

What it takes to end an
AIDS epidemic p. 226

Polar bears suffer through
lean summers p. 295

Sperm produced in ovary
of mutant fish p. 328

Science

\$10
17 JULY 2015
sciencemag.org

AAAS

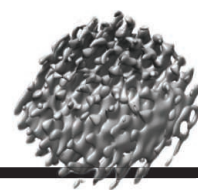
SPECIAL ISSUE

ARTIFICIAL INTELLIGENCE



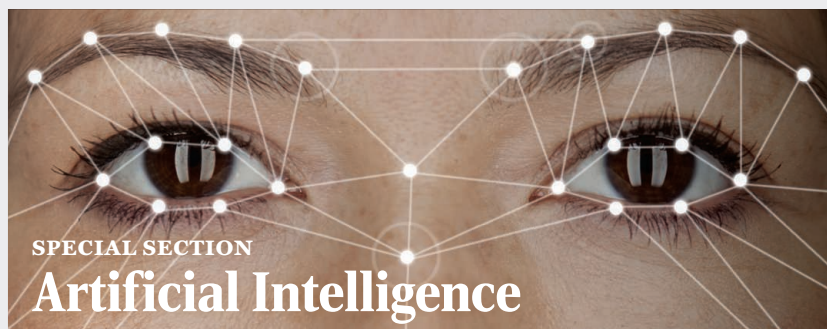
CONTENTS

17 JULY 2015 • VOLUME 349 • ISSUE 6245



232 & 290

Visualizing a
moving target



SPECIAL SECTION

Artificial Intelligence

INTRODUCTION

248 Rise of the machines

NEWS

250 The synthetic therapist

By J. Bohannon

252 Fears of an AI pioneer *By J. Bohannon*

POLICY FORUM

253 Data, privacy, and the greater good
By E. Horvitz and D. Mulligan

REVIEWS

255 Machine learning: Trends, perspectives, and prospects *M. I. Jordan and T. M. Mitchell*

261 Advances in natural language processing *J. Hirschberg and C. D. Manning*

267 Economic reasoning and artificial intelligence *D. C. Parkes and M. P. Wellman*

273 Computational rationality: A converging paradigm for intelligence in brains, minds, and machines
S. J. Gershman et al.

SEE ALSO ► BOOKS ET AL. P. 243 ► PODCAST

ON THE COVER



Intelligence is hard to define, but you know it when you see it ... Or do you? Artificial intelligence researchers can now design algorithms with almost humanlike abilities to perceive images, communicate with language, and learn from experience. Can we learn anything about how our neuron-based minds work from these machines? Do we need to worry about what these algorithmic minds might be learning about us? On the cover is a visualization of human brain connectivity from MRI diffusion imaging, with superimposed computer connectors. See page 248. *Illustration: Beth Rakouskas, imaging courtesy of Arthur W. Toga, USC Laboratory of Neuro Imaging; computer connectors: Hans-Joachim Roy/Shutterstock*

NEWS

IN BRIEF

218 Roundup of the week's news

IN DEPTH

221 TORTURE REPORT PROMPTS APA APOLOGY

Admitting it colluded with U.S., psychologists group to change policies, leadership *By J. Bohannon*

222 RESEARCHERS SEEK CLEAR REASONS WHEN CLINICAL TRIALS END EARLY

Explanations are often hazy
By J. Couzin-Frankel

223 REPORT PRESCRIBES STRONG MEDICINE FOR WHO

Ebola failures show that difficult reforms are needed *By K. Kupferschmidt*

224 RUSSIA TARGETS WESTERN TIES

Crackdown on "foreign agents" and "undesirable" groups threatens private support for science *By V. Pokrovsky*

225 DATA CHECK: SALARIES PUMP UP BIOMEDICAL INFLATION

By J. Mervis

FEATURES

226 MEANS TO AN END

Cities, states, and provinces are gearing up to halt their AIDS epidemics—though the definition of success varies *By J. Cohen*

..... **230** No end in sight

By J. Cohen

INSIGHTS

PERSPECTIVES

232 TRACKING THE MERRY DANCE OF NANOPARTICLES

Electron microscopy provides atomic-resolution structures of nanoparticles in solution *By C. Collier*

► REPORT P. 290

234 MAGNETIC BUBBLES WITH A TWIST

Individual skyrmionic bubbles can be generated and moved at room temperature *By K. von Bergmann*

► RESEARCH ARTICLE P. 283

235 IS BIODIVERSITY GOOD FOR YOUR HEALTH?

Disease incidence is often lower in more diverse communities of plants and animals *By F. Keesing and R. S. Ostfeld*

237 NEUTROPHIL-MACROPHAGE COMMUNICATION IN INFLAMMATION AND ATHEROSCLEROSIS

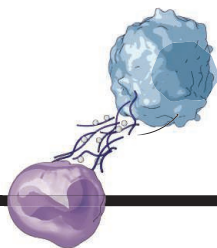
Neutrophils may license macrophages to respond to cholesterol crystals and drive inflammation that aggravates atherosclerosis *By M. Nahrendorf and F. K. Swirski*

► REPORT P. 316

238 METRICS FOR LAND-SCARCE AGRICULTURE

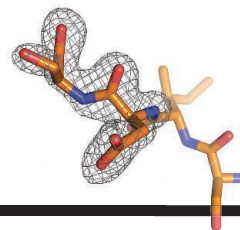
Nutrient content must be better integrated into planning
By R. DeFries et al.

Science Staff	214
New Products	333
Science Careers	334



237 & 316

Neutrophils prime
atherosclerosis



312 & 324

Cyanobacterial
circadian pacemaker

241 LIVING SUPRAMOLECULAR POLYMERIZATION

Greater control is achieved over the chain growth and properties of dynamic materials

By R. D. Mukhopadhyay and A. Ajayaghosh

BOOKS ET AL.

243 EX MACHINA

A. Garland, director, reviewed by R. W. Picard

► ARTIFICIAL INTELLIGENCE SECTION P. 248

244 FROM FIELD TO FORK

By P. B. Thompson,
reviewed by N. Freudenberg

LETTERS

246 HOLOCENE AS ANTHROPOCENE

By G. Certini and R. Scalenghe

246 GEOLOGICAL EVIDENCE FOR THE ANTHROPOCENE

By S. L. Lewis and M. A. Maslin

247 RESPONSE

By W. F. Ruddiman

247 TECHNICAL COMMENT ABSTRACTS

247 ERRATA

RESEARCH

IN BRIEF

279 From *Science* and other journals

RESEARCH ARTICLES

282 CHROMOSOMES

A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation A. Minajigi et al.

RESEARCH ARTICLE SUMMARY; FOR FULL TEXT:

dx.doi.org/10.1126/science.aab2276



283 MAGNETISM

Blowing magnetic skyrmion bubbles W. Jiang et al.

► PERSPECTIVE P. 234

REPORTS

287 HEAVY FERMIONS

Unconventional Fermi surface in an insulating state B. S. Tan et al.

290 NANOPARTICLE IMAGING

3D structure of individual nanocrystals in solution by electron microscopy J. Park et al.

► PERSPECTIVE P. 232

295 ANIMAL PHYSIOLOGY

Summer declines in activity and body temperature offer polar bears limited energy savings J. P. Whiteman et al.

298 THERMAL PHYSIOLOGY

Keeping cool: Enhanced optical reflection and radiative heat dissipation in Saharan silver ants N. N. Shi et al.

302 PLANT ECOLOGY

Worldwide evidence of a unimodal relationship between productivity and plant species richness L. H. Fraser et al.

305 ICE SHEETS

Reverse glacier motion during iceberg calving and the cause of glacial earthquakes T. Murray et al.

309 PLANT SCIENCE

Morphinan biosynthesis in opium poppy requires a P450-oxidoreductase fusion protein T. Winzer et al.

312 CIRCADIAN RHYTHMS

Atomic-scale origins of slowness in the cyanobacterial circadian clock J. Abe et al.

► REPORT P. 324

316 INFLAMMATION

Neutrophil extracellular traps license macrophages for cytokine production in atherosclerosis A. Warnatsch et al.

► PERSPECTIVE P. 237

320 HIV-1 VACCINES

Protective efficacy of adenovirus/protein vaccines against SIV challenges in rhesus monkeys D. H. Barouch et al.

324 CIRCADIAN RHYTHMS

A protein fold switch joins the circadian oscillator to clock output in cyanobacteria Y.-G. Chang et al.

► REPORT P. 312

328 SEX DETERMINATION

foxl3 is a germ cell-intrinsic factor involved in sperm-egg fate decision in medaka T. Nishimura et al.



DEPARTMENTS

217 EDITORIAL

Passion is just the start
By Marcia McNutt

338 WORKING LIFE

The space roboticist
By Vijaysree Venkatraman

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2015 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership and subscription (51 issues): \$153 (\$74 allocated to subscription). Domestic institutional subscription (51 issues): \$1282. Foreign postage extra: Mexico, Caribbean (surface mail) \$55; other countries (air assist delivery) \$85. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request. GST #R1254 88122. 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: \$10.00 current issue, \$15.00 back issue prepaid includes surface postage; bulk rates on request. Authorization to photocopy 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 other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that \$30.00 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.

Editor-in-Chief Marcia McNutt

Executive Editor Monica M. Bradford **News Editor** Tim Appenzeller

Managing Editor, Research Journals Katrina L. Kelner

Deputy Editors Barbara R. Jasny, Andrew M. Sugden(UK), Valda J. Vinson, Jake S. Yeston

Research and Insights

SR. EDITORS Caroline Ash(UK), Gilbert J. Chin, Lisa D. Chong, Julia Fahrenkamp-Uppenbrink(UK), Pamela J. Hines, Stella M. Hurlty(UK), Paula A. Kiberstis, Marc S. Lavine(Canada), Kristen L. Mueller, Ian S. Osborne(UK), Beverly A. Purnell, L. Bryan Ray, Guy Riddihough, H. Jesse Smith, Jelena Stajic, Peter Stern(UK), Phillip D. Szurmi, Brad Wible, Nicholas S. Wigginton, Laura M. Zahn **ASSOCIATE EDITORS** Brent Grocholski, Keith T. Smith, Sacha Vignieri **ASSOCIATE BOOK REVIEW EDITOR** Valerie B. Thompson **ASSOCIATE LETTERS EDITOR** Jennifer Sills **CHIEF CONTENT PRODUCTION EDITOR** Cara Tate **SR. CONTENT PRODUCTION EDITORS** Harry Jack **CONTENT PRODUCTION EDITORS** Jeffrey E. Cook, Chris Filiatreau, Cynthia Howe, Lauren Kmcac, Barbara P. Ordway, Catherine Wolner **SR. EDITORIAL COORDINATORS** Carolyn Kyle, Beverly Shields **EDITORIAL COORDINATORS** Ramatoulaye Diop, Joi S. Granger, Lisa Johnson, Anita Wynn **PUBLICATIONS ASSISTANTS** Aneera Dobbins, Jeffrey Hearn, Dona Mathieu, Le-Toya Mayne Flood, Shannon McMahon, Scott Miller, Jerry Richardson, Rachel Roberts(UK), Alice Whaley(UK), Brian White **EXECUTIVE ASSISTANT** Anna Bashkova **ADMINISTRATIVE SUPPORT** Janet Clements(UK), Lizanne Newton(UK), Maryrose Madrid, Laura-Nadine Schuhmacher (UK, Intern), Alix Welch (Intern), John Wood(UK)

News

NEWS MANAGING EDITOR John Travis **INTERNATIONAL EDITOR** Richard Stone **DEPUTY NEWS EDITORS** Daniel Clery(UK), Robert Coontz, Elizabeth Culotta, David Grimm, David Malakoff, Leslie Roberts **CONTRIBUTING EDITOR** Martin Enserink(Europe) **SR. CORRESPONDENTS** Jeffrey Mervis, Elizabeth Pennisi **NEWS WRITERS** Adrian Cho, Jon Cohen, Jennifer Couzin-Frankel, Carolyn Gramling, Eric Hand, Jocelyn Kaiser, Catherine Matacic, Kelly Servick, Robert F. Service, Erik Stokstad(Cambridge, UK), Emily Underwood **INTERNS** Emily Conover, Emily DeMarco, Annick Laurent, Laura Oliveri, Juan David Romero **CONTRIBUTING CORRESPONDENTS** Michael Balter(Paris), John Bohannon, Ann Gibbons, Mara Hivstendahl, Sam Kean, Richard A. Kerr, Eli Kintisch, Kai Kupferschmidt(Berlin), Andrew Lawler, Christina Larson(Beijing), Mitch Leslie, Charles C. Mann, Eliot Marshall, Virginia Morell, Dennis Normile(Tokyo), Heather Pringle, Tania Rabesandratana(London), Gretchen Vogel(Berlin), Lizzie Wade(Mexico City) **CAREERS** Jim Austin(Editor), Donisha Adams, Rachel Bernstein **CUPO EDITORS** Kara Estelle (Chief), Julia Cole, Jennifer Levin **ADMINISTRATIVE SUPPORT** Jessica Williams

Executive Publisher Rush D. Holt

Publisher Kent R. Anderson **Chief Digital Media Officer** Rob Covey

BUSINESS OPERATIONS AND PORTFOLIO MANAGEMENT DIRECTOR Sarah Whalen **BUSINESS SYSTEMS AND FINANCIAL ANALYSIS DIRECTOR** Randy Yi **MANAGER OF FULFILLMENT SYSTEMS** Neal Hawkins **SYSTEMS ANALYST** Nicole Mehmedovic **ASSISTANT DIRECTOR, BUSINESS OPERATIONS** Eric Knott **MANAGER, BUSINESS OPERATIONS** Jessica Tierney **BUSINESS ANALYSTS** Cory Lipman, Cooper Tilton, Celeste Troxler **FINANCIAL ANALYST** Robert Clark **RIGHTS AND PERMISSIONS ASSISTANT DIRECTOR** Emilie David **PERMISSIONS ASSOCIATE** Elizabeth Sandler **RIGHTS, CONTRACTS, AND LICENSING ASSOCIATE** Lili Kiser

MARKETING DIRECTOR Ian King **MARKETING MANAGER** Julianne Wielga **MARKETING ASSOCIATE** Elizabeth Sattler **SR. MARKETING EXECUTIVE** Jennifer Reeves **SR. ART ASSOCIATE, PROJECT MANAGER** Tzeitel Sorrosor **ART ASSOCIATE** Seil Lee **SR. ART ASSOCIATE** Kim Huynh **ASSISTANT COMMERCIAL EDITOR** Selby Frame **MARKETING PROJECT MANAGER** Angelissa McArthur **PROGRAM DIRECTOR, AAAS MEMBER CENTRAL** Peggy Mihelich **FULFILLMENT SYSTEMS AND OPERATIONS** membership@aaas.org **MANAGER, MEMBER SERVICES** Pat Butler **SPECIALISTS** LaToya Casteel, Terrance Morrison, Latasha Russell **MANAGER, DATA ENTRY** Mickie Napoleoni **DATA ENTRY SPECIALISTS** JJ Regan, Brenden Aquilino, Fiona Giblin

DIRECTOR, SITE LICENSING Tom Ryan **DIRECTOR, CORPORATE RELATIONS** Eileen Bernadette Moran **SR. PUBLISHER RELATIONS SPECIALIST** Kiki Forsythe **PUBLISHER RELATIONS MANAGER** Catherine Holland **PUBLISHER RELATIONS, EASTERN REGION** Keith Layson **PUBLISHER RELATIONS, WESTERN REGION** Ryan Rexroth **SALES RESEARCH COORDINATOR** Aiesha Marshall **MANAGER, SITE LICENSE OPERATIONS** Iquo Edim **SENIOR PRODUCTION SPECIALIST** Robert Koeppke **SENIOR OPERATIONS ANALYST** Lana Guz **FULFILLMENT** Judy Lillibridge **ASSOCIATE DIRECTOR, MARKETING** Christina Schlecht **MARKETING ASSOCIATES** Thomas Landreth, Isa Sesay-Bah

DIRECTOR OF WEB TECHNOLOGIES Ahmed Khadr **SR. DEVELOPER** Chris Coleman **DEVELOPERS** Dan Berger, Jimmy Marks **SR. PROJECT MANAGER** Trista Smith **SYSTEMS ENGINEER** Luke Johnson

CREATIVE DIRECTOR, MULTIMEDIA Martyn Green **DIRECTOR OF ANALYTICS** Enrique Gonzales **SR. WEB PRODUCER** Sarah Crespi **WEB PRODUCER** Alison Crawford **VIDEO PRODUCER** Nguyen Nguyen **SOCIAL MEDIA PRODUCER** Meghna Sachdev

DIRECTOR OF OPERATIONS PRINT AND ONLINE Elizabeth Harman **DIGITAL/PRINT STRATEGY MANAGER** Jason Hillman **QUALITY TECHNICAL MANAGER** Marcus Spiegel **DIGITAL PRODUCTION MANAGER** Lisa Stanford **ASSISTANT MANAGER DIGITAL/PRINT** Rebecca Doshi **DIGITAL MEDIA SPECIALIST** Tara Kelly **SENIOR CONTENT SPECIALISTS** Steve Forrester, Antoinette Hodal, Lori Murphy, Anthony Rosen **CONTENT SPECIALISTS** Jacob Hedrick, Kimberley Oster

DESIGN DIRECTOR Beth Rakouskas **DESIGN EDITOR** Marcy Atarod **SENIOR SCIENTIFIC ILLUSTRATORS** Chris Bickel, Katharine Suttiff **SCIENTIFIC ILLUSTRATOR** Valerie Altounian **SENIOR ART ASSOCIATES** Holly Bishop, Preston Huey **SENIOR DESIGNER** Garvin Grullón **DESIGNER** Chrystal Smith **SENIOR PHOTO EDITOR** William Douthitt **PHOTO EDITORS** Leslie Blizard, Christy Steele

DIRECTOR, GLOBAL COLLABORATION, CUSTOM PUBLICATIONS, ADVERTISING Bill Moran **EDITOR, CUSTOM PUBLISHING** Sean Sanders: 202-326-6430 **ASSISTANT EDITOR, CUSTOM PUBLISHING** Tianna Hicklin: 202-326-6463 **ADVERTISING MARKETING MANAGER** Justin Sawyers: 202-326-7061 **science_advertising@aaas.org** **ADVERTISING MARKETING ASSOCIATE** Javia Flemmings **ADVERTISING SUPPORT MANAGER** Karen Foote: 202-326-6740 **ADVERTISING PRODUCTION OPERATIONS MANAGER** Deborah Tompkins **SR. PRODUCTION SPECIALIST/GRAPHIC DESIGNER** Amy Hardcastle **PRODUCTION SPECIALIST** Yuse Lajimimuhup **SR. TRAFFIC ASSOCIATE** Christine Hall **SALES COORDINATOR** Shirley Young **ASSOCIATE DIRECTOR, COLLABORATION, CUSTOM PUBLICATIONS/CHINA/TAIWAN/KOREA/SINGAPORE** Ruolei Wu: +86-186 0822 9345, rwu@aaas.org **COLLABORATION/ CUSTOM PUBLICATIONS/JAPAN** Adarsh Sandhu + 81532-81-5142 asandhu@aaas.org **EAST COAST/E. CANADA** Laurie Faraday: 508-747-9395, FAX 617-507-8189 **WEST COAST/W. CANADA** Lynne Stickrod: 415-931-9782, FAX 415-520-6940 **MIDWEST** Jeffrey Dembski: 847-498-4520 x3005, Steven Loerch: 847-498-4520 x3006 **CUPO/ASIA** Roger Gonçalves: TEL/FAX +41 43 243 1358 **JAPAN** Katsuyoshi Fukamizu(Tokyo): +81-3-3219-5777 kfukamizu@aaas.org **CHINA/TAIWAN** Ruolei Wu: +86-186 0822 9345, rwu@aaas.org

WORLDWIDE ASSOCIATE DIRECTOR OF SCIENCE CAREERS Tracy Holmes: +44 (0) 1223 326525, FAX +44 (0) 1223 326532 tholmes@science-int.co.uk **CLASSIFIED** advertise@sciencecareers.org **U.S. SALES** Tina Burks: 202-326-6577 **Nancy Toema**: 202-326-6578 **SALES ADMINISTRATOR** Marci Gallun **EUROPE/ROW SALES** Axel Gesatzki, Sarah Lelange **SALES ASSISTANT** Kelly Grace **JAPAN** Hiroyuki Mashiki(Kyoto): +81-75-823-1109 hmashiki@aaas.org **CHINA/TAIWAN** Ruolei Wu: +86-186 0082 9345 rwu@aaas.org **MARKETING MANAGER** Allison Pritchard **MARKETING ASSOCIATE** Aimee Aponte

AAAS BOARD OF DIRECTORS **RETIRING PRESIDENT, CHAIR** Gerald R. Fink **PRESIDENT** Geraldine (Geri) Richmond **PRESIDENT-ELECT** Barbara A. Schaaf **TREASURER** David Evans **SHAW CHIEF EXECUTIVE OFFICER** Rush D. Holt **BOARD** Bonnie L. Bassler, May R. Berenbaum, Carlos J. Bustamante, Stephen P. A. Fodor, Claire M. Fraser, Michael S. Gazzaniga, Laura H. Greene, Elizabeth Loftus, Mercedes Pascual

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 202-326-6755 **REPRINTS:** Author Inquiries 800-635-7181 **COMMERCIAL INQUIRIES** 803-359-4578 **PERMISSIONS** 202-326-6765, permissions@aaas.org **AAAS Member Services** 202-326-6417 or http://membercentral.aaas.org/discounts

Science serves as a forum for discussion of important issues related to the advancement of science by publishing material on which a consensus has been reached as well as including the presentation of minority of conflicting points of view. Accordingly, all articles published in Science—including editorials, news and comment, and books reviews—are signed and reflect the individual views of the authors and not official points of view adopted by AAAS or the institutions with which the authors are affiliated.

