the date of delivery.

Asynt
For info: +44-(0)-1638-781709
www.asynt.com/product/550-filtration-fume-cabinet

Antibody and Protein Labeling Kit

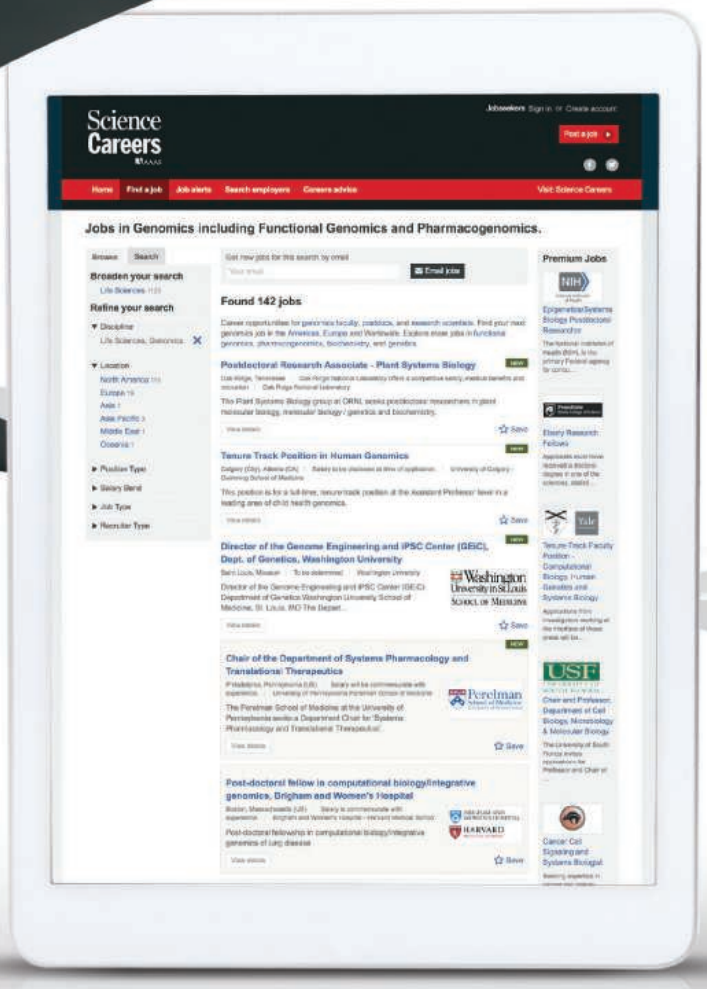
Lightning-Link enables direct labeling of antibodies or proteins for use in R&D applications, drug discovery, and the development of diagnostic kits. The kits require only 30 s hands-on time and there are no separation steps involved, meaning you will retain 100% of your materials. Simply pipette the antibody or other biomolecule into a vial of lyophilized mixture containing the label of interest and incubate for either 3 h (Lightning-Link range) or only 15 min (Lightning-Link Rapid range). Despite its apparent simplicity, the Lightning-Link process is sophisticated and generates conjugates with performance characteristics identical to, or better than, those prepared with laborious multistep conjugation procedures.

Innova Biosciences
For info: 855-466-6821
innovabiosciences.com

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/about/new-products-section for more information.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS



Step up your job search with *Science* Careers

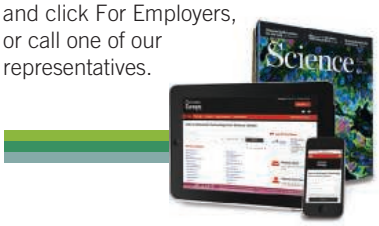
- Access thousands of job postings
- Sign up for job alerts
- Explore career development tools and resources

Search jobs on **ScienceCareers.org** today

Science Careers

SCIENCE CAREERS ADVERTISING

For full advertising details, go to ScienceCcareers.org and click For Employers, or call one of our representatives.



AMERICAS

+1 202 326-6577
+1 202 326-6578
advertise@sciencecareers.org

EUROPE, INDIA, AUSTRALIA, NEW ZEALAND, REST OF WORLD

+44 (0) 1223 326527
advertise@sciencecareers.org

CHINA, KOREA, SINGAPORE, TAIWAN, THAILAND

+86 131 4114 0012
advertise@sciencecareers.org

JAPAN

+81 3-6459-4174
advertise@sciencecareers.org

CUSTOMER SERVICE

AMERICAS

+1 202 326-6577

REST OF WORLD

+44 (0) 1223 326528

advertise@sciencecareers.org

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

ScienceCareers.org

JOB FOCUS: CHEMISTRY

FACULTY POSITION IN NUCLEAR CHEMISTRY AND/OR RADIOCHEMISTRY at the University of California, Irvine