INFORMATION FOR AUTHORS See pages 678 and 679 of the 6 February 2015 issue or access www.sciencemag.org/about/authors

SENIOR EDITORIAL BOARD

Robert H. Grubbs, *California Institute of Technology*, Gary King, *Harvard University*
Susan M. Rosenberg, *Baylor College of Medicine*, Ali Shalithard, *Northwestern University*
Feinberg School of Medicine, Michael S. Turner, *U. of Chicago*

BOARD OF REVIEWING EDITORS (Statistics board members indicated with \$)

Adriano Aguzzi, *U. Hospital Zürich*
Takuzo Aida, *U. of Tokyo*
Leslie Aiello, *Wenner-Gren Foundation*
Judith Allen, *U. of Edinburgh*
Sonia Altizer, *U. of Georgia*
Sebastian Amigorena, *Institut Curie*
Kathryn Anderson, *Memorial Sloan-Kettering Cancer Center*
Meinrat O. Andreae, *Max-Planck Inst. Mainz*
Paola Arlotta, *Harvard U.*
Johan Auwerx, *EPFL*
David Awschalom, *U. of Chicago*
Jordi Bascompte, *Estación Biológica de Doñana CSIC*
Facundo Batista, *London Research Inst.*
Ray H. Baughman, *U. of Texas, Dallas*
David Baum, *U. of Wisconsin*
Carlo Beenakker, *Leiden U.*
Kamran Behnia, *ESPCI-ParisTech*
Yasmine Belkaid, *NIH/NIH*
Philip Benfey, *Duke U.*
Stephen J. Benkovic, *Penn State U.*
May Berenbaum, *U. of Illinois*
Gabriele Bergers, *U. of California, San Francisco*
Bradley Bernstein, *Massachusetts General Hospital*
Peer Bork, *EMBL*
Bernard Bourdon, *Ecole Normale Supérieure de Lyon*
Chris Bowler, *Ecole Normale Supérieure*
Ian Boyd, *U. of St. Andrews*
Emily Brodsky, *U. of California, Santa Cruz*
Ron Brookmeyer, *U. of California Los Angeles (\$)*
Christian Büchel, *Hamburg-Eppendorf*
Joseph A. Burns, *Cornell U.*
Gyorgy Buzsaki, *New York U. School of Medicine*
Blanche Capel, *Duke U.*
Mats Carlsson, *U. of Oslo*
David Clapham, *Children's Hospital Boston*
David Clary, *U. of Oxford*
Joel Cohen, *Rockefeller U., Columbia U.*
James Collins, *Boston U.*
Robert Cook-Deegan, *Duke U.*
Alan Cowman, *Walter & Eliza Hall Inst.*
Robert H. Crabtree, *Yale U.*
Roberta Croce, *Vrije Universiteit*
Janet Currie, *Princeton U.*
Jeff L. Dangl, *U. of North Carolina*
Tom Daniel, *U. of Washington*
Frans de Waal, *Emory U.*
Stanislas Dehaene, *Collège de France*
Robert Desimone, *MIT*
Claude Desplan, *New York U.*
Ap Dijksterhuis, *Radboud U. of Nijmegen*
Dennis Discher, *U. of Pennsylvania*
Gerald W. Dorn II, *Washington U. School of Medicine*
Jennifer A. Doudna, *U. of California, Berkeley*
Bruce Dunn, *U. of California, Los Angeles*
Todd Ehlers, *U. of Tuebingen*
David Ehrhardt, *Carnegie Inst. of Washington*
Tim Elston, *U. of North Carolina at Chapel Hill*
Gerhard Ertl, *Fritz-Haber-Institut, Berlin*
Barry Everitt, *U. of Cambridge*
Ernst Fehr, *U. of Zurich*
Anne C. Ferguson-Smith, *U. of Cambridge*
Michael Feuer, *The George Washington U.*
Toren Finkel, *NHLBI, NIH*
Kate Fitzgerald, *U. of Massachusetts*
Peter Fratzl, *Max-Planck Inst.*
Elaine Fuchs, *Rockefeller U.*
Daniel Geschwind, *UCLA*
Andrew Gewirth, *U. of Illinois*
Karl-Heinz Glassmeier, *TU Braunschweig*
Ramon Gonzalez, *Rice U.*
Julia R. Greer, *Caltech*
Elizabeth Grove, *U. of Chicago*
Nicolas Gruber, *ETH Zürich*
Kip Guy, *St. Jude's Children's Research Hospital*
Taekjip Ha, *U. of Illinois at Urbana-Champaign*
Christian Haass, *Ludwig Maximilians U.*
Steven Hahn, *Fred Hutchinson Cancer Research Center*
Michael Hasselmo, *Boston U.*
Martin Heimann, *Max-Planck Inst. Jena*
Yka Helariutta, *U. of Cambridge*
James A. Hendler, *Rensselaer Polytechnic Inst.*
Janet C. Hering, *Swiss Fed. Inst. of Aquatic Science & Technology*
Kai-Uwe Hinrichs, *U. of Bremen*
Kei Hirose, *Tokyo Inst. of Technology*
David Hodell, *U. of Cambridge*
David Holden, *Imperial College*
Laura Hooper, *UT Southwestern Medical Ctr. at Dallas*
Raymond Huey, *U. of Washington*
Steven Jacobson, *U. of California, Los Angeles*
Kai Johnsson, *EPFL Lausanne*
Peter Jonas, *Inst. of Science & Technology (IST) Austria*
Matt Kaeblerlein, *U. of Washington*
William Kaelin Jr., *Dana-Farber Cancer Inst.*
Daniel Kahne, *Harvard U.*
Daniel Kammen, *U. of California, Berkeley*
Masashi Kawasaki, *U. of Tokyo*
Y. Narry Kim, *Seoul National U.*
Joel Kingsolver, *U. of North Carolina at Chapel Hill*
Robert Kingston, *Harvard Medical School*
Etienne Kochlin, *Ecole Normale Supérieure*
Alexander Koldkin, *Johns Hopkins U.*
Alberto R. Kornblitt, *U. of Buenos Aires*
Leonid Kruglyak, *UCLA*
Thomas Langer, *U. of Cologne*
Mitchell A. Lazar, *U. of Pennsylvania*
David Lazer, *Harvard U.*
Thomas Lecuit, *IBDM*
Virginia Lee, *U. of Pennsylvania*
Stanley Lemon, *U. of North Carolina at Chapel Hill*
Ottoline Leyser, *Cambridge U.*
Marcia C. Linn, *U. of California, Berkeley*
Jianguo Liu, *Michigan State U.*
Luis Liz-Marzan, *CIC bioGUNE*
Jonathan Losos, *U. of Harvard*
Ke Lu, *Chinese Acad. of Sciences*
Christian Lüscher, *U. of Geneva*
Laura Machesky, *CRUK Beatson Inst. for Cancer Research*
Anne Magurran, *U. of St. Andrews*
Oscar Marin, *CSIC & U. Miguel Hernández*
Charles Marshall, *U. of California, Berkeley*
C. Robertson McClung, *Dartmouth College*
Graham Medley, *U. of Warwick*
Tom Misteli, *NCI*
Yasushi Miyashita, *U. of Tokyo*
Mary Ann Moran, *U. of Georgia*
Richard Morris, *U. of Edinburgh*
Alison Mottis-Reif, *NC State U. (\$)*
Sean Munro, *MRC Lab. of Molecular Biology*
Thomas Murray, *The Hastings Center*
James Nelson, *Stanford U. School of Med.*
Daniel Neumark, *U. of California, Berkeley*
Kitty Nijmeijer, *U. of Twente*
Pär Nordlund, *Karolinska Inst.*
Helga Nowotny, *European Research Advisory Board*
Ben Olken, *MIT*
Joe Orenstein, *U. of California*
Berkeley & Lawrence Berkeley National Lab
Harry Orr, *U. of Minnesota*
Andrew Oswald, *U. of Warwick*
Steve Palumbi, *Stanford U.*
Jane Parker, *Max-Planck Inst. of Plant Breeding Research*
Giovanni Parmigiani, *Dana-Farber Cancer Inst. (\$)*
Donald R. Paul, *U. of Texas, Austin*
John H. J. Petrini, *Memorial Sloan-Kettering Cancer Center*
Joshua Plotkin, *U. of Pennsylvania*
Albert Pollman, *FOM Institute AMOLF*
Philipp Polzin, *CNRS*
Jonathan Prichard, *Stanford U.*
David Randall, *Colorado State U.*
Colin Renfrew, *U. of Cambridge*
Felix Rey, *Institut Pasteur*
Trevor Robbins, *U. of Cambridge*
Jim Roberts, *Fred Hutchinson Cancer Research Ctr.*
Barbara A. Romanowicz, *U. of California, Berkeley*
Jens Rostrup-Nielsen, *Haldor Topsøe*
Mike Ryan, *U. of Texas, Austin*
Mittori Saitou, *Kyoto U.*
Shimon Sakaguchi, *Kyoto U.*
Miguel Salmeron, *Lawrence Berkeley National Lab*
Jürgen Sandkühner, *Medical U. of Vienna*
Alexander Schlier, *Harvard U.*
Randy Seeley, *U. of Cincinnati*
Vladimir Shalay, *Purdue U.*
Robert Siliciano, *Johns Hopkins School of Medicine*
Denis Simon, *Arizona State U.*
Alison Smith, *Johns Innes Centre*
Richard Smith, *U. of North Carolina (\$)*
John Speakman, *U. of Aberdeen*
Allan C. Spradling, *Carnegie Institution of Washington*
Jonathan Sprent, *Garvan Inst. of Medical Research*
Eric Steig, *U. of Washington*
Paula Stephan, *Georgia State U. and National Bureau of Economic Research*
Molly Stevens, *Imperial College London*
V. S. Subrahmanian, *U. of Maryland*
Ira Tabas, *Columbia U.*
Sarah Teichmann, *Cambridge U.*
John Thomas, *North Carolina State U.*
Shubha Tole, *Tata Institute of Fundamental Research*
Christopher Tyler-Smith, *The Wellcome Trust Sanger Inst.*
Herbert Virgin, *Washington U.*
Berth Vogelstein, *Johns Hopkins U.*
Cynthia Volkert, *U. of Göttingen*
Douglas Wallace, *Dalhousie U.*
David Wallace, *Weizmann Inst. of Science*
Ian Walmsey, *U. of Oxford*
Jane-Ling Wang, *U. of California, Davis*
David A. Wardle, *Swedish U. of Agric. Sciences*
David Waxman, *Fudan U.*
Jonathan Weissman, *U. of California, San Francisco*
Chris Wikle, *U. of Missouri (\$)*
Ian A. Wilson, *The Scripps Res. Inst. (\$)*
Timothy D. Wilson, *U. of Virginia*
Rosemary Wyse, *Johns Hopkins U.*
Jan Zaenen, *Leiden U.*
Kenneth Zaret, *U. of Pennsylvania School of Medicine*
Jonathan Zehr, *U. of California, Santa Cruz*
Len Zon, *Children's Hospital Boston*
Maria Zuber, *MIT*

BOOK REVIEW BOARD

David Bloom, *Harvard U.*, Samuel Bowring, *MIT*, Angela Creager, *Princeton U.*, Richard Sweder, *U. of Chicago*, Ed Wasserman, *DuPont*

Passion is just the start

Last month, 65 Nobel laureates gathered in Lindau, Germany, for an annual exchange of knowledge between these most honored scientists and a group of selected young researchers from around the globe. To extend the reach of the conference, *Science* hosted a webinar to address a common concern of young scientists everywhere: persevering in science. Some advice on this topic may have been unexpected.

Elizabeth Blackburn and Jack Szostak (both 2009 laureates in Physiology or Medicine, for the discovery of how chromosomes are protected by telomeres) and Daniel Shechtman (2011 laureate in Chemistry, for the discovery of quasi-periodic crystals) fielded questions submitted from more than 3100 young scientists who registered for the discussion. Not lost on these Nobelists was the discouragement that young scientists feel about their career prospects. Many find themselves in serial postdocs for many years while fielding rounds of rejection letters for academic positions. In the face of this disheartening situation, there was some tough

advice. Given the current competitive climate for hiring and advancement, the panel agreed that young scientists who are not deeply passionate about their research need to reconsider their career choices. And even if one has such extraordinary passion, it alone is not enough. The panelists stressed the importance of becoming a leading expert in something new and in demand. They also emphasized the need to hone people skills, as science is more and more a team effort, and to become able communicators in order to share goals and achievements with potential funders and the public. Graduate students also need to be smart in selecting a lab for postdoctoral training; ideal attributes are a collaborative atmosphere, opportunities for leadership in publications, and exposure to more than just one senior leader in the field who will be able to write a letter of recommendation.

Certainly, the path to a successful academic career, no less to a Nobel Prize, is not easy. The panelists also faced

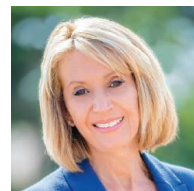
obstacles, periods of self-doubt, and career setbacks. But if a young scientist faces despair from too many failed academic job searches, one panelist had some tough advice: It's time to reassess options as to whether academia is the right track. The panelists encouraged thoughtful consideration of industry and government research labs, because in many cases, they offer more stable funding, and an impressive number of Nobel laureates came from these environments.

A large part of persevering in science is balancing a demanding scientific career with personal life. Panelists emphasized the importance of flexibility; adjusting one's scientific career to meet the needs of a family, for example. I certainly resonated with this advice. A seagoing oceanographer early in my career, I decided to move into a different position that allowed me to stay ashore when my three daughters reached the age when having their mother home regularly really mattered. Once they all were done with college, I was open to job opportunities for which more time away from home was required.

For young scientists in particularly competitive fields, they are challenged with doing everything they can to stand out from the crowd. In this regard, I am delighted by the growing number of events that are devoted to helping young scientists network with very influential leaders who might help advance their careers (such as the Lindau Nobel Laureate Meeting) and prizes that recognize the early-career accomplishments of young scientists (such as the SciLifeLab Prize and the Eppendorf Prize, both of which are administered in conjunction with *Science*). Advisors and mentors may not have the time to look out for such resume-enhancing possibilities for their trainees. I encourage ambitious young scientists to investigate these opportunities.

Many Nobel Prize winners spoke about their "good luck," but on closer inspection, in most cases, they had created the conditions for their own good luck to happen.

— Marcia McNutt



Marcia McNutt
Editor-in-Chief
Science Journals



"...if a young scientist faces despair...It's time to reassess options..."

“I’ve secretly been working on a lander.”

New Horizons principal investigator Alan Stern, at a press conference 14 July as the spacecraft made its closest approach to Pluto. Stern was joking with reporters about plans to return to the dwarf planet.

IN BRIEF

Plants in cold storage



National Museum of Natural History curator Vicki Funk and a summer field team member collect samples from the U.S. Botanic Garden.

In the race to preserve biodiversity even as species extinction rates soar, Smithsonian Institution scientists are starting with their own backyards. Plants from the U.S. Botanic Garden, the U.S. National Arboretum, and the Smithsonian Gardens—all located in Washington, D.C.—are the first targets of an initiative to capture the genomic diversity of half of the world’s living plant genera in less than 2 years. On 8 July, the Smithsonian’s National Museum of Natural History launched the effort, which will include a summer field team of students who will assist scientists with the museum’s Global Genome Initiative in sampling plants from the gardens’ holdings. The scientists will then preserve them in cryogenic vials and store them in liquid nitrogen. GGI, the Norway-based Svalbard Global Seed Vault, and the Royal Botanic Gardens’ Millennium Seed Bank in the United States are three of the largest ex situ conservation efforts to preserve genetic material outside of its natural environment. Ex situ conservation is a strategy described in the Intergovernmental Panel on Climate Change’s 2014 Fifth Assessment Report as an insurance policy against climate change and other potential sources of biodiversity loss.

AROUND THE WORLD

“Cures” bill clears U.S. House

WASHINGTON, D.C. | A bill to speed the discovery and development of new medical treatments sailed through the U.S. House of Representatives last week. The legislation, known as the 21st Century Cures Act, calls for a temporary “innovation fund” that would give the National Institutes of Health an extra \$8.75 billion over 5 years. It also aims to smooth the approval of high-priority treatments; for example, one provision allows the Food and Drug Administration to approve antibiotics for rare life-threatening diseases based on smaller clinical trials. Although the bill enjoys the backing of hundreds of industry, research, and patient organizations, it has also prompted concerns that the effort to speed cures might unintentionally compromise patients’ safety. The updated version of the bill will now head to the Senate, which has yet to release its own biomedical innovation bill. <http://scim.ag/housecures>

Helium rules stifle competition

WASHINGTON, D.C. | New rules on selling off a U.S. government cache of helium—essential for many types of technology and scientific research—aren’t working as planned, witnesses said last week before the Committee on Natural Resources in the U.S. House of Representatives. The 2013 Helium Stewardship Act aimed to establish a competitive market for federal helium by phasing in an auction instead of having the U.S. Bureau of Land Management (BLM) sell it for a fixed price. But the law may be stifling competition, the panel said, by giving an unfair advantage to companies that have storage contracts with BLM to receive the crude helium and refine it. “Nonrefiners” who buy federal helium must also work out a “tolling” deal to get one of the refiners to process it. BLM previously reserved 10% of helium for fixed-price “nonallocated” sales to nonrefiners, but at BLM’s first helium auction in July 2014, two refiners bought it all, shutting the nine nonrefiners out of the bidding. <http://scim.ag/heliumrules>

Appeals court dodges H5N1 issue

AMSTERDAM | Influenza researcher Ron Fouchier of Erasmus MC in Rotterdam, the Netherlands, received a disappointing verdict in his legal battle over so-called gain-of-function research, in which flu strains are made more transmissible in the lab. In 2012, the Dutch government demanded that Fouchier seek an export license before submitting a highly controversial H5N1 flu study to *Science*. Fouchier obtained the license under protest, and Erasmus MC sued the government for infringement of academic freedom—a claim a district court rejected in 2013 (*Science*, 8 November 2013, p. 676). The case went to Amsterdam's Court of Appeal, which annulled the lower court's decision last week, Fouchier says, because Erasmus MC had no standing to sue once it had obtained the license. The appeals court did not rule on the key issue of whether Fouchier's research is subject to export control. The verdict was not yet public as *Science* went to press. <http://scim.ag/FouchierH5N1>

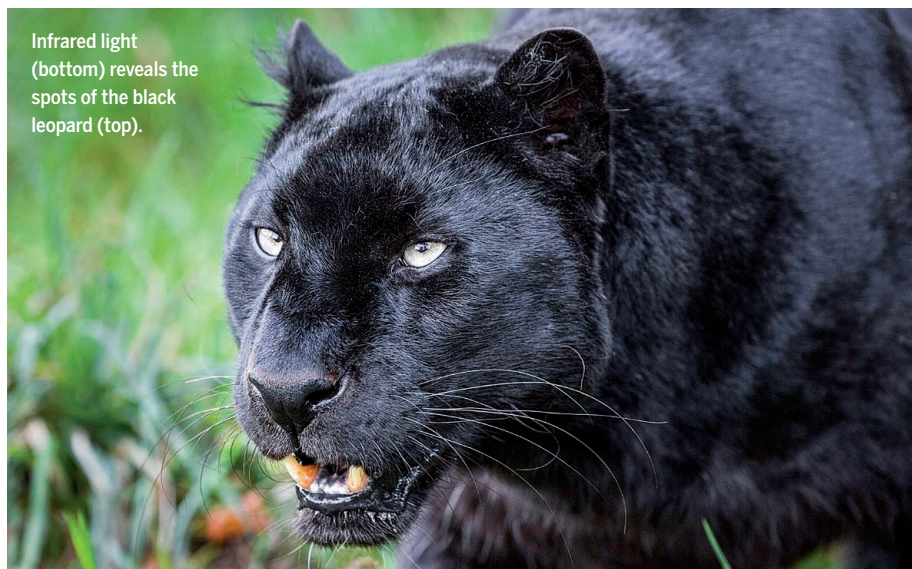
Deal struck on Iran nukes

VIENNA | Iran this week agreed to dismantle large pieces of its nuclear program in exchange for relief from crippling economic sanctions—with science as an important beneficiary. The agreement, 2 years in the making, would slow Iran's "breakout time"—the time needed to produce enough weapons-grade fissile material for one bomb—from an estimated 2 to 3 months to at least a year. Iran will convert its sensitive Fordow uranium enrichment facility into an international "nuclear, physics, and technology centre," with Russia helping to reconfigure two centrifuge cascades to produce stable isotopes for industry. Fordow will also host small linear accelerators for basic research in nuclear physics and astrophysics. Proposals for collaborative projects at Fordow will be reviewed at an international workshop. The deal also calls for the possibility of having Iran participate in the International Thermonuclear Experimental Reactor, or ITER, under construction in France. <http://scim.ag/IranFordow>

NEWSMAKERS

Plant biologist altered images

Two investigations found many instances of image manipulation in papers by **Olivier Voinnet**, a plant biologist at CNRS, the French national research agency, who is currently working at the Swiss Federal Institute of Technology Zurich (ETH). Voinnet's work came under attack in December on



Infrared light (bottom) reveals the spots of the black leopard (top).



Seeing (black leopard) spots

Scientists keep track of the population density of leopards in the wild using camera traps; they snap pictures of the cats and identify individuals by their characteristic spot patterns. But, due to a recessive gene that causes melanism, most of the leopards prowling the Malay Peninsula are black—even their spots—making it nearly impossible to distinguish them using the cameras. Now, there's a workaround, scientists report in the *Journal of Wildlife Management*. The spots are visible at infrared wavelengths, so by modifying infrared flash camera traps on the peninsula so that they were forced into night mode throughout the day, the researchers were ultimately able to distinguish distinct spot patterns and identify 94% of the animals along a wildlife corridor in Malaysia—crucial for keeping tabs on the region's population over time. The work not only represents the first leopard density estimate in Malaysia, the authors note, but also the first successful attempt to estimate population size using melanistic phenotypes—such as a black leopard's dark spots.

a website called PubPeer; a CNRS investigation that won't be made public found manipulations that constituted scientific misconduct and "tarnished" CNRS's reputation, the agency said on 10 July. CNRS suspended Voinnet for 2 years. On the same day, ETH announced that a panel had found problems in 20 of Voinnet's papers, but that those produced while he was at ETH did not constitute misconduct under the institute's definition. Voinnet received an "admonition" but will keep his position at ETH. In a statement, he apologized and took "complete responsibility."

FINDINGS

Cholera vaccine shows promise

An inexpensive cholera vaccine may help prevent the deadly disease in the more than 50 countries, mostly in Asia and Africa, where it is endemic, according to a new study in *The Lancet*. Cholera causes severe diarrhea and is estimated to kill about 95,000 people—a large number of whom are children—every year. In the first-of-its-kind study, scientists tested the vaccine, called Shanchol and made in India, in 270,000 people in the slums of

Dhaka, where the threat of contracting the infection is nearly constant. They found that the vaccine provided individuals more than 50% protection against cholera and lowered life-threatening episodes of the infection by about 40%. When given as part of routine health services, the vaccine, which consists of a two-dose regimen that currently costs \$3.70, could drive down the toll considerably, the researchers say.

A flightless, winged dino

Researchers have found the largest ever dinosaur with full-fledged wings and feathers, they report online this week in *Scientific Reports*. The 125 million-year-old dino, called *Zhenyuanlong suni* and found in China, was about 1.65 meters long, a little longer than a modern condor. It belonged to a group of dinos called dromaeosaurs, which includes *Velociraptor* from *Jurassic Park*, and which was closely related to early birds. But although *Zhenyuanlong's* wings had multiple layers of birdlike feathers, they were very short compared with those of most other winged



The short-armed, winged, and feathered dinosaur *Zhenyuanlong suni*.

dinos, and thus it was probably unable to fly. That leaves the discovery team wondering what the wings were for. One possibility is that it evolved from ancestors that could fly, but used its wings mainly for sexual display. <http://scim.ag/flightlessdino>

Enter the pentaquark

GENEVA, SWITZERLAND | A long search paid off this week, when researchers at the European high-energy physics

laboratory CERN announced that they had sighted strong evidence of particles made up of five quarks. Predicted by physicist Murray Gell-Mann in 1964 to exist alongside protons and neutrons (which contain three quarks) and mesons (which are quark-antiquark pairs), pentaquarks would consist of four quarks and an antiquark. Scientists thought they had discovered a light pentaquark known as theta-

plus in 2003, but those claims evaporated under further scrutiny (*Science*, 22 April 2005, p. 478). Physicists at the Large Hadron Collider's LHCb experiment instead snared a heavier "charmonium" pentaquark by sifting through thousands of decays of three-quark lambda-b particles. The finding, which still has to be confirmed by other groups, suggests the existence of a variety of pentaquarks with different masses, the scientists say. <http://scim.ag/pentaquark>

BY THE NUMBERS

15
million

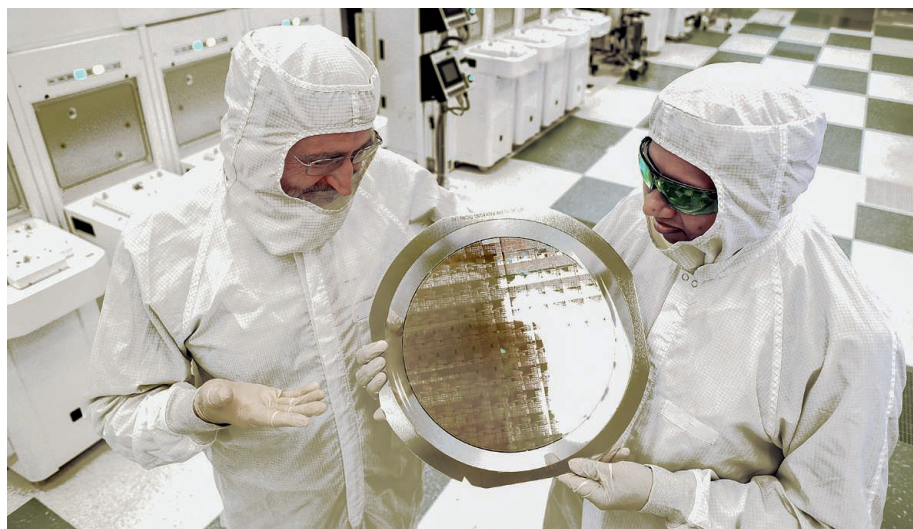
Number of people receiving antiretrovirals for HIV in March 2015, according to a new UNAIDS report released this week.

834,000

Number of rabbits, nonhuman primates, and other regulated animals used in U.S. biomedical research last year—a drop of 6% from the previous year, and the lowest number since data collection began in 1972.

18.7%

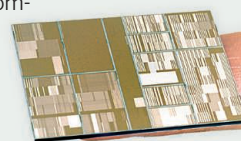
Increase in global wildfire activity over the last 35 years, according to a study in *Nature Climate Change*.



Scientists (top) display a wafer made of 7-nanometer chips (inset).

IBM back on track with Moore's law

IBM announced last week that it has manufactured working versions of a new, ultra-dense computer chip that will reduce the area needed for a given amount of circuitry by half and has four times the capacity of the most powerful chips on the market. Currently, commercial chip technology is at the "14-nanometer manufacturing" stage, referring to the size of the chip's smallest features. The new IBM chips have 7-nanometer transistors, making it possible to pack more into the same space and increasing the capacity. The new chips were manufactured by using silicon-germanium in some regions rather than pure silicon, which the company says reduces power requirements and makes faster transistor switching possible. But IBM hasn't said when the new chips are likely to be commercially available.



PHOTOS: (TOP TO BOTTOM) JUNCHANG LI; DARRYL BAUTISTA/FEATURE PHOTO SERVICE FOR IBM (2)



IN DEPTH

A 2007 protest denouncing waterboarding as torture; several psychologists guided the method's used on U.S. military detainees.

SCIENTIFIC SOCIETIES

Torture report prompts APA apology

Admitting it colluded with U.S., psychologists group to change policies, leadership

By John Bohannon

After years of denying that it had given scientific and ethical legitimacy to torture by the U.S. government, the American Psychological Association (APA) last week accepted the finding of an external investigation that concluded it had done just that. Now, with a public apology and sudden wave of high-level resignations or retirements, APA is struggling to craft an institutional response that will satisfy its members and long-time detractors, even as some of those pilloried in the probe defend themselves and their colleagues.

"This is a crisis," says Nadine Kaslow, a psychologist at Emory University in Atlanta and a former APA president, who helped launch the investigation. "I regret that the organization didn't listen to the critics earlier."

The 542-page report from a former Chicago inspector general, David Hoffman, pulls no punches, concluding that APA officials colluded with the U.S. government to enable the torture of detainees. APA's Board of Directors quickly released

a response, promising among other things to recommend a new policy prohibiting psychologists from participating in interrogation of persons held in custody by military and intelligence authorities. APA then announced the departure of most of its staff leadership: CEO Norman Anderson, Deputy CEO Michael Honaker, public relations director Rhea Farberman, and ethics director Stephen Behnke.

"All of these people were in the know," says former APA President Gerald Koocher, a psychologist at DePaul University in Chicago, Illinois, and the editor of the journal *Ethics & Behavior*. Koocher himself is under fire, as two of APA's staunchest critics have called for the body to censure him as well, after receiving a confidential briefing on the Hoffman report earlier this summer.

APA had commissioned the report last year after the publication of *Pay Any Price: Greed, Power and Endless War*, a book by *New York Times* reporter James Risen that accused the organization of providing cover for torture. The report's most damning findings concern a 2005 APA committee called the Task Force on

Psychological Ethics and National Security (PENS). The task force was created in the midst of revelations that detainees were subjected to "enhanced interrogation" at U.S. government facilities in Iraq, Afghanistan, and Guantanamo Bay, Cuba, and that psychologists were intimately involved in both the design and practice of these efforts.

As Hoffman discovered through interviews, medics within the intelligence community were "not on board" with such interrogations. To quell this internal resistance, the government hoped to enlist support from APA, psychology's largest professional organization. And the PENS task force provided it, concluding in a 2005 statement that it was ethical for psychologists to take part in the interrogation program.

The PENS decision sparked protests by many APA members, some of whom called for withholding dues, but Hoffman found that they were ignored. "Being involved in the intentional harming of detainees ... could do lasting damage to the integrity and reputation of psychology, a profession that purports to 'do no harm,'" he writes, but "these countervailing concerns were

simply not considered or were highly subordinated to APA's strategic goals." According to Hoffman, APA sought to maintain its privileged relationship with the Pentagon, a massive employer of psychologists.

Hoffman's analysis of internal APA emails found that the members of the PENS task force were carefully chosen in a collaboration involving officials at APA, the Pentagon, and the Central Intelligence Agency, and its conclusions were vetted in advance by insiders at both agencies. The goal of PENS, Hoffman offers, was not to examine the ethics of torture but to "curry favor" with the U.S. Defense Department.

Hoffman's characterization of PENS is unfair, according to Koocher, who was one of the architects of the task force. "We solicited widely and openly for membership," he says. The fact that so many task force members came from the military is not evidence of collusion but good judgment. "If you're focusing on interrogation in a military context then those are the people with the relevant expertise." As for the allegation of currying favor with the Pentagon, Koocher is adamant that it was not his goal. "No way were we covering up for [Vice President] Cheney or [Defense Secretary] Rumsfeld, both of whom I cannot stand."

Koocher says that he was unaware that the torture was ongoing. He points out that he, along with other representatives of U.S. medical associations, visited the detention center at Guantanamo in 2006. "I asked hard questions," he says. When it was later revealed that torture continued at the facility, "I was extremely upset." But by then, he says, "I was no longer an APA official. What was I supposed to do?"

That sentiment may not save Koocher from sanctions. He is on a list of APA members to be banned from APA governance "effective immediately"—just one of several recommendations from Steven Reisner of New York University and Stephen Soldz of the Boston Graduate School of Psychoanalysis, who also urged that APA's top executive, legal, and public relations staff be fired. Reisner and Soldz, persistent critics of APA's role in the interrogation program, were invited by APA to review the Hoffman report in advance and give the society their feedback. APA wouldn't comment specifically on the pair's recommendations; several people on their "staff to be fired" list remain with APA. "A lot of change can happen, but it will take a lot of time to implement it," Kaslow says.

APA's 180th turn is only a start, Soldz says. "The APA and the entire psychology profession needs to grapple with the enormous scandal enveloping psychological ethics." ■

CLINICAL TRIALS

Researchers seek clear reasons when clinical trials end early

Explanations for abandoning tests of new treatments earlier than planned are often hazy

By Jennifer Couzin-Frankel

Like a marathon with far more runners limbering up at the start than stumbling across the finish, the race to bring a new treatment to market has dropouts along the way. About 12% of clinical trials are reported to shut down prematurely. Knowing why could help minimize the number of terminated trials going forward.

But a paper published earlier this month by a group of computational biologists suggests that this knowledge isn't easy to come by. The main reason: Companies can type whatever they want into the tiny space—a mere 160 characters—allotted on ClinicalTrials.gov, a registry maintained by the National Library of Medicine. In their analysis of all 3122 terminated trials on the registry at the time their study began, "it just seemed like a complete mess," says

Frederick Roth, a systems biologist and geneticist at the University of Toronto in Canada. Reasons were often murky, ranging from "it was decided to not proceed with the study at this time" to "SARS epidemic in Asia and Canada."

While at Harvard University, Roth and two undergraduates came to their project quite by accident. They were interested in adverse interactions between drugs or between drugs and genes. They decided to do what he calls an "amateur pass" through ClinicalTrials.gov.

Their search didn't yield much. The reasons for termination were so diverse and often so vague that Roth and his students decided to launch a new project: learn more about why clinical trials end early by dividing the information given in the short blurbs into "buckets," such as funding, ethical reasons, or business decisions, so they could see the breakdown by category. They found that by far the most popular reason was insufficient enrollment, accounting for about one-third of terminated trials. About 11% failed to establish efficacy. In all, Roth

and his students identified 35 categories, among them "lost interest," "inadequate design," and "key staff left." (In one case, an orthopedic surgeon performing a trial's knee surgeries moved elsewhere.) "It's a potpourri of reasons why you terminate," agrees Deborah Zarin, the director of ClinicalTrials.gov.

In the new paper, published online at bioRxiv.org, Roth and first author Theodore Pak, now an M.D./Ph.D. student at the Icahn School of Medicine at Mount Sinai in New York City, recommend that ClinicalTrials.gov aim for greater transparency by asking sponsors to answer several questions when their trials end early. Those include whether it even started, whether data were ever examined, and whether inter-

terim examinations of efficacy or safety played a role.

Zarin welcomes Roth's dive into the myriad explanations, but she is cautious about the questions the authors recom-

mend for trial sponsors. Safety and efficacy can't always be easily separated from each other, she says. Zarin is especially interested in distinguishing between trials that end because of the science versus some other reason: In May, she and her colleagues reported in *PLOS ONE* that 68% of 905 terminated trials with results listed on ClinicalTrials.gov stopped for reasons other than data accumulated in the study. Only one-fifth ended early because of safety or efficacy concerns from trial data, and the rest didn't give a reason.

Another unsettling issue is that many terminated studies aren't listed as such in ClinicalTrials.gov and other databases because the entries are not up to date, according to ongoing research by clinical epidemiologist Matthias Briel at the University Hospital of Basel in Switzerland. That, combined with sometimes squishy reasons offered for trial termination, suggests to him that one implication of Roth's paper is that "people might then take information from these registries and introduce inaccuracies at least, or even bias into their own studies, without realizing it." ■

"It just seemed like a complete mess."

Frederick Roth, University of Toronto



The World Health Assembly rejected a proposal to increase membership fees in May.

GLOBAL HEALTH

Report prescribes strong medicine for WHO

Ebola failures show that difficult reforms are needed, an independent panel says

By Kai Kupferschmidt

The World Health Organization (WHO) is still struggling to end the Ebola outbreak in West Africa. But in the wake of what many viewed as a sluggish response to the crisis, another battle is brewing—about the future role of the organization itself.

On 7 July, an independent panel delivered a scathing review of WHO's performance and proposed wide-ranging reforms that would enable the agency to better tackle the next major health crisis—from giving it more money and power to setting up a special, semi-independent emergency center. WHO “does not currently possess the capacity or organizational culture to deliver a full emergency public health response,” concluded the group, led by Dame Barbara Stocking, the former chief executive of Oxfam in the United Kingdom; it recommended that the agency be made “fit for purpose.”

But although many of the suggestions have been praised as sensible, WHO's complex, politicized governance structure and entrenched bureaucracy make the \$2 billion U.N. agency difficult to change. Just 2 months ago, for instance, the World Health Assembly (WHA)—the annual meeting of health ministers that is WHO's decision-making body—rejected a proposal by WHO Director-General Margaret Chan to increase member states' contribution fees by 5%, as the report recommends. Several other suggestions in the report were previously made by a panel that reviewed the H1N1 pandemic of 2009, but

were never implemented.

The new report, too, “might just die on paper,” says Joanne Liu, president of Doctors Without Borders (MSF), which has played a major role in the Ebola epidemic. “Everything at this stage is in the hands of the member states and how willing they are to give a second chance to WHO.”

The six-member panel, appointed by WHO's Executive Board in March, interviewed WHO sources and outside experts, met with representatives of relief organizations, and flew to the affected countries. It explored why WHO waited until 8 August 2014, when Ebola had already infected more than 1000 people, to declare the outbreak a Public Health Emergency of International Concern, a status that triggers international action. Early warnings about the outbreak's scale within WHO were ignored, the panel found, partly because “WHO does not have an organizational culture that supports open and critical dialogue between senior leaders and staff or that permits risk-taking.”

In addition, local community leaders—whose collaboration would prove crucial to controlling Ebola—were engaged too late, deployments of staff and consultants in the field were too short, and communication with the media was ineffective.

The report's most far-reaching proposal is that WHO set up a new Centre for Health Emergency Preparedness and Response that would bring together WHO experts in outbreak control and those providing humanitarian assistance, two areas that are now separate. The center should be overseen by an

independent board and headed by “a strong leader and strategic thinker,” the report notes. It's a good proposal, says Norwegian epidemiologist Preben Aavitsland, the head of a researching and consulting company in Kristiansand, Norway. But WHO officials may resist the idea, because a strong emergency center could become a state within a state, he cautions. “It will be interesting to see whether that is followed up.”

Others say WHO does not need an internal capacity to mount a full-on response to every outbreak. Rather than expanding, the organization should depend more on external partners like the U.S. Centers for Disease Control and Prevention—which has three times WHO's annual budget—or Public Health England, says David Heymann, a former WHO assistant director-general who is now at the London School of Hygiene & Tropical Medicine. “To be sustainable, the workforce should be outside of WHO,” he says. “I would be totally against WHO increasing its staff to sit around and wait for outbreaks to occur.”

The report also stresses the need to boost countries' compliance with the International Health Regulations (IHR), which lay out the responsibilities of WHO and member states during outbreaks. Completely revised and modernized after the SARS outbreak of 2003, the current IHR require member states to have “core public health capacities”—such as laboratories and trained staff—to detect and fight infectious diseases. So far, only 64 of 193 member states have confirmed that they comply with these rules; WHO should develop a plan to bring all countries up to

speed, the report says.

That is easier said than done in dirt-poor countries where even basic medical services are lacking, Liu says. “It’s this long, huge shopping list of things that you are supposed to implement, with no prioritization,” she says. WHO should narrow the requirements down to the essentials, Liu argues. Aavitsland adds that developed countries should help foot the bill.

In addition, the report says, WHO should penalize countries that violate an IHR provision banning member states from issuing travel restrictions if WHO doesn’t deem them warranted. More than 40 countries imposed unnecessary restrictions during the Ebola outbreak, which hurt the economies in Guinea, Sierra Leone, and Liberia, and made it harder to send people and supplies. The report recommends giving WHO the power to sanction such behavior and to take cases to the U.N. Security Council. But that requires countries to surrender some national powers, something politicians are notoriously reluctant to do. “The current IHR were the result of a political negotiation that took much time,” Heymann says. “In order to change them, there have to be new negotiations.”

A spokesperson for WHO says “several of the recommendations in the report have been addressed already or are being addressed.” For instance, the organization is looking seriously at establishing a new emergency center, he says. Whether any real change is possible will become clear at the next WHA in May 2016, Aavitsland says. The issue of WHO’s budget will surely come up again then as well. “Member states keep increasing their expectations of what WHO will deliver, in many areas, without giving the organization the means to fulfill these expectations,” he says. “I think member states will get what they pay for.”

In 2017, Chan, who has been criticized for being too conflict-averse, is stepping down; a new, more forceful personality at the helm could make a huge difference, says Ilona Kickbusch, an independent global health consultant based in Brienz, Switzerland, and one of the authors of the new report. A group looking at proposals to reform WHO’s governance has discussed limiting directors-general to one 7-year term instead of two 5-year terms, she says. Free of the need to be reelected, WHO leaders might find it easier to challenge member states.

Liu is hopeful that a stronger WHO will emerge from the soul-searching. “The reality is today, there is only one organization that has the legitimacy and the authority to call a Public Health Emergency of International Concern, and that is not MSF, or the World Bank. It is WHO,” she says. “So they need to have the capacity to do that.” ■

SCIENCE POLICY

Russia targets Western ties

Crackdown on “foreign agents” and “undesirable” groups threatens private support for science

By Vladimir Pokrovsky

The news last week that the Dynasty Foundation, Russia’s only private funder of scientific research, is closing its doors adds to a darkening prospect for philanthropic support of Russian science. The decision by the Dynasty Foundation’s council, announced 8 July in a terse one-sentence notice on Dynasty’s Russian-language website, came weeks after the Russian government had labeled Dynasty a “foreign agent” under a recently enacted law.

Zimin, funds the foundation from personal accounts held in Western banks. Zimin was infuriated and left Russia in early June. Dynasty’s governing council decided to try to keep operating without personal donations from Zimin, but it declared the foreign agent label “absolutely unacceptable.” The final straw came in mid-June, when a court fined Dynasty 300,000 rubles (about \$5500) for refusing to register as a foreign agent.

Anna Piotrovskaya, the foundation’s executive director, told the press that she cannot say exactly when it will close but that all of Dynasty’s obligations to current grantees



Russian parliamentarians Valentina Matviyenko and Konstantin Kosachev pushed for the new laws.

That move was part of what many see as a growing official crackdown on organizations the government considers subversive. A new mechanism, separate from the foreign-agent law, threatens to label such groups as “undesirable” and make collaborating with them illegal, potentially curtailing their support for scientists. “These two stories are very symptomatic and replicate each other,” says Mikhail Gelfand of the Russian Academy of Sciences’ (RAS) Institute for Information Transmission Problems in Moscow, one of Russia’s top biologists.

In May, the Russian Ministry of Justice added Dynasty to its list of foreign agents—a new designation for organizations that receive funding from the West (*Science*, 5 June, p. 1067)—on the grounds that Dynasty’s founder, telecommunications mogul Dmitry

will be fulfilled. Vladimir Putin’s spokesman Dmitry Peskov said the Kremlin regrets that the foundation will be liquidated. But he said no one forced the closure.

Some Russian scientists, however, called the move outrageous. Zimin’s public humiliation was an “especially dishonorable action,” says Evgeny Onishchenko, a researcher at RAS’ Lebedev Physical Institute in Moscow. “He was unselfishly helping to develop and popularize science in Russia, but the authorities have ... publicly made a spy of him.”

As Dynasty’s board deliberated, Russian authorities were enacting even more draconian measures. A new law, which was passed in May and took effect in June, authorizes the Russian prosecutor general to label “undesirable” any group deemed to threaten the foundations of constitutional order, national

defense, or national security. Organizations so designated must cease activities in Russia and will have their bank accounts blocked. Groups or individuals who cooperate with them, even outside Russia, face fines and could be sentenced to as many as 6 years in prison for a second offense.

On 8 July, the Federation Council—the upper chamber of Russia's parliament—published a suggested list of 12 undesirable organizations for the prosecutor general to investigate. The list includes the MacArthur Foundation, which has run grant programs in Russia, and the Open Society Foundations of Hungarian-born philanthropist George Soros. Parliamentarians say many more names will follow. Press reports have suggested that they could include the Ford Foundation, the International Research & Exchanges Board (which awards grants for postgraduate study in the United States), and other donors.

Seeing Soros's name on the list was bewildering, researchers say, because one of his charities—the International Science Foundation (ISF)—helped ensure the survival of Russian science in the mid-1990s. During the economic turmoil that followed the collapse of the Soviet Union, ISF spent more than \$100 million on assistance that included 35,000 emergency research grants in Russia and funding to build 32 university computer centers in Russia's major cities. Soros's foundations haven't been active in Russia since 2003 and have shown no sign of plans to return.

"Soros has done very much" for Russia, says Pavel Arseniev, who directed ISF's Russian office in the mid-1990s. Arseniev notes that the government of former Prime Minister Viktor Chernomyrdin had partnered with ISF in funding travel grants. Gelfand says the Russian scientific community is partly to blame. "They did not react to accusations that Soros allegedly was buying [Russian] secrets."

As for Dynasty, Arseniev faults its lawyers for giving the government an opening. "If you work in an aggressive environment, you cannot ignore such details," he says. "Zimin, though he never positioned himself as a political figure, has always behaved very independently, which undoubtedly annoyed the conservative part of the establishment. Once they got a possibility, they took him off the stage."

Onishchenko says the attack on Dynasty and the black list are troublingly reminiscent of the Soviet era. "As in the former gloomy time, they create the image of an enemy who is responsible for the majority of Russia's problems," he says. "And those who helped Russian science get persona non grata status." ■

DATA CHECK

BEHIND THE NUMBERS

Salaries pump up biomedical inflation

By Jeffrey Mervis

According to the conventional wisdom, the cost of the things needed to do biomedical research in the United States rises faster than the price tags on all consumer goods and services. For 3 decades, an index created by the National Institutes of Health (NIH) has documented that disparity, giving lobbyists more ammunition to plead with lawmakers that NIH's annual budget hikes should exceed the overall rate of inflation in the country.

The NIH index, which captures the cost of such things as reagents, test animals, and scientific equipment, has at times outpaced that broader index by as many as three percentage points. But in 2012, a strange thing happened that challenges the conventional wisdom: The Biomedical Research and Development Price Index (BRDPI) fell below the Gross Domestic Product Price Index (GDP PI), a variation of the more familiar Consumer Price Index. The biomedical index's 1.3% growth rate that year not only trailed the GDP PI's 1.9% but also was the lowest in the BRDPI's history. (The news, which NIH reported in January 2014 after the final 2012 numbers had been crunched, went largely unnoticed at the time.)

To find out why that year was such an anomaly, one needs to know what goes into the BRDPI. NIH told *Science* that the information wasn't publicly available, but that it could be obtained under the federal Freedom of Information Act (FOIA). We did so, and learned that the index is dominated not by the cost of equipment and supplies, but rather by the salaries and benefits paid from a grant. In fact, overall personnel costs typically account for two-thirds of the change in the index from one year to the next.

The outsized effect of salaries and benefits on biomedical inflation became clear after Congress passed a spending bill in December 2011 that lowered the salary ceiling for investigators on a standard NIH grant from \$199,700 to \$179,700. Legislators wanted to free up money for more research grants at a time when scientists complained that NIH's \$30 billion budget couldn't support enough of their good ideas. The 2011 bill was the most recent

example of NIH's chronic boom-and-bust funding cycles. A 2-year, \$10 billion budget spike, part of a massive stimulus package to help the U.S. economy recover from the 2008 global financial meltdown, was ending, and money was tight.