The Department of Chemistry at the University of California, Irvine invites applications from outstanding individuals for a tenure-track position at the **Assistant Professor** level in the broad field of **Nuclear Chemistry and/or Radiochemistry**. Candidates must have a PhD in Chemistry or a related field; postdoctoral experience is desirable. The position requires the establishment of a vigorous research program involving any aspect of Nuclear Chemistry or Radiochemistry, development of ideas for utilization of the department's nuclear reactor facility, and a strong commitment to excellence in teaching areas of general chemistry and radiochemistry at both the undergraduate and graduate levels. Applications must be submitted electronically via the Internet at website: **http://recruit.ap.uci.edu/apply/JPF04764**. Applicants should upload a cover letter, a curriculum vitae (including a publication list), a concise statement of proposed research and a teaching statement. A separate statement that addresses past and/or potential contributions to diversity, equity and inclusion should also be included in the application materials. At least three letters of recommendation are required. Applications and supporting materials should be received by **September 30, 2018** for full consideration. *The University of California, Irvine is an Equal Opportunity/Affirmative Action Employer advancing inclusive excellence. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability, age, protected veteran status, or other protected categories covered by the UC nondiscrimination policy.*

POSITIONS OPEN

Postdoctoral Positions Helen Diller Family Comprehensive Cancer Center University of California San Francisco

Two post-doctoral Fellow positions are available immediately to work in the group of Dr. Allan Balmain, based in the Helen Diller Family Comprehensive Cancer Center at UCSF. One position will focus on targeting of cancer stem cells, and the other on development of new strategies for cancer immunotherapy. Preference will be given to applicants with strong experience in at least one of the following areas: normal and cancer stem biology, tumor immune responses, computational biology and bioinformatics analysis of gene expression. Candidates should have strong genetic and molecular/cell biology skills, preferably with some experience in con-focal imaging and FACS sorting of fluorescent cell populations from whole tissues and tumors. Highly developed communication skills are essential.

The work is based in the purpose-built UCSF, Helen Diller Family Cancer Research Building on the new Mission Bay campus in San Francisco. UCSF has an active post-doctoral training program with many benefits (http://postdocs.ucsf.edu/).

More information can be found by email: **allan.balmain@ucsf.edu**, or Nora Bayani; **nora.bayani@ucsf.edu**.

Interested applicants should email their curriculum vitae, brief statement of research interest, and names of three academic references to email: **nora.bayani@ucsf.edu**

The University of California is an Equal Opportunity Employer, dedicated to excellence through diversity.

UC San Francisco seeks candidates whose experience, teaching, research, or community service has prepared them to contribute to our commitment to diversity and excellence.

OIST OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY GRADUATE UNIVERSITY
沖縄科学技術大学院大学

Open Faculty Positions

Okinawa Institute of Science and Technology Graduate University (www.oist.jp) in Japan invites applications for new faculty positions as part of its planned expansion.

Targeted areas:

- **Ecology, Evolution, and Environmental Sciences**
- **Marine Science**
- **Mathematics**
- **Neuroscience**
- **Computer Science**
- **Life Sciences**
- **Engineering and Applied Science**
- **Physics and Materials Science**
- **Chemistry**

Excellent scholarship and creativity are the main criteria. Successful candidates are expected to establish an active and independent research program, supervise Ph.D. students, teach graduate courses, and perform university services as required. Generous research resources will be provided, which may be supplemented with external grants. Appointments will be Tenure-Track or Tenured. Starting date is flexible.

Suitably qualified applicants for other fields of science and technology should also apply. These applicants are requested to apply through the Marine Science portal, regardless of subject area.

Application documents:

1) Cover letter
2) CV and publication list
3) Research statement
4) Teaching statement
5) Three contact references

Information and instructions regarding application submission and deadline:
https://groups.oist.jp/facultypositions

OIST offers a world-class research environment with an international research community and opportunities for interdisciplinary research. Research and teaching are conducted in English. The campus is situated in a beautiful subtropical setting in Okinawa, Japan. See the video "University of the Future": https://www.oist.jp/news-center/videos/university-future

OIST Graduate University is an equal opportunity educator and employer that is actively working to increase the diversity of its faculty, students and staff. OIST has implemented policies designed to promote a culture of diversity and inclusion and a family friendly working environment. These include considerations for hiring and supporting dual career couples, "stop the clock" options for new parents, an on site child care facility for children from 2 months to 6 years of age, and after-school and holiday programs for K-6 children. The University strongly encourages applications from women and other underrepresented groups.