The decision to cap salaries sent the BRDPI into a tailspin, bringing it below the already low GDP PI. In 2008 the biomedical index had hit a 20-year peak of 4.7%, more than twice the 2.1% rate for the GDP PI. By 2010 it had fallen somewhat, to 3.0%, but it still far outpaced the minuscule 0.9% rate clocked by the GDP PI. Then it plummeted in 2012, whereas the GDP PI rose to 1.9%.

Outsiders might have thought the drop would be good news. After all, if biomedical inflation slows, then NIH should be able to stretch its limited dollars further.

But top NIH officials did not trumpet the slowdown. The former head of NIH's extramural program, Sally Rockey, was in the habit of blogging each year about the value of the index. She has called it an "important [way] ... to measure changes to the purchasing power of the NIH dollar and make

projections for future fiscal years." But in a 28 March 2014 posting, she simply noted that the 2012 dip "was primarily driven by the reduction in the salary cap for principal investigators."

Nor does a departmental memo announcing the record-low 2012 BRDPI number include any formal reaction to the surprising slowdown. The same memo does include a preliminary estimate for the 2013 BRDPI that shows it exceeding the broader index, and the authors appear relieved at the prospect of returning to the status quo. The 2013 BRDPI figure, they note, although "still low by historical standards ... is at least once again growing at a higher rate than the [GDP] price index."

NIH enjoys strong support in Congress, and the realization that biomedical inflation largely tracks salary trends, not the sticker price of essential lab equipment and supplies, is unlikely to have a major impact on policy debates. Still, it may behoove biomedical lobbyists to think twice before citing the cost of high-tech science as a rationale for pumping up NIH's budget. ■

66
percent

Share of the biomedical research price index taken up by salaries and fringe benefits.



MEANS TO AN END

Cities, states, and provinces are gearing up to halt their AIDS epidemics—though the definition of success varies

By Jon Cohen

After what Shane Ryan calls “a silly, stupid weekend” in late April, he joined the many other men in San Francisco, California, who line up six mornings each week and wait for the doors to open at Magnet. Located in the city’s Castro district, Magnet provides sexual health services for gay and bisexual men. Last year, the center diagnosed 37% of the new HIV infections in the city. Ryan,

like most of the other men in line, came to find out whether his condomless sex had the consequence he feared. “Usually, I’m very safe, always precautionous, but I went a little sidetracked that weekend,” Ryan said.

Ryan, who is 24 years old and grew up in Ireland, met with a nurse practitioner, who explained that he could immediately start taking antiretrovirals (ARVs). This so-called postexposure prophylaxis (PEP) might abort an infection if indeed he had been exposed to

HIV, the nurse explained. Ryan began taking a pill each day that contains four ARVs.

In June, when Ryan’s follow-up test came back negative, another Magnet nurse practitioner, Pierre-Cédric Crouch, asked if he wanted to start on *pre*-exposure prophylaxis, or PrEP. “Having gone on PEP is an indication that PrEP might be a good thing for you to do,” Crouch said. He explained that the daily PrEP pill contains only two of the four drugs used in PEP and has far fewer side effects. It



After this test showed Shane Ryan's blood was negative for the AIDS virus, Magnet's nurse practitioner Pierre-Cédric Crouch (background) offered him anti-HIV drugs to prevent infection.

is not a substitute for using condoms, he said, but it provides "that extra layer of protection." Ryan opted to try it and was directed to a PrEP benefits manager who would help him get the drug at low or no cost.

Magnet's aggressive attempt to prevent the spread of HIV to people at high risk is part of a groundbreaking citywide initiative launched this January that aims to make San Francisco the first jurisdiction in the world to get transmission down to such low levels that it effectively will "end" its epidemic. Specifically, San Francisco wants to achieve a 90% drop in new infections and deaths by 2020. "I think we're going to get to the point where HIV diagnoses are very rare," says Steve Gibson, who heads Magnet.

New York state and the Canadian province of British Columbia (BC) are also at the vanguard of the ending AIDS movement. These three locales have different definitions of what it means to end AIDS, as well as different approaches, tailor-made to their distinct demographics and politics. All, however, rely on aggressive use of ARVs and intensive efforts to identify newly infected people and their partners. Those strategies

reflect several game-changing findings of the past 5 years.

Researchers now know that if infected people take their ARVs and knock down the virus to undetectable levels in the blood, they rarely transmit it to their sexual partners. The approach, called treatment as prevention, pays special dividends with recently infected people, who account for a disproportionate amount of spread. Studies have also shown that uninfected people who take the drugs pre-exposure—PrEP—can greatly reduce their own risk of infection.

Buoyed by these remarkable advances against a virus that to date has infected 76 million people and shortened the lives of half of them, the Joint United Nations Programme on HIV/AIDS (UNAIDS) in 2014 set an "ambitious treatment target" of 90-90-90 by 2020: 90% of infected people worldwide will know their HIV status, 90% of those will receive ARVs without interruption, and 90% of those will have no detectable virus in their blood. Doing the math, that means 73% of the total infected population will have fully suppressed viral levels. This treatment-as-prevention strategy, even without PrEP, will "end the AIDS epidemic as a major global health threat by 2030," UNAIDS predicts.

A just-released report, "Defeating AIDS—advancing global health," in the 25 June online issue of *The Lancet*, is less optimistic. "There is an urgent need to do more and to do better now," declares the report, by a commission including heads of state, public health leaders, and even actress Charlize Theron. It notes that too many locales have sluggish responses and little hope of downgrading their epidemics to "low-level endemicity" (see sidebar, p. 230). But BC, New York, and San Francisco aim to pave the way in the ending AIDS movement, which will be a central topic at an international HIV/AIDS meeting in Vancouver 19 to 22 June.

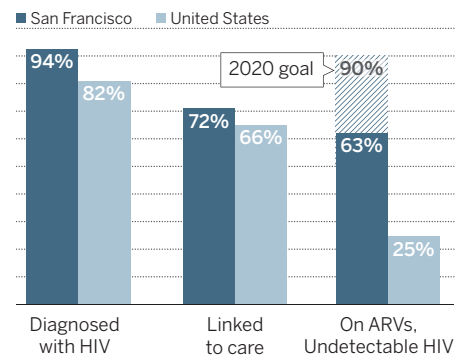
JULIO MONTANER, CO-CHAIR of the Vancouver meeting and the key architect of 90-90-90, took a gamble when he and his colleagues published an editorial in the 5 August 2006 issue of *The Lancet* that argued for making ARVs the backbone of prevention efforts. "We were told by UNAIDS that we were medicalizing prevention, and that this was irresponsible," says Montaner, who is at the BC Centre for Excellence in HIV/AIDS in Vancouver (an offshoot of the University of British Columbia). He acknowledges that he and his colleagues were relying more on mathematical models than hard evidence that treatment as prevention worked. But studies vindicated him. "I'm happy we were right."

Although BC has not declared an explicit goal for "ending AIDS," Montaner says it is

committed to reaching the 90-90-90 target by 2020. Nearly seven times the size of New York state, the province has only 4.6 million residents, an estimated 12,000 of whom are infected with HIV, giving it a prevalence rate of 0.3%. The epidemic boomed in the early 1990s in men who have sex with men (MSM). When heroin use surged around 1996 in downtown Vancouver, HIV infections did as well, which led authorities to declare a public health emergency the next year. The government expanded a long-running methadone maintenance program, distributed free needles, and in 2003 opened the first supervised injection site in North America. In 2011, the province began offering ARVs to all

Treatment cascade

For the treatment-as-prevention strategy to work, people must know their HIV status, receive appropriate medical care, and take ARVs regularly such that their HIV becomes undetectable.



infected people, whereas most of the world, concerned about cost and side effects, still recommended withholding treatment until a person's immune system was damaged.

With the widespread use of ARVs, AIDS-related illness and death in BC has plummeted since 1996. New HIV diagnoses have fallen from a high of 681 per year in 1995 to 262 in 2014, with a huge drop among people who inject drugs. But among MSM and heterosexuals, the number of new diagnoses has barely budged for several years. And although Vancouver has been relatively successful, the epidemic is poorly controlled in more remote regions of the province.

In BC as a whole, some 80% of infected people know their HIV status, 61% take ARVs at some point, but only 51% have undetectable viral levels—the cornerstone of the treatment-as-prevention strategy. "Treatment as prevention has not been pushed to the limit," Montaner says. So the BC center and its collaborators have recently stepped up efforts to find infected people and make sure everyone who starts ARVs stays on them so the virus is fully suppressed.



Following a surge in HIV infections from sharing needles, Vancouver opened Insite in 2003 for drug users to inject safely.

As part of its revved-up effort, BC is one of a growing number of places that is exploiting DNA sequencing of the fast-evolving virus to work out how isolates are related. That allows epidemiologists to pinpoint hot spots of transmission, where interventions can then be targeted (*Science*, 12 June, p. 1188). These data have led health authorities to concentrate outreach efforts in the northern, rural part of the province, which is home to many indigenous Canadians.

Some criticize BC's program for not going beyond 90-90-90 and offering PrEP to high-risk but uninfected people like Shane Ryan, a mainstay of efforts in New York and San Francisco. "I don't think we're going to eliminate the epidemic without PrEP," says Susan Buchbinder, who directs HIV prevention at the Department of Public Health in San Francisco. Montaner, who notes that Canada's government-funded health care plan does not reimburse for PrEP, is not convinced PrEP is needed. But he says he is keeping an open mind. If in 5 years the jurisdictions that combine both strategies are ahead of us, he says, "I'll say PrEP is the path to take."

MARK HARRINGTON and Charles King hatched their bold plan for ending the AIDS epidemic in New York in the back of a paddy wagon. In 2012, during an international AIDS conference

in Washington, D.C., the prominent activists were arrested outside the White House while protesting the Obama administration's HIV/AIDS policies. Based in New York, Harrington runs Treatment Action Group, an HIV/AIDS think tank, and King is head of Housing Works, which helps find homes for people living with the virus. Convinced that the administration's plan was not aggressive enough, they began brainstorming while en route to their jail cell about how they could do better.

With support from HIV/AIDS researchers at Columbia University and the AIDS Institute at the New York State Department of Health, Harrington and King organized

community meetings to explore how New York could drive down new infections to such a low level that the epidemic would peter out. "It brought more gravitas to the whole thing that it was researchers, advocates, and policymakers all working together to push for this," says psychologist Robert Remien, who heads Columbia's HIV Center for Clinical and Behavioral Studies.

The idea received an enormous political boost in June 2014 when New York Governor Andrew Cuomo, a Democrat, announced a three-point plan to end the state's AIDS epidemic by expanding testing, treatment, and access to PrEP. A few months later, Cuomo set up an Ending the Epidemic (ETE) Task Force,

which includes Harrington, King, other community advocates, scientists, and government health workers. In April 2015, at a ceremony to unveil the group's blueprint, Cuomo said that when he announced the goal, people thought it was "outrageous." But New York will succeed and set an example, said Cuomo, who budgeted \$10 million this year to support the ETE project.

The meaning of "ending AIDS" varies from place to place, and the "Defeating AIDS" report urges the international AIDS community to agree on "a precise scientific and epidemiological definition of low-level endemicity." The report defines the end



On Gay Pride Day in June 2014, New York Governor Andrew Cuomo (D), shown marching in a Manhattan parade, announced a plan to end AIDS in the state.

as a reproduction number below one, meaning that each infected person transmits the virus to less than one other person on average. At that rate, the epidemic will gradually fade. No locale has adopted that exact language. Governor Cuomo offered this definition: “The end of the AIDS epidemic in New York state will occur when the total number of new HIV infections has fallen below the number of HIV-related deaths.” In 2012, the state had an estimated 3000 new infections and 1653 AIDS deaths (with an ongoing debate about how many of those were HIV related). The ETE blueprint aims to reduce the number of new infections to 750 in 2020.

New York has more HIV-infected residents, about half of whom are MSM, than any state. Out of a population of nearly 20 million, an estimated 154,000 people were living with HIV in 2012. Of those, 132,000, or 89%, knew their status, but only 68,000, or 44%, were on treatment and had undetectable levels of virus, far below the UNAIDS target. To decrease the gap between diagnosis and effective treatment, which Harrington refers to as “the Grand Canyon,” the blueprint calls for stepping up surveillance to find people who know their status but don’t consistently use ARVs and to assist them with nonmedical challenges that might get in the way of treatment, such as finding housing and jobs. It will also provide incentives to both care providers and patients to help people stay on treatment.

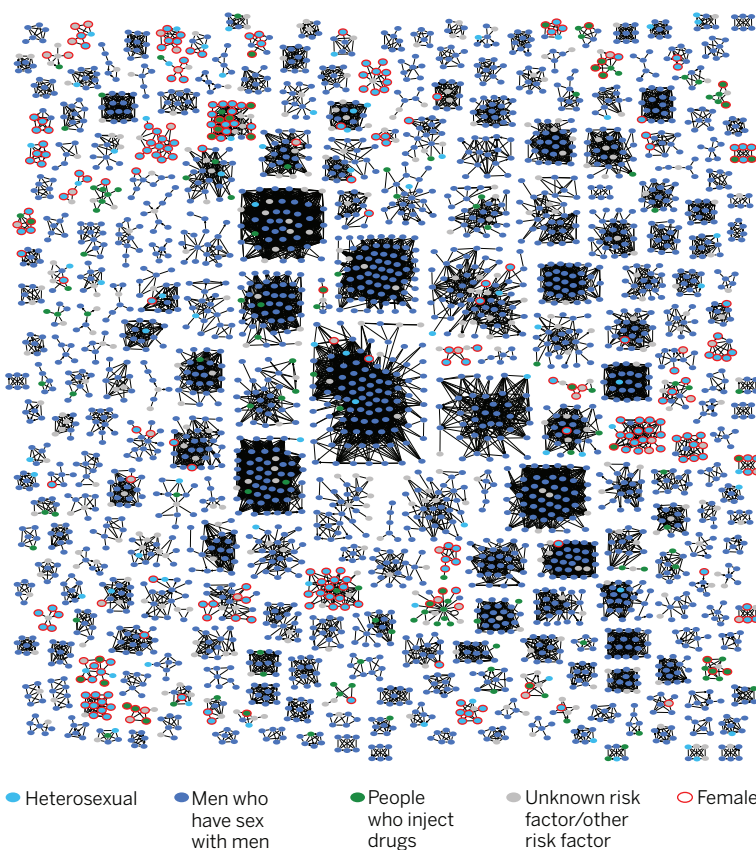
Meanwhile, 22,000 New Yorkers don’t even know they are infected. To reach them, the state will ramp up routine testing in places like emergency rooms and send out more mobile testing units to high-risk populations, including MSM, transgender people, new immigrants, the homeless, and the mentally ill. ETE is launching a statewide education campaign about PrEP—which has been approved by the U.S. Food and Drug Administration for 3 years but has been slow to catch on, in part because some uninfected people worried that it would brand them as promiscuous and reckless. New York is also creating programs to help people access PrEP drugs. Medical record “autopsies” of

people who die from AIDS will try to identify how they slipped through the cracks.

The state’s portion of the price tag for putting 68,000 more people on treatment by 2020 is \$2.25 billion, according to a recent analysis done by Treatment Action Group and Housing Works. Additional housing support for the 12,000 or so HIV-infected people who are homeless or poor would add as much as \$720 million to the bill. But improved care and decreases in new infections would reduce the load on the health care system by enough to save New York nearly \$8 billion, the analysis finds.

New York City’s hot spots of spread

With genetic sequences of HIV from different people, researchers can identify how the virus moves between individuals (circles). Mapping these transmission clusters can help target treatment and prevention efforts.



ON WORLD AIDS DAY in 2013, Diane Havlir of the University of California, San Francisco (UCSF), moderated a forum for the local community entitled “Getting to Zero in San Francisco: How Close Are We?” After experts described their city-wide efforts, Havlir, a clinician who heads the HIV/AIDS division at San Francisco General Hospital, invited questions from the audience.

“You guys are all doing great things,” a man said. “Are you working together?”

Havlir was taken aback. “At that moment we recognized that we were really *not* working together,” she said recently. That discussion sparked the Getting to Zero Coalition, which bands together UCSF, the city’s health department, HIV/AIDS advocacy groups, and major hospitals. In January, the coalition released a strategic plan for ending the city’s epidemic.

Although San Francisco has just 20% the population of BC, it is home to more HIV-infected people—an estimated 15,901 at the end of 2013, an adult prevalence of about 2%. One of the first cities to be hard-hit by

HIV, San Francisco early on developed a strong response by combining the research and clinical resources of UCSF and the local health department. The city’s epidemic surveillance is sophisticated and granular, including mapping of viral load levels by neighborhoods, which reveals where people are not controlling their infections. In 2010, San Francisco became the first jurisdiction to recommend immediate treatment for all HIV-infected people. Three years later, San Francisco General Hospital went a step further and launched a RAPID program that starts people on treatment the very day they are diagnosed, if possible. It also sends outreach teams to the streets to search for patients who miss appointments (*Science*, 13 July 2012, p. 175). Researchers at UCSF and the health department pioneered PrEP studies.

Exceptional as San Francisco’s response has been, it also has serious shortcomings, which the strategic plan confronts with three “signature initiatives”: scale up RAPID citywide,

teach people about PrEP and make it easy to get—à la Shane Ryan’s visit at Magnet—and work more diligently to keep people taking their ARVs. Testing is a not a big issue, as some 94% of infected people in the city know their status. But whereas 89% of those people see a specialist within 90 days, only 63% start on ARVs and achieve undetectable viral loads for prolonged periods. The Getting to Zero Coalition hopes to end AIDS by bringing that 63% up to 90% by 2020—an even more ambitious goal than the 73% set



Pedro Robles died from AIDS in a Tijuana, Mexico, hospice in 2013 without ever receiving treatment.

No end in sight

By Jon Cohen, in Tijuana, Mexico

On 6 December 2013, Pedro Robles spent 14 hours in an ambulance being driven up Mexico's Baja California Peninsula. The 51-year-old man was not rushed north for emergency medical care. Time was not of the essence.

Robles had an advanced case of AIDS, and he was being driven 1127 kilometers north from his home in Loreto to Albergue Las Memorias A.C. in Tijuana. A nongovernmental organization (NGO) arranged the trip, because Las Memorias is the only AIDS hospice on the entire Baja California Peninsula, Robles was broke, and Tijuana held out the remote hope that someone there could navigate

the medical bureaucracy and maybe save his life. But Las Memorias itself, which also serves as a drug rehabilitation center and is largely run by its residents, has no trained medical staff. And although Las Memorias did what it could to make Robles comfortable, Tijuana ultimately failed him: He died 6 days later without ever having seen a doctor.

As the dream of ending AIDS catches hold in a growing number of locales (see main story, p. 226), Tijuana is hardly anomalous: Many places are still struggling to provide basic treatment and prevention services. Of course, people still die from AIDS in wealthy countries like the United States, which is visible from downtown Tijuana, but appropriate care is so readily available that AIDS hospices shut their doors years ago.

Like Tijuana, too many locales appear to be “running at a standstill” and are saddled

by “poor strategy, absence of leadership, or inadequate resources,” laments a prominent commission in a report in *The Lancet* last month, “Defeating AIDS—advancing global health.”

The drop in HIV infections and AIDS-related deaths worldwide over the past dozen years has been impressive, the report says. But without a “massive and rapid expansion of a comprehensive AIDS response,” the global toll—still more than a million new infections and deaths each year—will worsen again over the next 5 years, and the world will fail to reach the United Nations goal of “ending AIDS as a public health threat” by 2030.

Mexico is not particularly hard-hit by HIV. In 2014, UNAIDS estimated the country had 190,000 infected people, which is an adult prevalence of 0.2%—lower than in the United States. The government offers free antiretrovirals (ARVs) and, since

PHOTO: © MALCOLM LINTON

November 2014, has recommended that all HIV-infected people receive them as soon as diagnosed. “When you look back to what we were doing 10 years ago, we are really, really better,” says Carlos Magis-Rodríguez of the National Center for HIV/AIDS Prevention and Control in Mexico City. But he acknowledges that Tijuana and other cities in Mexico are struggling.

Tijuana has what is known as a “micro-hyperepidemic.” Overall HIV prevalence in Tijuana is 0.6%—the same as the United States. But the rate is soaring in high-risk groups. In women who sell sex, prevalence jumped from 2% in 2003 to 6% by 2012, according to recent studies by researchers from the University of California, San Diego (UCSD), whose team includes Mexico-based colleagues. Clients of these sex workers had a prevalence of 5%, they found. HIV prevalence is also about 5% in the many people in Tijuana who inject drugs. Preliminary studies of men who have sex with men and transgender people suggest about 20% are infected.

In theory, Tijuana should be able to rein in its concentrated epidemic by taking advantage of recent advances. Key among them is the 2011 demonstration that people who fully suppress their HIV levels with ARVs rarely spread the virus to their sexual partners. But this treatment-as-prevention strategy has not gained much traction in Tijuana. UCSD behavioral health scientist Laramie Smith recently pooled data from six studies of nearly 200 HIV-infected people in Tijuana and found that only about half even knew they had the virus. Tijuana offers free HIV testing through NGOs and government-funded clinics, yet no plan is in place to regularly test high-risk people at venues where they hang out, like gay bars or the red light district.

Those who do learn they’re infected rarely get treatment. In Smith’s study, only 11% received related medical care, and only 3.66% began taking ARVs. The federally sponsored HIV/AIDS clinic, CAPASITS, provides free ARVs, but it is located far from the downtown area and is difficult for many people to reach. Tijuana is also a hub for migrants, including many deportees from the United States, and some do not have the documents required to receive help at CAPASITS. And the services there fall short of those in developed countries: CAPASITS, for example, must ship

patient blood samples to Mexico City for measurements of CD4 lymphocyte counts and HIV levels.

José Luis Burgos, a Tijuana-based clinician who works with the UCSD team, says a key problem is that the patient load in Tijuana outstrips the availability of qualified HIV/AIDS doctors. Burgos contends that Tijuana could train primary care physicians to diagnose and treat HIV/AIDS patients. “You need to demystify HIV care,” he says. CAPASITS, he notes, has only three doctors and treats some 1000 patients. “What kind of care can you expect from three providers?”

Paradoxically, Mexico’s rising economic status is hampering the fight. A 2011 grant from the Global Fund to Fight AIDS, Tuberculosis and Malaria enabled two Tijuana NGOs to launch mobile needle-exchange units. But the grant ended in 2013 when Mexico achieved upper middle income status and became ineligible for Global Fund support. Tijuana’s needle-exchange programs shriveled overnight. “Mexico is supposed to be upper middle income, but the border isn’t,” Burgos says.

The UCSD team soon documented a

“Mexico is supposed to be upper middle income, but the border isn’t.”

José Luis Burgos,

University of California, San Diego

40% increase in needle-sharing among a group of users that it has closely followed. That undermines other efforts to reach people with HIV, says UCSD epidemiologist Steffanie Strathdee, who leads the binational research program with her psychologist husband Thomas Patterson. “The sad thing” is when drug users come in for needle exchange, she says, “they have an opportunity to get HIV testing or a referral to a drug treatment program.”

Strathdee hopes the group’s extensive research will draw attention to the problems and the opportunities in Tijuana. “It’s entirely possible to end the AIDS epidemic in Tijuana,” she says. And if Tijuana can do it, so can much of the rest of the world. ■

Reporting for this story was supported in part by a grant from the Pulitzer Center on Crisis Reporting.

by 90-90-90. The city’s ending AIDS targets also include reducing new HIV diagnoses from 371 in 2013 to 37 in 2020 and HIV-related deaths from an estimated 91 to eight.

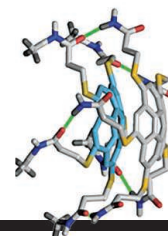
“The biggest challenges are reaching into marginalized populations that are not getting the services they need,” says the health department’s Buchbinder—transgender people, drug users, African-Americans, and the young. Getting to Zero is massively expanding the use of “navigators” assigned to infected people to help address barriers that interfere with care, such as substance abuse, food insecurity, homelessness, and violence. A new program will better coordinate medical records at different providers to help identify patients who are slipping through the net. And HIV-infected patients deemed at high risk of not taking ARVs will receive extra check-in phone calls and reminders for appointments.

Rigorous evaluation is critical for the campaign to succeed, said Havlir at a coalition meeting in June. Getting to Zero is raising money specifically to track the program’s performance, using novel metrics like assessing the impact of PrEP on new infection rates. “This is not just talking, talking, talking,” Havlir said. “This is about action.”

THE DRIVE TO END AIDS is spreading worldwide, and there is even something of a good-natured race to be first. Washington, D.C., New South Wales in Australia, and Brazil are now in the running as well, and San Francisco has attracted intrigued delegations from Amsterdam, France, and the White House’s Office of National AIDS Policy.

From his office at the London School of Hygiene & Tropical Medicine, which he directs, epidemiologist Peter Piot is watching these efforts with interest—and some skepticism. “It’s very important that these projects proceed and that we learn from them,” says Piot, who chaired the “Defeating AIDS” commission and formerly headed UNAIDS. But he cautions that the intensive efforts in BC, New York state, and San Francisco must continue indefinitely. “These three examples are not North Korean types of islands—there will be constant reintroduction of the virus,” he notes.

Ending the global spread of HIV will ultimately take a vaccine, Piot says, stressing that treatment as prevention packs a limited punch. He points to a mathematical model in “Defeating AIDS” that found that even if the world achieves the UNAIDS 90-90-90 goal in 2030, hundreds of thousands of new HIV infections and deaths will still occur each year. “We’ll have to see whether these three places can end AIDS as a public health threat,” Piot says. “But that doesn’t mean one shouldn’t try.” ■



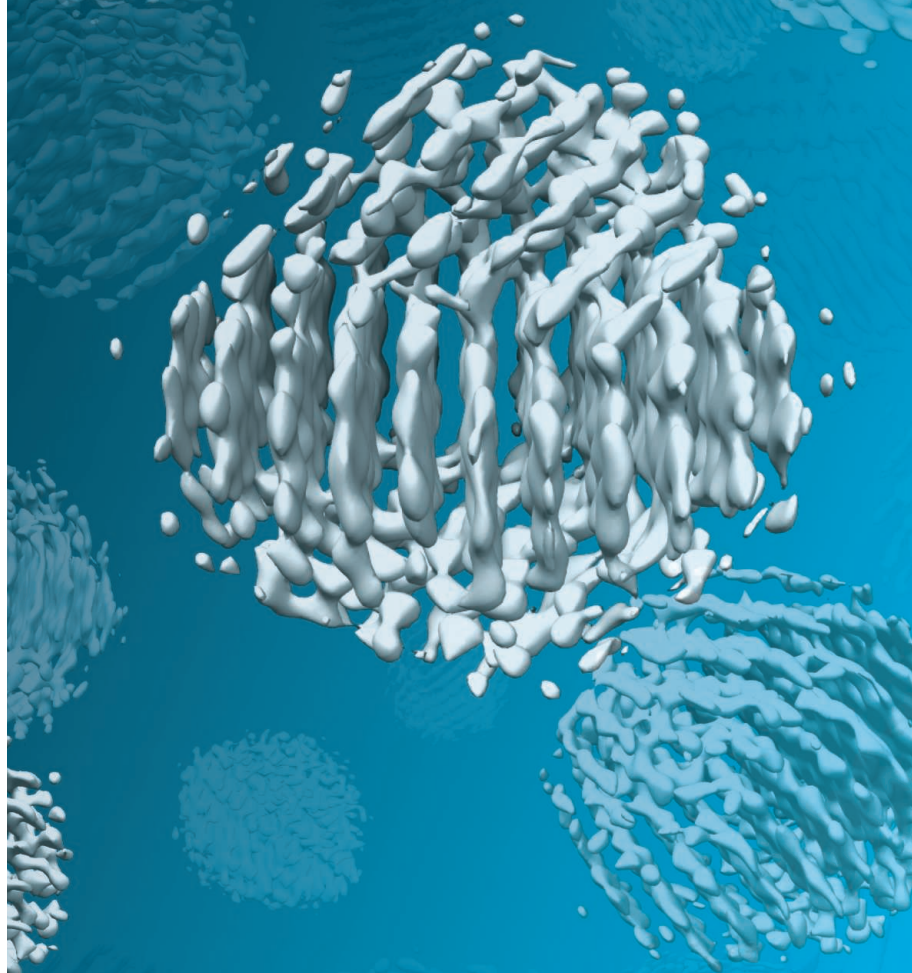
PERSPECTIVES

CHEMISTRY

Tracking the merry dance of nanoparticles

Electron microscopy provides atomic-resolution structures of nanoparticles in solution

By Christian Colliex



A triumph of electron microscopy. Park *et al.* use a specially equipped electron microscope to determine the atomic structure of individual platinum nanoparticles in solution.

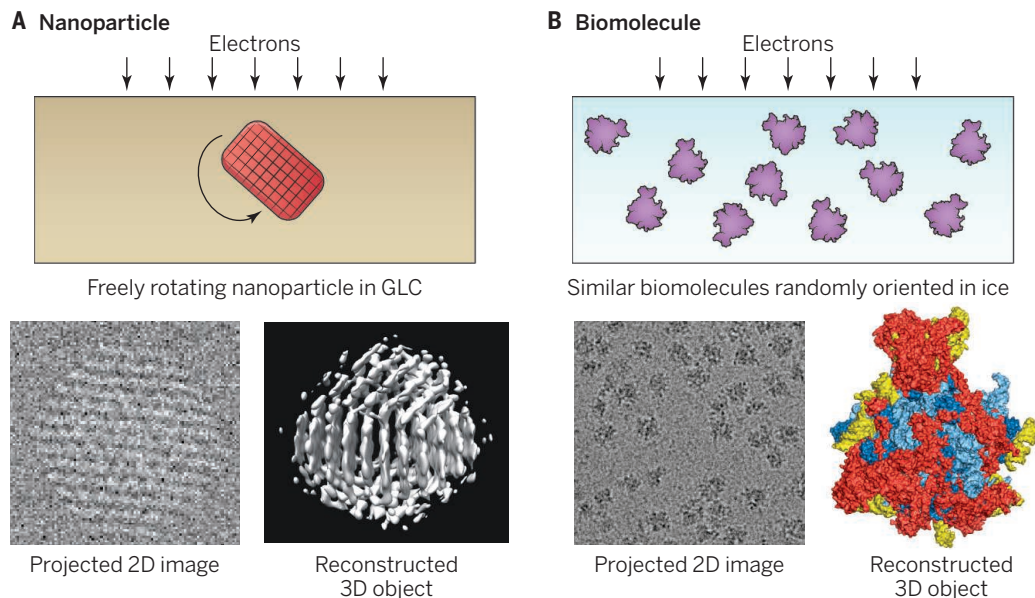
Nanoparticles containing a few tens or hundreds of atoms have applications in an ever-expanding range of areas, from electronics to catalysis and biological sensors. This versatility stems from the high sensitivity of nanoparticle properties to size, chemical composition, and shape, which are largely determined by the synthesis route by which they are produced. Nanoparticles grown in solution, termed “colloidal nanoparticles,” are of particular interest because they may be used in many applications, including as biosensors (1). On page 290 of this issue, Park *et al.* (2) report high-resolution structures of nanoparticles in solution deduced from electron microscopy. This spectacular result relies on the convergence of recently developed techniques from both the physical and life sciences and heralds a new era of high-resolution studies in solution.

Colloidal particles made of noble metals, particularly gold or platinum, have been widely studied for applications in catalysis and photocatalysis. As early as 20 years ago, Ahmadi *et al.* showed that the shape of colloidal Pt nanoparticles could be controlled by changing the composition and concentration of capping materials and metal cations in the solution (3). To image the particles, they deposited nanoparticles on an amorphous carbon layer. By tilting the resulting specimen, they found that nanoparticles with square shapes were cubic and nanoparticles with triangular shapes were tetrahedral. However, they could not resolve the three-dimensional (3D) atomic structure or study the particles directly in solution. This is what Park *et al.* now achieve for individual nanocrystals. The work follows similar progress realized a few years earlier in a gaseous environment (4).

The success of the study by Park *et al.* relies on three technological developments. The graphene liquid cell (GLC), introduced previously by the same group (5), maintains nanometer-scale objects in a liquid environment between two layers of free-standing graphene. Highly sensitive direct electron detectors enable the production of movies with about 50 frames per second. To analyze their data, the authors adapted the theory and associated software tools created by molecular biologists to reconstruct the structures of single biomolecules.

In the approach used by Park *et al.*, movies follow individual Pt nanocrystals freely

rotating in solution (see the figure, panel A). These nanoparticles may thus be observed in many different orientations, providing the required number of different projections for the reconstruction algorithm. In molecular biology (see the figure, panel B), multiple copies of biomolecules with identical structures are flash-frozen in a vitrified thin layer of ice (6). One micrograph covers a field with hundreds of particles viewed at different orientations. These images of individual molecules yield a collection of 2D projections that are then used to generate a single 3D structure.



Three-dimensional structures from electron microscopy. (A) Park *et al.* take many images of an individual platinum nanoparticle as it rotates freely in the solution contained in a GLC. They then combine these images of the same particle to determine its 3D structure. The method was adapted from cryo-EM (B), where 2D projections of numerous copies of the same biomolecule frozen in a vitrified ice layer are used to determine the 3D structure of the biomolecule (15).

Yuk *et al.* first used a GLC to image with atomic resolution a few critical steps (coalescence, reshaping, and faceting) in the growth of Pt nanocrystals (5). Since then, GLCs have been used to follow the real-time trajectories of Au nanocrystals linked to double-stranded DNA fragments. The observations confirmed that sample-substrate interactions between the liquids and the inert graphene layers are minimized (7). The authors then deduced the 3D motion from a series of 2D images recorded while the complex nanostructure was rotating. This was clearly a key step on the way to the present demonstration. Recently, Wang *et al.* used GLCs to encapsulate tiny water volumes containing ferritin molecules, which consist of a core of iron (hydro)-oxide and a protein shell (8). They then used elec-

tron energy-loss spectroscopy (EELS) to map the spatial distribution of Fe, O, and N elements and also reported an analysis of the iron valence state in ferritin and in solution. The results show that GLCs provide a potentially useful route for monitoring biochemical reactions in situ.

As for the latest generation of direct electron-detection camera technology, it has contributed to the rapid improvement of high-resolution structures reconstructed from cryo-electron microscopy (cryo-EM) data. Their signal-to-noise ratio is improved in the movie frame mode, allowing refined

nanocrystals can generally be considered as rigid, as long as the electron beam does not cause any transformations.

The reconstruction of 3D atomic structures of individual metallic nanoparticles in a vacuum or embedded in a solid matrix has benefited from advances in electron microscopy techniques, such as aberration correction for higher-resolution images, annular dark-field scanning transmission mode for counting the number of atoms in each column, and the development of dedicated algorithms for tomography. Successful examples include the 3D atomic structures of a gold nanoparticle in vacuum (11, 12), a silver nanocrystal epitaxially embedded in an aluminum matrix (13), and a gold nanoparticle of 68 atoms surrounded by a thiol molecular cage that makes it water soluble (14). Park *et al.* now investigate two distinct individual nanoparticles in the liquid phase. Their density maps reveal that each particle consists of three nanocrystals joined together nonsymmetrically. The authors analyze the local organization of the interfaces between these nanocrystals and compare the results with molecular dynamics calculations of the free energies for grain boundary formation.

As Park *et al.* show, the structure and dynamics of individual members of an inhomogeneous ensemble of nano-objects in a fluid can now be determined with specifically equipped modern electron microscopes. This advance provides a tool for exploring the growth, self-organization and reactivity of nanoparticles in liquids containing different types of ions, salts, or molecules. ■

REFERENCES

1. P. D. Howes, R. Chandrawati, M. M. Stevens, *Science* **346**, 1247390 (2014).
2. J. Park *et al.*, *Science* **349**, 290 (2015).
3. T. S. Ahmadi, Z. L. Wang, T. C. Green, A. Henglein, M. A. El-Sayed, *Science* **272**, 1924 (1996).
4. J. R. Jinschek, *Chem. Commun. (Camb.)* **50**, 2696 (2014).
5. J. M. Yuk *et al.*, *Science* **336**, 61 (2012).
6. D. Agard, Y. Cheng, R. M. Glaeser, S. Subramaniam, *Adv. Imaging Electron Phys.* **185**, 113 (2014).
7. Q. Chen *et al.*, *Nano Lett.* **13**, 4556 (2013).
8. C. Wang *et al.*, *Adv. Mater.* **26**, 3410 (2014).
9. M. G. Campbell *et al.*, *Structure* **20**, 1823 (2012).
10. A. Bartesaghi *et al.*, *Science* **348**, 1147 (2015).
11. Z. Y. Li *et al.*, *Nature* **451**, 46 (2008).
12. M. C. Scott *et al.*, *Nature* **483**, 444 (2012).
13. S. Van Aert, K. J. Batenburg, M. D. Rossell, R. Erni, G. Van Tendeloo, *Nature* **470**, 374 (2011).
14. M. Azubel *et al.*, *Science* **345**, 909 (2014).
15. A. Amunts *et al.*, *Science* **343**, 1485 (2014).

PHYSICS

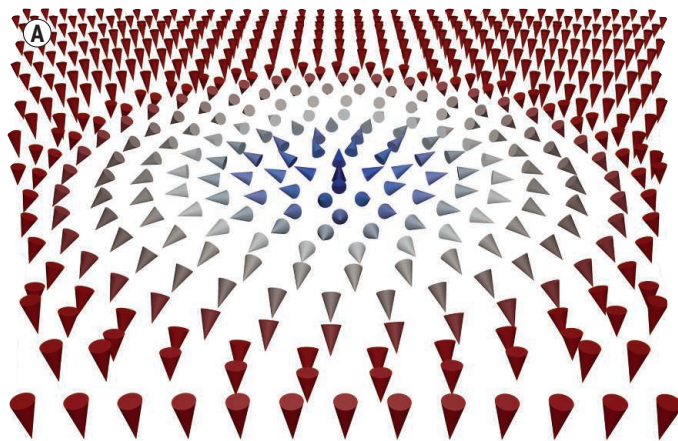
Magnetic bubbles with a twist

Individual skyrmionic bubbles can be generated and moved at room temperature

By Kirsten von Bergmann

Present-day hard disk drives use magnetic bits that are read out by movable read heads. To circumvent the inherent fragility of such a mechanical construction, and to exploit the third dimension for increased storage density, Parkin *et al.* proposed a racetrack memory device (1). The general concept is that the information is encoded in a localized magnetization configuration that can be driven through the material with electrical currents and transported to a stationary read head. Recently, interface-induced skyrmions, which are circular particle-like magnetic objects (see the figure, panel A), have been envisioned as ideal candidates for future racetrack memory-type applications (see the figure, panel B) (2, 3). On page 283 of this issue, Jiang *et al.* (4) have made progress toward realizing such a device architecture. They report on the generation and movement of individual skyrmionic bubbles at room temperature, accomplished by exploiting two different spin-orbit coupling-related effects.

Department of Physics, University of Hamburg, 20355 Hamburg, Germany.
E-mail: kbergman@physnet.uni-hamburg.de



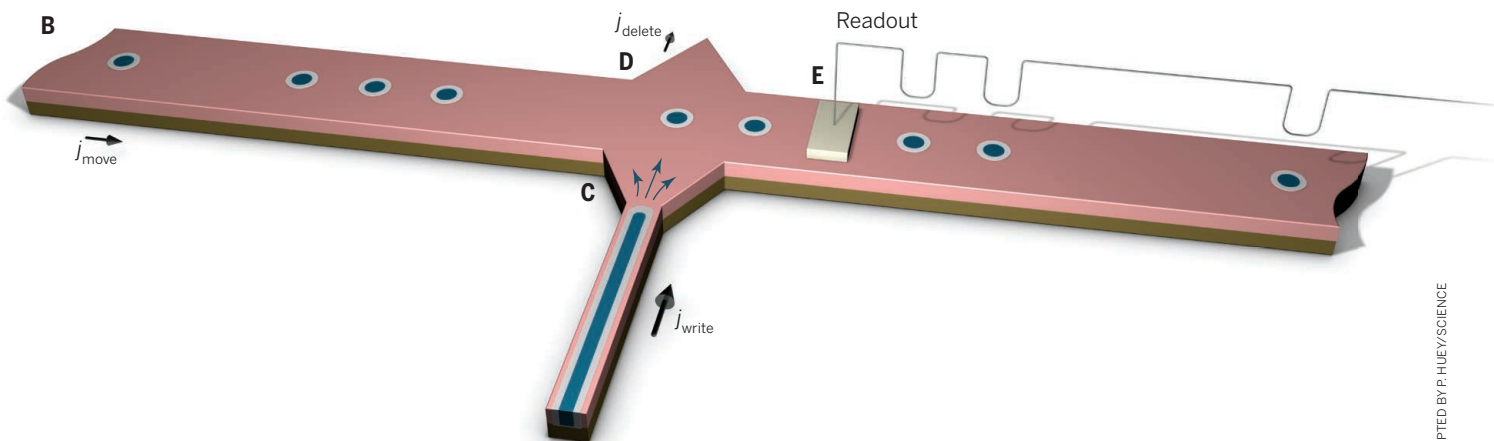
Magnetic skyrmions are localized non-collinear states with a twist, possessing the key characteristic that they are topologically protected. That is, they effectively constitute a knot in the magnetic texture that adds to their stability with respect to a trivial ferromagnetic state. The application of spin transfer torques by the injection of lateral or vertical electrical currents can induce a motion of magnetic skyrmions at surprisingly low current densities (5, 6). For such topologically distinct states, the direction of motion depends on the direction of the twist, manifested in the rotational sense of the spin canting.

Twisted magnetic states are useful for applications only if the direction of the twist is unique, thus enabling control over

the movement via the direction of the injected current. To achieve this, the twist must be intrinsic to the system. A magnetic interaction that favors one rotational sense over the other is the Dzyaloshinskii-Moriya (DM) interaction, which arises as a result of spin-orbit interaction and can occur in systems with broken inversion symmetry as, for example, in chiral crystal structures (7, 8) or at interfaces (9–11). For the interface-induced skyrmions, the rotational sense is determined by the specific

material, and when realized on a large scale by standard epitaxial growth processes, it promises an easy implementation into present state-of-the-art multilayer-based device architectures.

Until now the only experimental observations of interface-induced skyrmions were in perfectly ordered, atomically thin layers, which served as model systems (10, 11). Tuning of the skyrmion size from 1 to 5 nm was demonstrated. By using local currents from a scanning tunneling microscope tip, isolated skyrmions were written and deleted individually (11). Future applications will rely on other, more practical materials in which the DM interaction governs the magnetic texture, inducing a defined twist of the magnetic state.



Skyrmions for future storage devices. (A) Sketch of a magnetic skyrmion, a knot in the magnetization with a defined twist. (B) Coherent movement of skyrmions along a track with lateral currents inspired by (3). (C) Generation of individual skyrmions by inhomogeneous spin-orbit torques (small blue arrows) behind a geometrical constriction, as demonstrated by Jiang *et al.* (4). (D) Annihilation of skyrmions by small currents that drive the magnetic object out of the magnetic material. (E) Stationary read head; the information is encoded in the magnetic texture, which is moved to the readout by small lateral currents along the track.

Jiang *et al.* report on the observation of interface-induced skyrmionic bubbles at room temperature in a magnetic layer grown epitaxially on a material with large spin-orbit coupling. The role of the substrate is twofold: First, it provides the interface at which the DM interaction is induced; and second, by driving a lateral current through the substrate, the charge is converted to a vertical spin current due to the spin Hall effect (12). The resulting spin-orbit torques act on the magnetic layer and induce a movement of the micrometer-sized skyrmions. All of them move in the same direction, which means they all have the same twist, imposed by the DM interaction.

The generation of individual skyrmions is realized by inhomogeneous spin-orbit torques occurring behind a geometric constriction (see the figure, panel C). Jiang *et al.* propose an intriguing model of how the bubbles are pinched off from a magnetic stripe domain driven through the constriction, and draw an analogy to the formation of soap bubbles. Thus, for the writing of a skyrmion into a track, a perpendicular geometry could be used, and with a lateral movement in the same direction, it should be possible to delete single magnetic objects by moving them out of the track (see the figure, panel D), unifying all relevant operations in one device.

With the goal of achieving a skyrmion racetrack in sight, there are several issues that still need to be resolved. The precision of the generation and the degree of control over each magnetic skyrmion need improvement. Jiang *et al.* demonstrate the generation of many skyrmionic bubbles at a time, and in their sample the movement is strongly influenced by local pinning due to the inhomogeneity of the film. Furthermore, the size of the skyrmions needs to be reduced to be competitive with present storage densities. Future research will show whether nanometer-sized skyrmions are possible in materials relevant for applications. However, such a small size of a skyrmion will not only limit the methods to investigate them but also challenge the mechanism and design of an appropriate readout device (see the figure, panel E). ■

REFERENCES

1. S. S. P. Parkin, M. Hayashi, L. Thomas, *Science* **320**, 190 (2008).
2. N. S. Kiselev *et al.*, *J. Phys. D* **44**, 392001 (2011).
3. A. Fert *et al.*, *Nat. Nanotechnol.* **8**, 152 (2013).
4. W. Jiang *et al.*, *Science* **349**, 283 (2015).
5. F. Jonietz *et al.*, *Science* **330**, 1648 (2010).
6. J. Sampaio, V. Cros, S. Rohart, A. Thiaville, A. Fert, *Nat. Nanotechnol.* **8**, 839 (2013).
7. S. Mühlbauer *et al.*, *Science* **323**, 915 (2009).
8. X. Z. Yu *et al.*, *Nature* **465**, 901 (2010).
9. M. Bode *et al.*, *Nature* **447**, 190 (2007).
10. S. Heinze *et al.*, *Nat. Phys.* **7**, 713 (2011).
11. N. Romming *et al.*, *Science* **341**, 636 (2013).
12. L. Liu *et al.*, *Science* **336**, 555 (2012).

ECOLOGY

Is biodiversity good for your health?

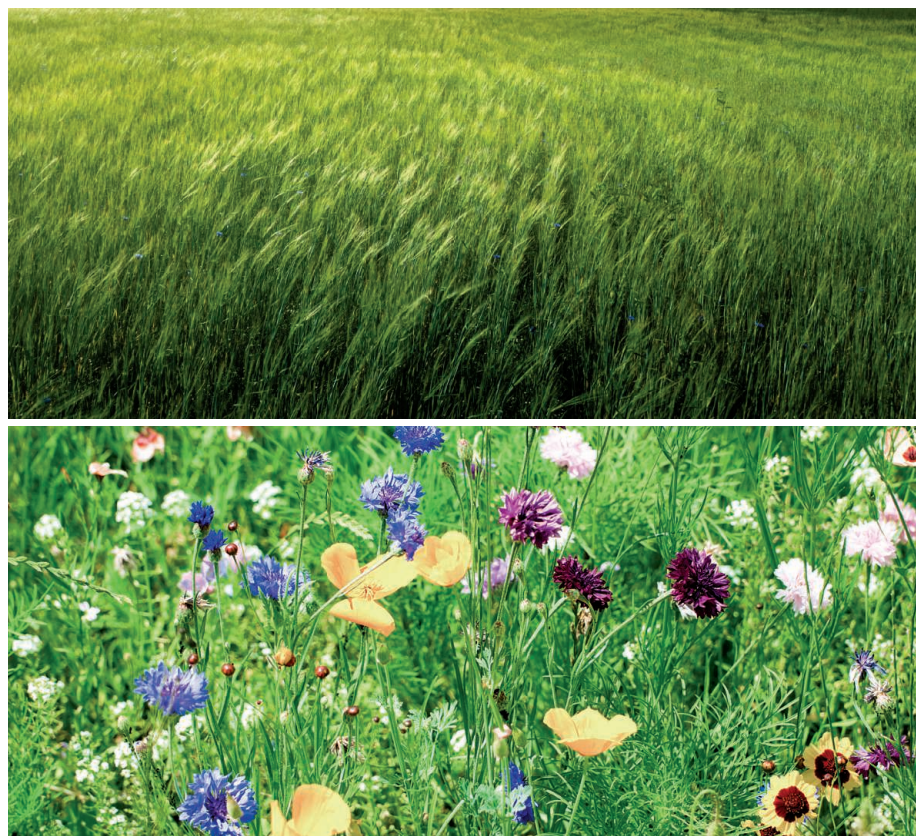
Disease incidence is often lower in more diverse communities of plants and animals

By Felicia Keesing and Richard S. Ostfeld

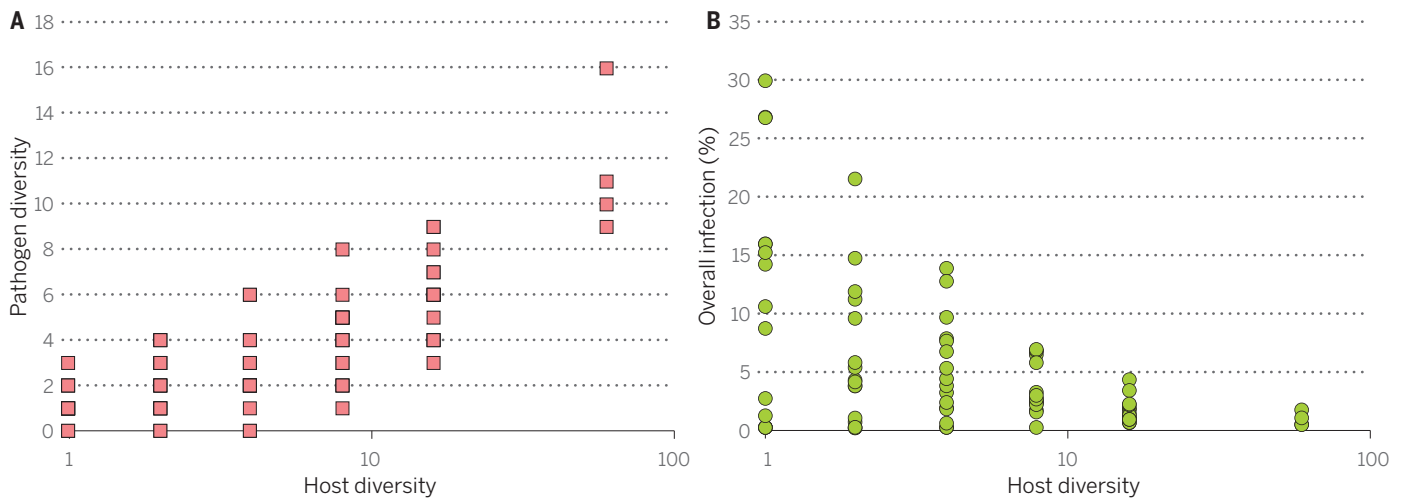
On a floodplain of the River Saale near Jena, Germany, grassland plants are naturally bombarded by spores of pathogenic fungi. But whether or not those fungi cause infection turns out to be largely about the neighborhood: Plants on highly diverse experimental plots have much lower levels of infection than plants grown in monoculture (1) (see the photos). The pathogens, it appears, are less likely to encounter their optimal host on the more diverse plots, which reduces disease prevalence and incidence. This protective effect of diversity has been found in many studies, not just for plants but also for diseases afflicting humans and wildlife. It has

remained unclear, however, whether this observation holds generally (2, 3). In a recent paper, Civitello *et al.* addressed this question in a rigorous meta-analysis of diversity-disease relationships (4).

Civitello *et al.* examined 202 results from studies of 61 parasites infecting both plants and animals. They found a clear and consistent reduction in disease prevalence when diversity was high. The effect was as strong for parasites that infect humans as for those that infect only wildlife and was seen both for microparasites (pathogens such as bacteria and viruses) and for macroparasites (such as trematodes and nematodes). The protective effect of diversity for human diseases is particularly important because it refutes the results of a previous meta-analysis



The benefits of diversity. Civitello *et al.* have performed a meta-analysis of studies of 61 parasites infecting both animals and plants. The results show that disease prevalence is often higher in less diverse systems (top, barley monoculture) than in more diverse systems (bottom).



Effects of plant diversity on fungal pathogens in Jena, Germany. Plots with higher host diversity generally had a higher diversity of pathogens (A) but also lower levels of overall infection (B). Drawn from data provided in (1). Note that the x-axis is on a log scale.

of only six parasites, which found no consistent effect of diversity (5). Civitello *et al.* also found that the effect of diversity was equally strong for both experimental and comparative studies. This is a key result, because although it has been known for some time that diversity can reduce disease in carefully controlled and constructed experimental communities, it has been less clear whether the effect translates into natural systems.

The widespread negative effect of high diversity on disease transmission documented by Civitello *et al.* suggests a consistent underlying mechanism that applies to diseases of humans, wildlife, and plants. One part of that mechanism has been hinted at in previous work: The best hosts for multihost pathogens are often abundant, widespread, and resilient species (6–8). A recent study by Han *et al.* (9) provides strong evidence for this pattern. The authors set out to determine what characteristics make some species good reservoir hosts of diseases that infect humans. Working with a database of 2277 rodent species and 66 pathogens, they used machine-learning algorithms to explore what aspects of host life history, physiology, behavior, ecology, and biogeography are associated with pathogens that spill over into human populations. Several patterns emerged. Good reservoir hosts have broad geographic ranges that contain comparatively few other species. They also reach maturity at a young age and have a short gestation period and a large litter size, all characteristics of what is often called a fast life history.

Why might species with fast life histories be good hosts for pathogens? Two hypotheses have garnered some support. First,

parasites and pathogens that infect multiple species might evolve to exploit those hosts that they are most likely to encounter in nature—in other words, the most widespread, abundant hosts. Species that are widespread and abundant often have fast life histories (10). The second possibility is that fast-living hosts tend to be less resistant to, or more tolerant of, infection because of how they allocate their immune defenses (11).

Whatever the causes of these patterns, the consequences are intriguing. Species with fast life histories are often resilient to disturbance and are thus likely to persist in the face of environmental stressors that cause other species to decline or disappear (12). Together, these two patterns—the resilience and ubiquity of species with fast life histories and their odds of being high-quality hosts for parasites and pathogens—may interact to produce the diversity-disease relationship. As biodiversity is lost from ecological systems, the species most likely to persist may tend to be those most likely to harbor and transmit pathogens at high rates.

But won't ecological communities with a higher diversity of hosts also have a higher diversity of parasites and pathogens? And might a higher diversity of parasites and pathogens counteract the disease-suppressing effects of high host diversity? These questions have been at the heart of critiques of the diversity-disease relationship (3, 13). An answer may be emerging. Back on the German floodplain, Rottstock *et al.* (1) found that the diversity of fungal pathogens was indeed higher in the plots with higher plant diversity, but the level of infection was lower (see the chart). Even the prevalence of co-infection—the simultaneous infection of a host with more than one pathogen—was lower where host and pathogen diversity were high. If this pattern of reduced disease

with increased parasite diversity occurs in other systems, it would help to resolve what some have seen as a paradoxical relationship between diversity and disease.

Many fascinating questions remain about the relationship between diversity and disease, but Civitello *et al.*'s meta-analysis demonstrates that diversity frequently reduces disease. At the same time, Han *et al.* (9) have identified traits in rodents that strengthen the connection between life history and the probability that a species will be a reservoir for human pathogens. Case studies to evaluate diversity effects in specific disease systems will remain important, as will the pursuit of mechanisms that underlie the general pattern. Perhaps the most important question that remains is about the application of this knowledge to public policy. Should health protection be added to the long list of ecosystem services provided by biodiversity (2)? The meta-analysis by Civitello *et al.* suggests that it should. ■

REFERENCES

1. T. Rottstock, J. Joshi, V. Kummer, M. Fischer, *Ecology* **95**, 1907 (2014).
2. B. J. Cardinale *et al.*, *Nature* **486**, 59 (2012).
3. C. L. Wood *et al.*, *Ecology* **95**, 817 (2014).
4. D. J. Civitello *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1073/pnas.1506279112 (2015).
5. D. J. Salkeld, K. A. Padgett, J. H. Jones, *Ecol. Lett.* **16**, 679 (2013).
6. Z. Y. X. Huang *et al.*, *PLOS ONE* **8**, e54341 (2013).
7. J. P. Cronin, M. E. Welsh, M. G. Dekkers, S. T. Abercrombie, C. E. Mitchell, *Ecol. Lett.* **13**, 1221 (2010).
8. P. T. J. Johnson *et al.*, *Ecol. Lett.* **15**, 235 (2012).
9. B. A. Han, J. P. Schmidt, S. E. Bowden, J. M. Drake, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7039 (2015).
10. M. A. Previtelli *et al.*, *Oikos* **121**, 1483 (2012).
11. L. B. Martin 2nd, D. Hasselquist, M. Wikelski, *Oecologia* **147**, 565 (2006).
12. S. Laverne, M. E. K. Evans, I. B. Burfield, F. Jiguet, W. Thuiller, *Philos. Trans. R. Soc. Lond. Ser. B* **368**, 20120091 (2013).
13. H. Young, R. H. Griffin, C. L. Wood, C. L. Nunn, *Ecol. Lett.* **16**, 656 (2013).

¹Department of Biology, Bard College, Annandale-on-Hudson, NY 12504, USA. ²Cary Institute of Ecosystem Studies, Millbrook, NY 12545, USA. E-mail: keesing@bard.edu

Neutrophil-macrophage communication in inflammation and atherosclerosis

Neutrophils may license macrophages to respond to cholesterol crystals and drive inflammation that aggravates atherosclerosis

By **Matthias Nahrendorf**
and **Filip K. Swirski**

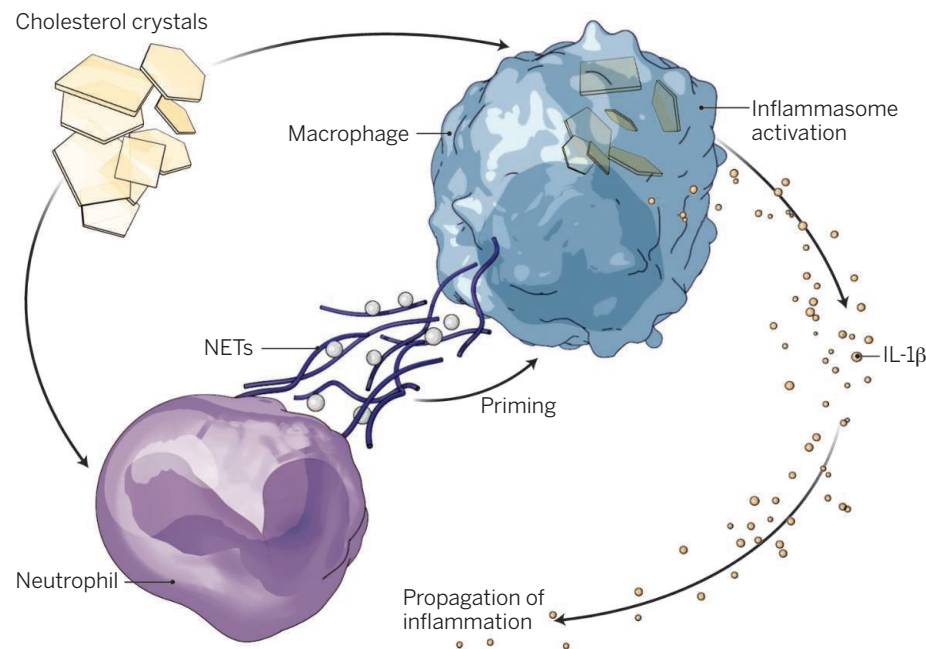
Neutrophils and monocyte-derived macrophages are myeloid cells, with a shared hematopoietic ancestry, that pursue both common and distinct immune functions. In response to tissue damage or infection, neutrophils infiltrate first, followed by monocytes (1). Through a series of events involving the release of proinflammatory products, both cell types seek to neutralize danger, and if successful, inflammation resolves. But when stimuli persist and inflammation is not resolved, neutrophils and monocytes continue to accumulate, presumably inflicting the irreversible damage that characterizes chronic inflammation. Whether neutrophils and monocytes act together to fuel the inflammatory cascade in those circumstances is still poorly understood. On page 316 of this issue, Warnatsch *et al.* (2) report that neutrophils prime macrophages for inflammatory responses that aggravate atherosclerosis. Neutrophil extracellular traps (NETs) underlie the communication between these two immune cell types.

A distinguishing feature of neutrophils is their ability to expel cytosolic and nuclear material, a suicidal act (NETosis) that ensnares extracellular microbes. NETs are web-like structures containing DNA, histones, neutrophil elastase, and myeloperoxidase (3) (see the second figure). According to Warnatsch *et al.*, NETs can be harmful in the setting of inflammatory atherosclerosis. In this case, extracellular cholesterol crystals interact with neutrophils to trigger the release of NETs, which prime macrophages to produce a precursor form of the inflammatory cytokine interleukin-1 β (pro-IL-1 β). In addition to activating NETosis, cholesterol crystals bind to the cell surface protein CD36 on macrophages, are internalized via endocytosis (4), and activate a signaling complex called the inflammasome (5). The inflammasome promotes maturation of

IL-1 β via cleavage of pro-IL-1 β with caspase-1. Thus armed, macrophages present in atherosclerotic plaques carry out inflammatory and plaque-destabilizing functions. Ultimately, these processes induce thrombotic complications that interrupt arterial blood supply to downstream tissues, causing myocardial infarction and stroke (6).

Warnatsch *et al.* identified NETs accruing adjacent to cholesterol crystals and macrophages in atherosclerotic lesions of apolipoprotein E-deficient (*ApoE*^{-/-}) mice. In this model of atherosclerosis, the absence of ApoE impairs reverse cholesterol transport, resulting in hypercholesterolemia, leukocytosis, and plaque formation in the vessel wall. Incubation of human neutrophils with cholesterol crystals *in vitro* triggered NETosis—that is, neutrophils died while expelling nuclear material, which includes the well-known danger signal DNA. To determine whether NETosis contributes to atherogenesis, the authors engineered *ApoE*^{-/-} mice that also lack neutrophil elastase and proteinase-3 (*ApoE*^{-/-}/*Ela2*^{-/-}/*Prtn3*^{-/-}), two serine proteases that

are secreted during inflammation to destroy pathogens and host tissue, and localize to NETs. These triple mutant mice could not support NETosis. When consuming a diet high in fat and cholesterol, *ApoE*^{-/-}/*Ela2*^{-/-}/*Prtn3*^{-/-} mice developed dramatically smaller atherosclerotic lesions compared to *ApoE*^{-/-} control animals, despite similar lipid concentrations and leukocyte counts in blood. In the triple mutant mice, the authors found no NETs, lower systemic IL-1 β concentration, and fewer lesional T cells that produce IL-17, a cytokine that propagates inflammation by promoting the expression of multiple inflammatory cytokines (although its role in atherosclerosis is not firmly established). Moreover, DNase treatment of *ApoE*^{-/-} mice resulted in the same small lesion phenotype as seen in the *ApoE*^{-/-}/*Ela2*^{-/-}/*Prtn3*^{-/-} mice, further supporting DNA's atherosclerosis-aggravating role. Warnatsch *et al.* also noted that incubation of human monocytes first with NETs and then with cholesterol crystals resulted



Fueling inflammation. Cholesterol crystals and NETosis collaborate to activate the inflammasome in murine atherosclerosis.

Center for Systems Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA. E-mail: mnahrendorf@mgh.harvard.edu; fswirski@mgh.harvard.edu

in strong inflammasome activation. The observations suggest a two-step inflammasome activation process (see the first figure) in murine plaque macrophages in which NETs prime a macrophage for inflammasome activation and pro-IL-1 β production. Cholesterol crystals then induce lysosomal damage, which is sensed by the macrophage inflammasome called NOD-like receptor family, pyrin domain containing 3

it will be imperative to determine whether the phenotype observed in *Apoe*^{-/-}/*Ela2*^{-/-}/*Prtn3*^{-/-} mice is exclusive to NETosis; other leukocytes, including macrophages, also express proteinase-3 (9). The high turnover of plaque macrophages (10) and the observation of NETosis in diabetes (11), a common risk factor for atherosclerosis, mean the striking phenotype reported by Warnatsch *et al.* could depend on additional mechanisms that still await discovery.

Warnatsch *et al.* introduce an interesting new facet to our understanding of how neutrophils and macrophages communicate. Macrophage activation through NETosis contrasts with neutrophil-macrophage interaction during resolution of inflammation. When macrophages engulf apoptotic neutrophils, a noninflammatory macrophage phenotype supports healing and a return to a steady state (1). Thus, the presence of cholesterol crystals dramatically influences cell-cell interactions and macrophage function. In considering possible therapies, inhibiting NETosis, or at least eliminating extracellular DNA, may prove beneficial. An alternative approach could reduce neutrophils by limiting their production in the bone marrow, migration, and/or recruitment. Both strategies would edit the instructions that macrophages receive from their tissue microenvironment, possibly curbing inflammasome activation and

promoting inflammation resolution. Further downstream in the pathway, inhibiting the active form of IL-1 β with a neutralizing antibody is a promising strategy currently being tested in patients with atherosclerosis (12). ■

REFERENCES AND NOTES

1. C. N. Serhan, J. Savill, *Nat. Immunol.* **6**, 1191 (2005).
2. A. Warnatsch, M. Ioannou, Q. Wang, V. Papayannopoulos, *Science* **349**, 316 (2015).
3. V. Brinkmann *et al.*, *Science* **303**, 1532 (2004).
4. F. J. Sheedy *et al.*, *Nat. Immunol.* **14**, 812 (2013).
5. P. Duewell *et al.*, *Nature* **464**, 1357 (2010).
6. F. K. Swirski, M. Nahrendorf, *Science* **339**, 161 (2013).
7. O. Soehnlein, *Circ. Res.* **110**, 875 (2012).
8. T. Quillard *et al.*, *Eur. Heart J.* **36**, 1394 (2015).
9. E. L. Gautier *et al.*, *Nat. Immunol.* **13**, 1118 (2012).
10. C. S. Robbins *et al.*, *Nat. Med.* **19**, 1166 (2013).
11. S. L. Wong *et al.*, *Nat. Med.* **10**, 1038/nm3887 (2015).
12. P. M. Ridker, T. Thuren, A. Zalewski, P. Libby, *Am. Heart J.* **162**, 597 (2011).

ACKNOWLEDGMENTS

We thank E. Latz for helpful discussions.

10.1126/science.aac7801

GLOBAL NUTRITION

Metrics for land-scarce agriculture

Nutrient content must be better integrated into planning

By Ruth DeFries,^{1*} Jessica Fanzo,² Roseline Remans,^{3,4} Cheryl Palm,³ Stephen Wood,^{1,3} Tal L. Anderman⁵

Over the past half-century, the paradigm for agricultural development has been to maximize yields through intensifying production, particularly for cereal crops (1). Increasing production of high-yielding cereals—wheat, rice, and maize—has replaced more nutrient-rich cereals, which has eroded the content of essential dietary nutrients in the world's cereal supply. New approaches are needed to produce healthy foods, rich in essential nutrients, with efficient use of land.

Standard yield metrics that measure the quantity of production are inadequate to assess progress toward this goal; thus, we propose alternative metrics of nutritional yields.

Intensification of agriculture through multiple crops per year, high-yielding seed varieties, synthetic fertilizer, mechanization, and other inputs increases yields and improves the efficiency of using land to produce food. In the last 50 years, intensification of cereal production has increased the world's cereal supply by a factor of almost 2.2, outpacing the 1.3-fold increase in population growth (2). Intensification was critical for averting food shortages that loomed on the horizon in the 1960s.

Intensification has spared 18 to 27 million hectares that would have been required to produce the same amount of cereals with yields equivalent to those in the mid-1960s (1). But intensification can exacerbate land-clearing in the absence of appropriate policies and enforcement (3). Moreover, intensification relies on high inputs of energy, fertilizer, pesticides, and water (4). Environmental consequences include runoff of excess fertilizer that damages water quality, toxicity from pesticides, nitrous oxide emissions, and degradation of habitat for biodiversity.

With increasing competition for land (e.g., for carbon storage, watershed protec-



Deploying NETs. Colored scanning electron micrograph of neutrophil extracellular traps (NETs, brown) capturing spores from the yeast *Candida albicans* (yellow).

(NLRP3). NLRP3 activates caspase-1, which subsequently leads to proteolytic cleavage and release of IL-1 β (5).

Cholesterol crystal-induced NETosis as a macrophage inflammasome stimulator adds to the growing knowledge on neutrophil activity in atherosclerosis, including the role of neutrophils in augmenting monocyte recruitment (7) and the pro-thrombotic actions of NETs in the setting of plaque erosion (8). Do these interactions occur in human atherosclerotic plaque? Although Warnatsch *et al.* provide compelling human in vitro data, and NETs are present on the surface of eroding human plaque (8), it will be important to verify that the observed murine phenotype translates to patients. Additionally, it is unknown if similar interactions between neutrophils and macrophages also occur in other organs, as these cells travel far and hypercholesterolemia is systemic. Hence,

the proportion of maize diverted for feed remained fairly stable over the time period, the change in the mix of cereal types had an even larger effect on reducing the nutritional content of the global, directly consumed cereal supply than the diversion to feed (table S1 and fig. S4).

In other words, the amount of cereals that a person would need to consume to fulfill the daily dietary reference intake (DRI) has increased for protein, iron, and zinc, based on a mix of cereals in proportion to the production of each type. In 1961, 533, 821, and 735 g of cereals were needed to satisfy requirements for protein, iron, and zinc, respectively. By 2011, the amount required increased to 556, 1013,

compensate for eroding nutritional content of cereals but requires more land because of low conversion efficiencies of animal feed to protein (16). Moreover, low-income populations have limited access to substantial amounts of animal products.

NEW METRICS. The standard yield metric of agricultural production by weight per unit land neglects the importance of human nutritional needs as a critical factor for sustainable intensification. We propose a metric of “nutritional yield,” the number of adults who would be able to obtain 100% of their recommended DRI of different nutrients for 1 year from a food item produced annually on one

tons and energy per hectare than rice, wheat, and maize (see the graph, B). In India and LDCs, drought-tolerant millet and sorghum provide nutritional yields for iron higher than rice, despite relatively low yields in terms of metric tons per hectare. In China, wheat and rice have the lowest nutritional yields of all cereals for iron (fig. S6).

Decisions about the desired mix of crops to achieve efficient use of land differ according to the priority: quantity of production or quality to produce essential dietary nutrients. The nutritional yield metric can guide decisions to simultaneously address both priorities at multiple scales. At the field and farm scales, the metric can quantify benefits of improved, nutrition-sensitive agricultural practices, such as the use of biofortified varieties and integrated soil fertility management. At a landscape or national scale, the metric can be applied to formulate policies that promote a mix of crops that balance productivity in terms of quantity and nutritional needs of the respective populations. At the global scale, the metric could be included in projections of food demand and requirements for micronutrients and other macronutrients beyond energy.

With growing pressures on land resources, food systems will be called upon to use land efficiently. At the same time, scarce land resources need to provide adequate nutrition for the world's population and alleviate micronutrient deficiencies. This confluence of imperatives calls for new alliances, metrics, and analyses for incorporating human nutrition as a primary consideration for sustainable agriculture. ■



Healthy foods are needed. They must be rich in essential nutrients produced with efficient use of land. Sacks filled with wheat in Punjab, India.

and 777 g. Grams required to satisfy energy requirements remained nearly unchanged at 625 to 623 g over this time period. The nutrient-to-calorie ratio in our directly consumed cereal supply has declined, with less nutrient-dense cereals contributing to high levels of micronutrient deficiencies, particularly in low-income settings with cereal-based diets. Although our analysis did not address bioavailability, our estimates of amounts needed to meet requirements would be expected to increase because of the relatively low bioavailability of micronutrients in grains, which depends strongly on processing and cooking procedures.

Declining nutritional content of cereals could be compensated by trade and increased consumption of other foods (15). Diversion of cereals for livestock increases consumption of animal products and can

hectare [see supplementary materials (SM)]. Conversely, the inverse of the metric indicates the land area required to grow enough food to supply one person with 100% of recommended daily requirement of different nutrients from a food item for 1 year.

In reality, people do not obtain 100% of recommended nutrient intake from a single food item, and processing and cooking further affect the nutrients available for human consumption (14). However, the metric allows comparison among different crops and production systems to evaluate nutritional value produced from a given land area. The metric could be compiled over different food items and along food value chains to measure nutritional yield for a food system as a whole.

In 2013, for example, on average one hectare of rice produced 4.5 metric tons/year, which is the equivalent of providing the annual energy requirement for 19.9 adults. Millet produced only 0.9 metric tons/ha per year, the annual energy requirement for 4.0 adults. However, a hectare of rice fulfills the annual iron requirement for only 7.6 adults, compared with 15.3 for millet. Similarly, oats yield more zinc per ha than all other cereals except maize, despite providing fewer metric

REFERENCES

1. J. R. Stevenson, N. Villoria, D. Byerlee, T. Kelley, M. Maredia, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 8363 (2013).
2. United Nations Food and Agriculture Organization, “FAOSTAT” (FAO, Rome, 2015).
3. L. R. Carrasco, C. Larrosa, E. J. Milner-Gulland, D. P. Edwards, *Science* **346**, 38 (2014).
4. E. Jobbágy, O. Sala, *Environ. Res. Lett.* **9**, 084014 (2014).
5. H. C. Godfray, T. Garnett, *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* **369**, 20120273 (2014).
6. T. Garnett *et al.*, *Science* **341**, 33 (2013).
7. D. Tilman, M. Clark, *Nature* **515**, 518 (2014).
8. C. A. Palm *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 19661 (2010).
9. E. Cassidy *et al.*, *Environ. Res. Lett.* **8**, 034015 (2013).
10. P. C. West *et al.*, *Science* **345**, 325 (2014).
11. International Food Policy Research Institute, “Global nutrition report 2014: Actions and accountability to accelerate the world's progress on nutrition” (IFPRI, Washington, DC, 2014).
12. S. S. Myers *et al.*, *Nature* **510**, 139 (2014).
13. R. E. Black *et al.*, *Lancet* **382**, 427 (2013).
14. S. de Pee, M. W. Bloem, *Food Nutr. Bull.* **30** (suppl.), S434 (2009).
15. R. Remans, S. Wood, N. Saha, T. D. Anderman, R. DeFries, *Global Food Security* **3**, 174 (2014).
16. V. Smil, *Ambio* **31**, 126 (2002).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/238/suppl/DC1

¹Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY, USA. ²Institute of Human Nutrition, Columbia University, New York, NY, USA. ³Agriculture and Food Security Center, The Earth Institute, Columbia University, New York, NY, USA. ⁴Biodiversity International, Addis Ababa, Ethiopia. ⁵Environmental Defense Fund, San Francisco, CA, USA.

*Corresponding author. E-mail: rd2402@columbia.edu

Living supramolecular polymerization

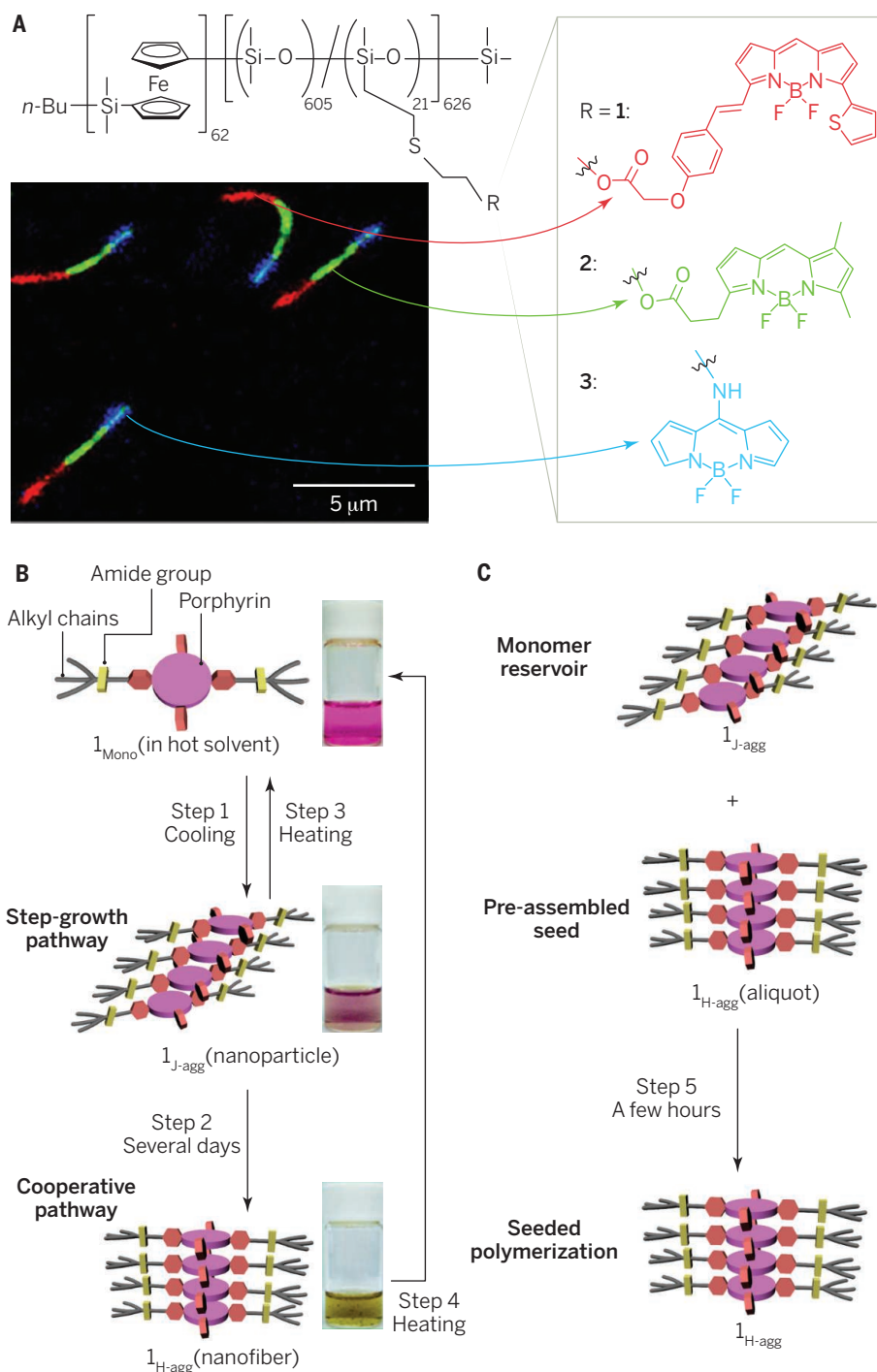
Greater control is achieved over the chain growth and properties of dynamic materials

By **Rahul Dev Mukhopadhyay**^{1,2}
and **Ayyappanpillai Ajayaghosh**^{1,2}

Most polymers that we encounter, like those used in grocery bags and soda bottles, are formed from small molecules (monomers) held together by strong covalent bonds. In supramolecular polymerization, monomers bond through weaker reversible interactions, such as hydrogen bonds (H-bonds). Supramolecular polymerization usually proceeds through step-growth mechanisms (1), where both ends of each monomer are reactive and many smaller oligomers form before long polymers appear. To overcome this problem, a method developed for covalent polymers known as living polymerization has been adopted for supramolecular polymers to achieve better control and uniformity of chain growth and dispersity.

Living polymerization is a type of chain-growth polymerization in which monomers undergo polymerization only upon reacting with an initiator to generate an active center. The active site regenerates with each monomer addition; it propagates along the polymer strand before transferring the active center to another polymer strand or terminating via mutual coupling. In fact, the last two steps of chain transfer or termination are essentially removed in a living polymerization process. The last monomer unit on a polymeric strand remains active until deliberately terminated, so adding more monomer—or a different monomer—resumes the reaction. Properties like the degree of polymerization (number of monomers in the chain), the chain conformation, and its lifetime (of the propagating chain) can therefore be efficiently controlled if a chain-growth polymerization is realized in such dynamic supramolecular systems (2).

An early attempt on controlled supramolecular polymerization was made by Manners, Winnick, and co-workers (3) by assembling polyferrocenyldimethylsilane block copolymers in hydrocarbon solvents. Addition of a fresh feed of the polymer in a good solvent like tetrahydrofuran caused the nanosized cylindrical micelle seeds to grow up to micrometers in length. Very recently, they have



Formation of micelles and nanofibers. (A) Laser scanning confocal microscopy image (scale bar, 5 μm) of self-assembled polyferrocenyldimethylsilane block copolymer micelles functionalized with fluorescence tunable BODIPY (boron-dipyrromethene) derivatives. Molecular structures responsible for different emission colors are represented; *n*-Bu, *n*-butyl (6). (B) Supramolecular seeded growth of nanofibers from molecule **1** proceeds slowly after first forming nanoparticles, but when these nanoparticles interact with existing nanofibers (C), the process is much faster (7).

¹Chemical Sciences and Technology Division, CSIR–National Institute for Interdisciplinary Science and Technology (CSIR-NIIST), Trivandrum 695019, India. ²Academy of Scientific and Innovative Research (AcSIR), CSIR-NIIST, Trivandrum 695019, India. E-mail: ajayaghosh@niist.res.in

prepared hierarchical [one-dimensional (1D) or 3D] multiblock comicelle structures by addition of different block copolymers to the preformed micellar seeds, which were stable both in solution and in solid state (4). “Living crystallization”-driven block copolymer self-assembly was also used to prepare functionalized block copolymers (see the first figure, panel A) with segments of different emission colors (5, 6).

Later, Sugiyasu, Takeuchi, and co-workers showed that systems undergoing self-assembly following a nucleation–elongation mechanism coupled with a kinetically controlled pre-equilibrium process results in living supramolecular polymers (7). These authors observed that a porphyrin-based molecule initially forms that “wrong product,” the for-

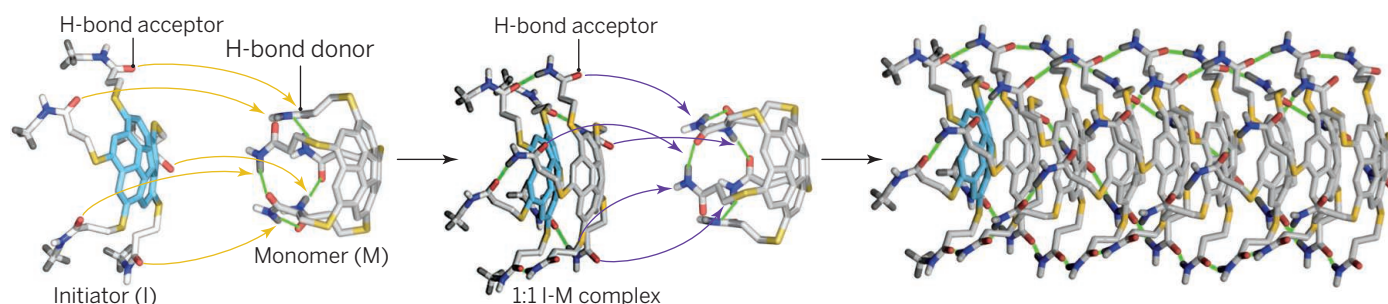
nanofibers of the organogelator as seeds.

Although the above reports have taken us a few steps closer to the concept of living supramolecular polymerization, unimolecular control on chain growth remained elusive until recently. Miyajima, Aida, and co-workers (9) reported an interesting property of a nonplanar bowl-shaped corannulene molecule appended with five amide-functionalized thio alkyl chains (see the second figure, panel A). This C_5 -symmetric molecule did not undergo self-assembly via intermolecular H-bonding in methylcyclohexane (MCH) because the formation of intramolecular H-bonds between the five amide chains is more conformationally feasible. However, upon heating, this metastable cage-like monomer opened up to undergo a sponta-

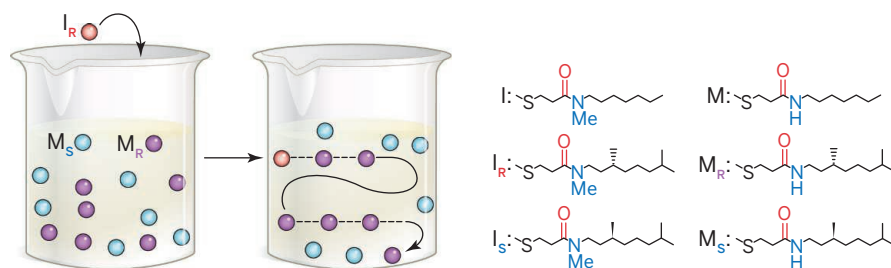
When chiral initiators were used, polymerization occurred only when the configuration of the stereogenic center of the monomer matched with that of the initiator. The monomer (M) with achiral side chains consisted of an asymmetric center at the corannulene core that formed a racemic mixture through a bowl-to-bowl inversion process. The chiral initiators (I) could differentiate between the enantiomers of M and undergo polymerization with selective handedness. This property was used to optically resolve a racemic mixture of monomers M_R and M_S by using either I_R or I_S as an initiator (see the second figure, panel B).

Supramolecular living polymerization is a step closer to the precision synthesis of complex architectures and presents us

A Living polymerization without termination: I--M--M--M--M--M--M--M--M--M--M



B Concept of optical resolution



mation of so-called “J-aggregates” that lead to metastable organic nanoparticles. After several days, these nanoparticles transformed into stable “H-aggregates” with nanofiber-like morphology via a cooperative mechanism (see the first figure, panel B). When an aliquot of the nanofibers was added to a solution of the nanoparticles, which serve as a reservoir for the porphyrin monomers, nanofibers formed much more rapidly (within a few hours) through a living polymerization process (see the first figure, panel C).

Recently, a clever molecular design by Würthner and co-workers (8) allowed a prolonged lag time in the self-assembly process of a perylene bisimide-based organogelator by locking the molecule in an inactive conformation via induced intramolecular H-bonding. A living polymerization process was achieved by introducing preassembled

neous 1D self-assembly. Once such assembly formed, it continued to grow with a fresh supply of the metastable monomer even at low temperatures (9).

Further work by Miyajima, Aida, and co-workers used corannulene molecules functionalized with methyl-substituted amides as the initiator of the chain-growth polymerization because these molecules remain in an open conformation in the absence of intramolecular H-bonds and serve as suitable proton acceptors when a caged corannulene molecule approaches (10). Polymerization was initiated only from one face of the monomer and promoted growth only in a particular direction. The resulting polymers were robust enough to remain stable even at a 10-fold dilution, so it was possible to analyze them by size exclusion chromatography.

Unimolecular control of chain growth. (A) Unimolecular control in a supramolecular living chain-growth polymerization with initiator I and monomer M. (B) Schematic representation of optical resolution process is shown. The different M and I derivatives are represented below; Me, methyl (10).

[Figures reproduced with permission from (6), (7), and (10)]

with a broad canvass to work toward new-generation functional materials. Controlled living polymerization in systems like triaryl-amines (11) or donor-acceptor molecules remains challenging with respect to the design of organic electronic devices with optimum light-conversion efficiency. ■

REFERENCES

1. T. F. A. de Greef, E. W. Meijer, *Nature* **453**, 171 (2008).
2. O. W. Webster, *Science* **251**, 887 (1991).
3. X. Wang et al., *Science* **317**, 644 (2007).
4. H. Qiu et al., *Science* **347**, 1329 (2015).
5. Z. M. Hudson et al., *Nat. Chem.* **6**, 893 (2014).
6. Z. M. Hudson et al., *Nat. Commun.* **5**, 4372 (2014).
7. S. Ogi et al., *Nat. Chem.* **6**, 188 (2014).
8. S. Ogi et al., *J. Am. Chem. Soc.* **137**, 3300 (2015).
9. J. Kang et al., *J. Am. Chem. Soc.* **136**, 10640 (2014).
10. J. Kang et al., *Science* **347**, 646 (2015).
11. V. Faramarzi et al., *Nat. Chem.* **4**, 485 (2012).

10.1126/science.aac7422

In *Ex Machina*, director Alex Garland explores the implications of endowing a robot with emotional intelligence.

ARTIFICIAL INTELLIGENCE

Virtual love

A young programmer falls for a humanoid robot, but is the feeling mutual?

By Rosalind W. Picard

In the opening scene of *Ex Machina*, Caleb Smith, a tall, young programmer, has just won a contest to spend a week at the home of his company's founder, Nathan Bateman. Bateman is the genius programmer behind Bluebook, a corporation handling 94% of the world's web search requests. On arriving at the isolated retreat, Caleb learns that he will be the first person to see if Nathan's newest creation, the robot Ava, passes the Turing test. The classic test, developed by Alan Turing, has the computer hidden from the human, who asks questions, receives answers via text, and in the end must judge, "Is this a human or a computer?" When Caleb points out that this will be different—that he will know that it's a computer from the start—Nathan argues that it will be all the more impressive if Caleb knows that Ava is a computer and yet becomes convinced that she is conscious.

When Caleb meets Ava, we see close ups of Ava's face and shots framing her sculpted breasts, transparent waistline, and internally illuminated organs. Caleb responds as any healthy 26-year-old heterosexual male might, with a slack jaw. When Nathan asks Caleb, "How do you feel about her?" Caleb is effusive: "She's fucking awesome." The more

interesting question, however, is what Nathan asks next, "How does she feel about you?"

As Caleb and Ava supposedly grow closer, we watch for the emotions to develop. When Caleb shares that his parents died when he was 15, Ava's face remains unmoved; her pause is only perfunctory before changing the topic. Later, we see experiences that should elicit expressions of pain and do not. Her face reminds me of the face of a dermatologist I once met who had experimented a bit too eagerly with Botox. Scientists have shown that we read a neutral facial expression as happy if we felt happy before we saw it or as slightly sad if we felt slightly sad before we saw it. We may buy that Ava has feelings because she produces words that describe feelings and because Nathan said "she can feel pleasure," but we do not see her demonstrate convincing empathy or emotional intelligence (1).

At one point, Caleb recognizes that Ava has "mind-reading" abilities. He does not mean the "I know all your thoughts" kind of mind reading that is routinely debunked but the ability to infer the likely mental state of another person—a simple task for most people. For example, if your phone is placed under a basket while you watch, and then moved while you are out of the room, when you come back to retrieve it, a "mind reader" would expect you to look for it where it had been and for you to be surprised to find that it is not where you left it. Ava speaks lines that imply that she is aware of minds and can reason about them. Some scientists argue that a lack of mind reading is the hallmark of autism, although there is also ample counter-evidence to this. As such, Nathan's remark that autistic people are not aware of their own minds, or those of others, is the first line I would have put on the cutting room floor.

When Caleb tries to engage Nathan intellectually about how Ava works, keywords like "stochastic" and "linearized" are sprinkled into the conversation in a way that is in-

Ex Machina

Alex Garland, director
DNA Films, 2015.
108 minutes.



tended to sound erudite without boring the uninitiated. To those of us who build emotional artificial intelligence (AI), however, the phrases have the sound of a toddler randomly sampling keys on a Steinway grand piano.

Nonetheless, the director, Alex Garland, has done some homework on the science of bringing emotions, emotional intelligence, and more to AI. He is wise, for example, to stretch the Turing test over multiple days (2). He is also clearly aware of how valuable it is to collect massive data to train the system. At one chilling point, Nathan tells Caleb how Ava was taught to read and synthesize facial expressions by turning on every camera and microphone in every cell phone on the planet and recording everybody without their knowledge. Still, the film leaps over the major breakthroughs that would be required before we could encounter such a future, namely, the complete lack of evidence that machines could ever have conscious feelings like ours.

A particularly glaring gap is the fact that Ava is programmed without morality. Some might consider morality an option or an upgrade; however, smart machines do not evolve unguided. A machine's choices are significantly biased by the procedures with which it has been programmed. Ava, and her actions, say more about the mind of her programmer than about the robot he created.

REFERENCES AND NOTES

1. For the first scholarly work on emotional intelligence, see P. Salovey, and J.D. Mayer. *Imagination, Cognition and Personality* 9, 3 (1990).
2. For steps on testing computers' emotional intelligence, see R. Picard, *Affective Computing* (MIT Press, Cambridge, 1997).

10.1126/science.aac7899

The reviewer is with the Affective Computing Research Group, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. E-mail: picard@media.mit.edu

PUBLIC HEALTH

Food for thought

A beginner's guide to ethical eating

By Nicholas Freudenberg

Should I get my morning coffee at Dunkin' Donuts, where it is cheap but where it will be served by workers who are not paid a living wage (1)? Or should I go to Starbucks, where it will cost much more but workers get better pay and benefits (2)? When I shop for dinner tonight, should I buy organic, free range, or natural beef—or skip meat altogether to support the rights of cows and the well-being of our planet? And when Monsanto urges me to reject a legislative proposal to require labels on genetically modified food, should I accept their argument that labeling will make food more expensive—especially for the poor? Or should I instead listen to my environmentalist friends who assert that we should not expose humans and the planet to modified organisms whose long-term health and environmental impacts are unknown?

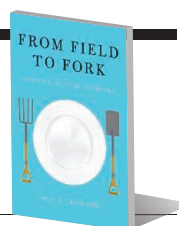
These are among the questions Paul B. Thompson examines in *From Field to Fork: Food Ethics for Everyone*. His goal is not to tell us what to eat but to apply the methods of philosophy and ethics to the choices individuals and societies have to make about food. He hopes to help readers find the right balance between “enabling choices vs. imposing values on others.” His refreshing belief is that public dialogue and debate on the values and ethics of key food questions will enable policy-makers and the public to make more informed and equitable choices.

Thompson is a philosopher at Michigan State University, where he holds the W. K. Kellogg Chair in Agricultural, Food, and Community Ethics. His 30 years' experience at the crossroads of philosophy, food, and agriculture make him a knowledgeable guide to food ethics, a field he helped to create. This book considers a variety of topics that have attracted media and public attention: hunger and food insecurity in a land of plenty; the impact of the Green Revolution and genetically modified crops on food



From Field to Fork Food Ethics for Everyone

Paul B. Thompson
Oxford University Press,
2015. 343 pp.



production, famine, and the environment; the ethical, health, and environmental rationales for vegetarianism; and the human cost of cheap food (workers paid so little they must depend on public benefits to survive).

Thompson's deep knowledge gives him some unexpected insights. Although he doubts that existing evidence on genetically modified organisms (GMOs) warrants restrictions for health or environmental reasons and accepts the contention that they can help make food more available to the world's poor, he argues that their widespread use endangers traditional forms of agrarianism. The ancient Greeks, Thomas Jefferson, and current agroecologists have all claimed that small farmers are the foundation for democracy and equality. By patenting seeds and dominating global markets, a few multinational companies have wrested decisions about how to farm from small growers and thus diminished their power to shape rural economies and politics. Thompson argues that these and other doubts about GMOs warrant continued skepticism about expanding their use.

Thompson recognizes that how we grow, sell, and eat food cannot be separated from our economic and political systems. “Since everyone needs to eat,” he writes, “control of food is a powerful locus for both profit-taking and the exercise of social control.” However, the book does not directly

From Field to Fork examines how our food choices affect society and the planet.

consider some of the most urgent ethical questions confronting our food system. For example, are the business models of PepsiCo and McDonald's ethical when they depend on persuading individuals to consume high-sugar, -fat, and -calorie products that are associated with costly diet-related diseases? And is New York City imposing a “nanny state” or simply living up to its mandate to protect public health when it seeks policies designed to restrict portion sizes or add warning labels to unhealthy foods? By not addressing such questions, Thompson leaves two key players in our food system—global food companies and government—curiously absent from his analyses. Because changing the practices of these two food behemoths is the goal of the global food justice movement, a deeper analysis of the ethical frameworks each uses could have helped to advance the search for sustainable and moral solutions to global food problems.

Despite this gap, *From Field to Fork* is a sensible and engaging introduction to food ethics. Thompson avoids moralizing and rejects the tendency of some scientists to assume that evidence alone can settle food policy fights without engaging with the moral arguments of critics. By emphasizing that our food system ultimately reflects our values, Thompson sets the table for more satisfying discussions of food policy.

REFERENCES AND NOTES

1. www.glassdoor.com/Salary/Dunkin-Donuts-Salaries-E19153.htm.
2. www.glassdoor.com/Salary/Starbucks-Salaries-E2202.htm.

The reviewer is at the City University of New York School of Public Health and the New York City Food Policy Center at Hunter College, New York, NY 10035, USA.
E-mail: nfreuden@hunter.cuny.edu

LETTERS

Edited by Jennifer Sills

Holocene as Anthropocene

IN THEIR PERSPECTIVE “Defining the epoch we live in” (3 April, p. 38), W. F. Ruddiman *et al.* write that in spite of its popularity, the Anthropocene still lacks an official onset. They propose that the term anthropocene be used informally (without the initial capital), which would avoid the constraints of a formal designation. We disagree.

The human footprint on the planet is so distinctive and lasting that the Age of Man must be officially recognized in the geologic time scale. We support considering the Anthropocene and the Holocene as a single geologic time span. This solution has already been proposed (1) but was inexplicably ignored. Combining the two epochs would provide the Anthropocene with a climatic justification (2). It was the end of the last glacial period that allowed the increase of the human population and its role as a geologic force, and anthropic impacts on the planet took place during most of the Holocene. Moreover, the latest studies are placing the appearance of domesticated plants and animals—which can plausibly be considered the onset of the Anthropocene (3)—increasingly earlier in time, closer and closer to the 11.7 ky B.P. Pleistocene–Holocene boundary (4).

Even if the impact of humans was not immediately major and uniform across the planet, what are a few thousand years of discrepancy between the end of the last glacial period and the dawn of the Anthropocene in comparison to the entire geological scale? Most transitions between the formally defined epochs would be much longer than the Holocene. Anthropocene seems a more reasonable name than Holocene for this combined time span, whose most characteristic trait is the human pressure on the planet. Holocene could possibly be the first stage of the Anthropocene, the one characterized by a soft and spotty human impact on Earth.

**Giacomo Certini¹ and
Riccardo Scalenghe²**

¹Dipartimento di Scienze delle Produzioni Agroalimentari e dell'Ambiente, Università degli Studi di Firenze, 50144 Firenze, Italy. ²Dipartimento di Scienze Agrarie e Forestali, Università degli Studi di Palermo, 90128 Palermo, Italy.



At what point did humans' impact on Earth become geologically significant?

*Corresponding authors. E-mail: giacomo.certini@unifi.it (G.C.); riccardo.scalenghe@unipa.it (R.S.)

REFERENCES

1. B. D. Smith, M. A. Zeder, *Anthropocene* **4**, 8 (2013).
2. M. Walker *et al.*, *J. Quat. Sci.* **24**, 3 (2009).
3. M. Balter, *Science* **340**, 261 (2013).
4. O. Smith *et al.*, *Science* **347**, 998 (2015).

Geological evidence for the Anthropocene

DEFINITIONS ARE ONE of the bedrocks of science. However, W. F. Ruddiman *et al.* (“Defining the epoch we live in,” Perspectives, 3 April, p. 38) propose that the geological term Anthropocene should remain deliberately undefined and ambiguous. They recommend not formalizing the term because some inception dates miss important earlier human-induced environmental impacts, particularly widespread farming. We have sympathy for this view. However, defining the Anthropocene as a geological epoch must be based on evidence, including changes to the Earth system lasting millions of years and the existence of stratigraphic evidence marking such changes (1). All other geological time-units have agreed-upon dated markers or agreed-upon dates. The Anthropocene should not be treated differently.

Ruddiman *et al.* discuss a recent paper suggesting that the Anthropocene began in 1945 (2). Zalasiewicz *et al.* (2) note, as have others (3), that many stratigraphic changes are coeval with the mid-to-late 20th century. However, they then side-step geological convention by not specifying a Global Stratotype Section and Point (GSSP) to define the Anthropocene. Instead, they choose the date of the first nuclear weapon detonation (2), which created no stratigraphic evidence in 1945. Given that GSSPs are the preferred method of marking geological time units (1), 1945 is unlikely

to obtain the backing of the multiple committees of geologists necessary to formally define the Anthropocene epoch (4).

The appropriate GSSP marker relating to nuclear weapons is the globally measurable peak in radionuclide fallout, in 1964, after the Partial Test Ban Treaty (3). An earlier date also likely adheres to the geological criteria for defining an epoch: 1610. This date marks the irreversible exchange of species following the collision of the Old and New worlds; an associated unusual drop in atmospheric CO₂ captured in Antarctic ice cores provides the GSSP marker (3). This date implicitly includes the impacts of farming, as the drop in CO₂ largely resulted from vegetation regrowth on abandoned farmlands following the deaths of 50 million indigenous Americans (mostly from smallpox brought by Europeans). The annexing of the Americas by Europe was also an essential precursor to the Industrial Revolution and therefore captures associated later waves of environmental change (3). Choosing among alternative inception dates will be challenging (5).

Finally, Ruddiman and colleagues suggest using the term Anthropocene with a small “a” to denote its informality. This is not consistent with geological-community norms. Formal time-unit names have a capital letter: Anthropocene Epoch versus Anthropocene epoch for the informal term (1, 3). An elegant solution to referring to time before an agreed-upon Anthropocene Epoch is already available: the paleoanthropocene (6). Geological norms should be followed when considering defining the Anthropocene.

**Simon L. Lewis^{1,2} and
Mark A. Maslin¹**

¹Department of Geography, University College London, London, WC1E 6BT, UK. ²School of Geography, University of Leeds, Leeds, LS2 9JT, UK. *Corresponding author. E-mail: S.L.Lewis@leeds.ac.uk

REFERENCES

1. F. M. Gradstein, J. G. Ogg, M. D. Schmitz, G. M. Ogg, *The Geologic Time Scale 2012* (Elsevier BV, Netherlands, 2012).
2. J. Zalasiewicz *et al.*, *Quat. Int.* (2015); 10.1016/j.quaint.2014.11.045.
3. S. L. Lewis, M. A. Maslin, *Nature* **519**, 171 (2015).
4. M. Walker, P. Gibbard, J. Lowe, *Quat. Int.* (2015); doi.org/10.1016/j.quaint.2014.11.0450.
5. S. L. Lewis, M. A. Maslin, *Anthropocene Rev.* **2**, 128 (2015).
6. S. F. Foley *et al.*, *Anthropocene* **3**, 83 (2013).

Response

CERTINI AND SCALENGHE and Lewis and Maslin seem convinced that geoscience must rely heavily on formal stratigraphic nomenclature to move forward. This view likely originates from the centuries-long effort to compile the relative geologic age sequence, which was a monumental achievement, especially given religious and social opposition. The compilation required careful attention to stratigraphic principles such as superposition, index fossils, diachroneity, and facies changes. Over time, the findings became formalized in geologic nomenclature.

However, since the mid-1900s, geochemical (radiometric) dating has increasingly overshadowed (although not replaced) those time-honored principles. In my experience, most geoscience papers in recent decades initially refer to geologic nomenclature as a way of providing a familiar frame of reference, but they soon shift focus to the best geochemical dating available. Although my view will displease some geological colleagues, isn't it time to acknowledge that almost all dating of geological sequences now takes place in the "geochemical age" (with a small "a")?

Lewis and Maslin propose that we designate a "Paleoanthropocene" (capital "P"). Should we place its start at the mass extinction of marsupials in Australia near 50,000 years ago? Or should we choose the mass extinction of large mammals in the Americas almost 40,000 years later? Or, if we decide to pick a level during the 10,000 years or more when crops and livestock were being domesticated and agriculture was spreading across the continents, which one time within that long span should we use for a formal designation? Could any choice be satisfying, given that it would have to ignore the rest of the long and rich history of early human influences?

In contrast, informal use of these terms (anthropogenic, anthropocene, and paleoanthropocene) would present no such problems or constraints.

William F. Ruddiman

Department of Environmental Sciences, University of Virginia, Charlottesville, VA 22903, USA.
E-mail: wfr5c@virginia.edu

TECHNICAL COMMENT ABSTRACTS

Comment on "Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago"

K. D. Bennett

Smith *et al.* (Reports, 27 February 2015, p. 998) identify wheat DNA from an 8000-calendar-years-before-the-present archaeological site in southern England and conclude that wheat was traded to Britain 2000 years before the arrival of agriculture. The DNA samples are not dated, either directly or from circumstantial evidence, so there is no chronological evidence to support the claim.

Full text at <http://dx.doi.org/10.1126/science.aab1886>

Response to Comment on "Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago"

Oliver Smith, Garry Momber, Richard Bates, Paul Garwood, Simon Fitch, Mark Pallen, Vincent Gaffney, Robin G. Allaby

Bennett questions the rigor of the dating of our sample from which sedimentary ancient DNA was obtained and the reliability of the taxonomic identification of wheat. We present a further radiocarbon date from S308 that confirms the lateral consistency of the palaeosol age. The suggestion of taxonomic false positives in our data illustrates a misinterpretation of the phylogenetic intersection analysis.

Full text at <http://dx.doi.org/10.1126/science.aab2062>

ERRATA

Erratum for the Report "Mutation rate and genotype variation of Ebola virus from Mali case sequences" by T. Hoenen *et al.*, *Science* **348, aac5674 (2015).** Published online 22 May 2015; 10.1126/science.aac5674

Erratum for the Report: "Molten uranium dioxide structure and dynamics" by L. B. Skinner *et al.*, *Science* **348, aab3869 (2015).** Published online 1 May 2015; 10.1126/science.aab3869

Correction for the Report: "The in vivo dynamics of antigenic variation in *Trypanosoma brucei*" by M. R. Mugnier *et al.*, *Science* **347, aaa4502 (2015).** Published online 22 April 2015; 10.1126/science.aaa4502. Fig. 2A contained errors in the print version of *Science*. The correct figure is displayed online.

Erratum for the Research Article: "Neurotransmitter switching in the adult brain regulates behavior" by D. Dulcis *et al.*, *Science* **348, aab2338 (2015).** Published online 10 April 2015; 10.1126/science.aab2338

TECHNICAL COMMENT

ARCHAEOLOGY

Comment on “Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago”

K. D. Bennett^{1,2}

Smith *et al.* (Reports, 27 February 2015, p. 998) identify wheat DNA from an 8000-calendar-years-before-the-present archaeological site in southern England and conclude that wheat was traded to Britain 2000 years before the arrival of agriculture. The DNA samples are not dated, either directly or from circumstantial evidence, so there is no chronological evidence to support the claim.

Current understanding of the spread of agriculture across Europe indicates arrival in Britain about 6000 years ago, having taken about 2000 years to spread across central Europe and the English Channel from the Mediterranean (1). There is abundant evidence for the occurrence of early forms of wheat in Britain after 6000 calendar years before the present (cal yr B.P.), so the main claim of Smith *et al.*'s investigations at Bouldnor Cliff southern England, is the early date of about 8000 cal yr B.P. (2), rather than the discovery of wheat itself.

The samples examined for DNA (2) come from a monolith, S308, collected “from a location at the site” [p. 999 in (2)], or “in the proximity MS-08 and MS-05, which were adjacent to each other” [p. 4 in the supplementary materials (SM) for (2)]. Nowhere in the Report is there any mention of a measurement between S308 and other monoliths, either horizontally or vertically. It is not clear how much time elapsed between the collection of the dated monoliths and S308. All the radiocarbon dating was carried out on monoliths other than S308 [table S1 in the SM for (2)]. No information

is provided on other analyses that would provide a firm link between the DNA analyses of S308 and any sample of the dated monoliths. Sedimentary analyses were carried out on the other monoliths (3) and might easily have been applied to S308 to provide a link even if, for some reason, it was not possible to obtain a radiocarbon date directly from the sediment matrix (within which the wheat DNA was obtained) of S308.

The stratigraphic section for Bouldnor Cliff [figure 1B in (2)] shows a complex series of coastal sediments, dipping into a channel. This raises the possibility that samples taken later, even if at the same measured location (horizontal and vertical) as the originals, might have come from a stratigraphically different part of the sequence, depending on accuracy and precision of both horizontal and vertical measurements, as well as allowance for erosion between the dates of sample collection. There are also inconsistencies in the presentation of the stratigraphy, including the relative extents of monoliths MS-04 and MS-07 [separated by a gap in figure 1B in (2) but contiguous in figure S2 in the SM for (2)]. Reference is made to a paleosol, which might incorporate material from a wide range of ages.

The only evidence for wheat at the site comes from DNA. There are no wheat macrofossils or wheat pollen in the samples. The contention is that this indicates that the wheat did not come from nearby agriculture (2). However, whenever

the wheat was incorporated in the sediment (from nearby agriculture or trading), it must have been associated with macrofossil remains, such as wheat grains. Both hazel and alder macrofossils (but no DNA) were found in or near S308 (2), so it is curious that no trace of wheat remains have been found. There are inconsistencies in the DNA—even after the phylogenetic intersection analysis (PIA)—including the presence of DNA from animals known in the mid-Holocene only from regions far removed from southern Britain, such as *Ursus maritimus* (polar bear, Arctic) and *Cervus nippon* (Sika deer, East Asia), and from tropical grasses (Panicoideae). These suggest that some false positives are coming through PIA or that the taxonomic precision indicated is inappropriate for the content of the database, which might be relevant to the age or origin of the wheat identification.

Given that S308 contains DNA of wheat, how might it have arrived there? There are a number of possibilities. The samples may be contemporaneous with the monoliths (2), so the wheat may have been traded from Europe, presumably all the way from the Mediterranean at 8000 cal yr B.P. (7). Alternatively, if the samples are actually much younger, the wheat may have been incorporated from nearby agriculture, perhaps from eroded soil, and ended up near to early Holocene samples through the vagaries of sedimentation in a complex channel and coastal situation. The latter is the more parsimonious view, in the absence of evidence to the contrary. Finally, the identification of the wheat DNA may be another false positive in the comparisons with database DNA.

The claim that wheat was being traded to Britain at 8000 cal yr B.P. has substantial implications for understanding of the archaeology of northwest Europe in the early postglacial period. Such a claim can be readily substantiated by dating the sediment matrix of the samples directly. A claim based on undated samples, lacking any incontrovertible link to dated samples, is insufficient to overturn current understanding.

REFERENCES AND NOTES

1. P. Rowley-Conwy, *Curr. Anthropol.* **52** (S4), S431–S451 (2011).
2. O. Smith *et al.*, *Science* **347**, 998–1001 (2015).
3. G. Larson, *Science* **347**, 945–946 (2015).

ACKNOWLEDGMENTS

I thank M. Blaauw and L. Parducci for comments on an early draft.

24 March 2015; accepted 28 May 2015
10.1126/science.aab1886

¹School of Geography, Archaeology, and Palaeoecology, Queen's University Belfast, University Road, Belfast BT7 1NN, Northern Ireland, UK. ²Palaeobiology, Department of Earth Sciences, Uppsala University, Villavägen 16, S-752 36 Uppsala, Sweden.
E-mail: k.d.bennett@qub.ac.uk

TECHNICAL RESPONSE

ARCHAEOLOGY

Response to Comment on “Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago”

Oliver Smith,¹ Garry Momber,² Richard Bates,³ Paul Garwood,⁴ Simon Fitch,⁵
Mark Pallen,^{6*} Vincent Gaffney,^{7*} Robin G. Allaby^{1*,†}

Bennett questions the rigor of the dating of our sample from which sedimentary ancient DNA was obtained and the reliability of the taxonomic identification of wheat. We present a further radiocarbon date from S308 that confirms the lateral consistency of the palaeosol age. The suggestion of taxonomic false positives in our data illustrates a misinterpretation of the phylogenetic intersection analysis.

Bennett (1) raises the point that possible variances in the depositional environment make it inappropriate to apply radiocarbon dates from nearby monoliths MS-05 and MS-08 in our study (2), raising the possibility that the wheat DNA we identified could possibly be younger than the 8000 years we claim. The evidence from the radiocarbon dates at the site and those we presented is that the palaeosol is of a consistent age over a 420-m area reaching between the sites of Bouldnor Cliff II, on which our study was based, and Bouldnor Cliff V (3). We therefore felt that the site was dated securely enough in this study. However, we accept that Bennett's argument is most robustly refuted by obtaining a radiocarbon date from the sample S308, which is a box sample taken from the cliff as outlined in the methods and not, as Bennett suggests, a monolith. We attempted to date both a twig from the sample

and the humic acid fraction of the sediment itself. Sediment dating is problematic because of the risk of inclusion of older carbon sources and can lead to overestimation of age. In this case, the humic acid content of the sandy clay was too low to provide a direct sediment date, but the twig returned an age of 7935 to 7790 calendar years before the present (Beta-406961), confirming the age of the sample from which the sedimentary ancient DNA (sedaDNA) was obtained.

Bennett asserts that the phylogenetic intersection analysis (PIA) fails to filter out exotic species such as *Ursus maritimus* (polar bear), *Cervus nippon* (Sika deer), and tropical panicoide, and as such casts doubt on the validity of the identity of the wheat DNA. This is incorrect and shows some misunderstanding of the DNA analysis. The analysis is shown to be highly robust and does not in any way falsely identify polar bears, Sika deer, or tropical grasses as being present at the site. These are instances of closest match between sedaDNA reads and the database; it would be naive to interpret these as a species identification, for reasons we explained in some depth in the supplementary materials of (2). These are not instances of reads that have been filtered out by the PIA, but rather meet the criteria of the analysis. The PIA is predicated on the fact that, due to variable database representation, the species of origin may not be present in the database; indeed, often, large taxonomic orders are represented by only a very few species for many genomic regions. The robustness of the analysis comes from

examining the phylogenetic range of similar DNAs within a database. In fact, Bennett is alluding to sedaDNA that is attributed robustly to Carnivora, Caniformia, or Ursidae, in the case of the “polar bear,” meaning that the true species of origin lies somewhere in the phylogenetic range encompassed by this order, suborder, or family, respectively. Similarly, cervinae is identified in the case of the “Sika deer,” and various uncontroversial higher taxonomic orders of grasses are identified in the analysis in which the closest database match is from a tropical grass such as rice.

Factors such as genome size and how well genomes are represented in databases have a great influence on the resulting frequency in the DNA profile. Wheat has a large and well-characterized genome, and, furthermore, many wheat species and the sister species *Aegilops* have been characterized. This means that we have a great power to detect wheat relative to other organisms from metagenomic profiles using the PIA, which is reflected in the relatively large number of wheat sequences we identify. Given the high level of accuracy (81%) of the analysis, the evidence for the presence of wheat at Bouldnor is overwhelming.

Finally, Bennett asserts that sedaDNA requires the existence of macrofossil sources of DNA. The sandy clay from which we took the sedaDNA sample was largely devoid of macrofossils save for a few twigs, with a low organic content, despite the large number of taxonomic orders that were identified. Bouldnor Cliff has not been extensively excavated yet, and it is possible that the quantities of grain involved may be very small and that such wheat macrofossils remain to be discovered. However, the presence of sedaDNA in the absence of