Inquiries:
Dr. Milind Purohit
Dean of Faculty Affairs
faculty-recruiting@oist.jp

Career Feature:

Postdoc Careers

Issue date: August 31

Book ad by August 16

Ads accepted until August 24 if space allows

129,562
subscribers in print every week

503,472
monthly unique browsers on ScienceCareers.org

56 %
of our weekly readers are Ph.D.s

Once their position ends, postdocs come to a fork in the road—to stay or leave research. This special must-read feature focuses on a path less traveled—careers in career development. Some work in Postdoc Affairs or Graduate Affairs divisions and departments or Careers Services organizations. If you're hiring, be sure to promote your job opportunities to *Science*'s highly qualified readership of grad students, postdocs and researchers.

What makes *Science* the best choice for recruiting?

- Read and respected by 400,000 readers around the globe
- 80% of readers read *Science* more often than any other journal
- Your ad dollars support AAAS and its programs, which strengthens the global scientific community.

Why choose this Postdoc Feature for your advertisement?

- Relevant ads lead off the career section with a special "Postdoc" banner
- Link on the job board homepage directly to postdoc positions
- Dedicated landing page for postdoc positions.

To book your ad:
advertise@sciencecareers.org

The Americas
+ 202 326 6577

Europe
+44 (0) 1223 326527

Japan
+81 3 6459 4174

China/Korea/Singapore/ Taiwan
+86 131 4114 0012

Produced by the *Science*/AAAS Custom Publishing Office.

SCIENCECAREERS.ORG

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

FOR RECRUITMENT IN SCIENCE, THERE'S ONLY ONE *SCIENCE*.

Department of Pharmacology
Yale University

Assistant or Associate Professor of Pharmacology

The Department of Pharmacology at Yale University School of Medicine invites applications for Assistant or Associate Professor positions in the tenure track. Applicants must have an M.D. and/or Ph.D. or equivalent degree and have demonstrated excellent qualifications in research and education. Areas of interest include studies of molecular function such as proteomics, chemical biology, structural biology (including electron microscopy), and computational biology in different areas of drug discovery, cell signaling, metabolism, genetic diseases, and epigenetic mechanisms.

Responsibilities include establishing a vigorous and independently funded research program in cancer biology while supervising and mentoring students with diverse backgrounds plus contributing to the graduate and medical school educational missions. We seek individuals with strong records of independent creative accomplishments, who will interact productively with colleagues within the Department and across Yale School of Medicine.

Pharmacology at Yale University has a rich history of pre-eminence in drug discovery, indeed the first cancer chemotherapy clinical trials were conducted here. Today, the research in our Department encompasses a broad range of fields, with researchers aiming to both understand and treat human disease, with particular strengths in drug discovery, integrative cell signaling, structural biology, and neuroscience. For more details, see https://medicine.yale.edu/pharm/

Please submit a letter describing qualifications, along with a CV, a two-page summary of current and proposed research, and three letters of reference to the following Interfolio website: **http://apply.interfolio.com/52275**. Informal inquiries may be submitted electronically to Prof. Joseph Schlessinger, Chair of Pharmacology, c/o Nathan Kucera (nathan.kucera@yale.edu). Consideration of applications will begin **October 1st, 2018**.

Yale University is an Equal Opportunity/Affirmative Action Employer. We seek candidates who embrace and reflect diversity in the broadest sense. Yale values diversity among its students, staff, and faculty and strongly welcomes applications from women, persons with disabilities, protected veterans, and underrepresented minorities.

Jefferson Science Fellowship

USAID
FROM THE AMERICAN PEOPLE

The National Academies of Sciences, Engineering, and Medicine is pleased to announce a call for applications for the 2019 Jefferson Science Fellows (JSF) program. Initiated by the Secretary of State in 2003, this fellowship program engages the American academic science, technology, engineering and medical communities in the design and implementation of U.S. foreign policy.

Jefferson Science Fellows spend one year at the U.S. Department of State or the U.S. Agency for International Development (USAID) for an on-site assignment in Washington, D.C. that may also involve extended stays at U.S. foreign embassies and/or missions.

The fellowship is open to tenured, or similarly ranked, academic scientists, engineers, and physicians from U.S. institutions of higher learning. Applicants must hold U.S. citizenship and will be required to obtain a security clearance.

The deadline for applications for the 2019-2020 program year is **October 31, 2018**. To learn more about the Jefferson Science Fellows program and to apply, visit **www.nas.edu/jsf**.

The Jefferson Science Fellows program is administered by the National Academies of Sciences, Engineering, and Medicine and supported by the U.S. Department of State and the United States Agency for International Development.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Cleveland Clinic

Assistant or Associate Professor
Genitourinary Malignancy Research Center

The newly formed Genitourinary Malignancy Research Center at Cleveland Clinic is seeking to recruit a physician-scientist or scientist with a primary interest in prostate, kidney or bladder cancer. The position presents an exceptional opportunity to partner with outstanding clinical urology and oncology programs in a new center where a principal goal is to translate basic discoveries rapidly to the clinic. Cleveland Clinic is ranked the #1 hospital for urology and #2 hospital overall by US News and World Report. The overarching goal of the Center is to define the underlying biochemical and clinical behavior of prostate, kidney, and bladder cancer in order to identify new strategies to prevent symptoms and death.

Applicants must have an M.D., M.D./Ph.D., or Ph.D., an exceptional track record of accomplishment, and be able to function well as part of a multidisciplinary team. The successful applicant will be supported by generous start-up funds and joint appointments in Cleveland Clinic's Lerner Research Institute, Glickman Urological and Kidney Institute and Taussig Cancer Institute (part of the NCI-designated Case Comprehensive Cancer Center).

Candidates should submit a CV, summary of research interests, and three references, via e-mail to Nima Sharifi, M.D. c/o Kali Casale, **casalek@ccf.org**

Cleveland Clinic is pleased to be an Equal Employment/Affirmative Action Employer: Women/Minorities/Veterans/Individuals with Disabilities. Smoke/drug-free environment.

WORKING LIFE

By **Vanessa Sung**

Let your stars shine

A few months ago, one of my fellow graduate students was offered an incredible career-building opportunity. His efforts on a student-led campaign to increase federal funding for basic research had led to an invitation to speak on a conference panel about health research advocacy. He would have a prominent platform to share his work with a large audience, and he would get to meet and network with leaders in the field. It would look great on his CV. I was happy for him, of course; he deserved it. But my feelings were complicated. I had been an equal partner in the campaign. I felt that I deserved the same opportunity. Yet I hadn't been invited, and I wasn't sure why.

Not long ago, I wouldn't have done anything about it. Growing up, I had been taught that the way to success was to put my head down and work hard. If my work was good enough, it would speak for itself and I would be rewarded. This, coupled with a natural shyness, meant that I rarely worked up the nerve to put myself forward.

Then last year happened. My long-term relationship was in shambles. I was in the midst of a significant health scare and battling doubts about where my career was headed. When the dust settled, I decided that I needed to take control of my life again—by becoming my own biggest advocate.

"I can't be recognized or rewarded for my work if I'm invisible."

So, with my heart pounding, I asked my colleague to ask the panel organizers to include me. My mind flooded with a litany of well-practiced self-doubt. Maybe I didn't deserve it. Maybe I was being too ambitious. But my colleague immediately contacted the organizers, who quickly agreed to add me to the panel.

I've come to understand that I can't be recognized or rewarded for my work if I'm invisible. This isn't a revolutionary idea, I know, but as an introvert this is my personal Everest. The biggest lesson I've learned is simply to ask for what you want. Ask to be nominated for an award. Ask to give a guest lecture. Ask for a reference letter from that person you think is beyond your reach. The worst that can happen is that the answer is no, but I've found that people are generally happy to oblige.

I am also working to get more comfortable promoting my accomplishments. A couple of years ago, when a mentor agreed to write a recommendation letter for me but asked me to provide a draft as a starting point, I agonized over it. Writing nice things about myself felt completely unnatural. With every complimentary statement I committed to paper came mounting anxiety that my mentor would disapprove of my lack of humility. I never got any feedback about that letter, so I don't know for sure what my mentor thought of it. Nonetheless, I learned an important lesson: I can't control what other people think, so I might as well share my accomplishments as best I can. These days, whenever I ask for a recommendation letter, I offer to provide a draft. I know best how to highlight my skills and accomplishments for a particular application, and it doesn't hurt to offer.

I've also embraced social media as a way to share my accomplishments more widely. I was hesitant at first, worried that I might cross the line from self-promotion to bragging. But when I asked myself whether I was turned off by other scientists sharing their publications, awards, projects, and achievements, the answer was clear: nope. In fact, their self-promotion helped me discover amazing work and find new professional connections. Last year, for example, I was intrigued to see two young female scientists promoting their new podcast on Twitter. I commented on their post, suggesting science policy as an episode topic, and they ended up interviewing me on their show! This unique opportunity would never have materialized without a little online self-promotion on both our ends.

I think advocating for myself will always make me a little uncomfortable. But I do plenty of other uncomfortable things to ensure I'm living my best life, like getting flu shots and (ugh) going to the gym. I suppose I can do the same for my career. ■

Vanessa Sung is a Ph.D. candidate at McGill University in Montreal, Canada, and co-president of Science & Policy Exchange. Do you have an interesting career story? Send it to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER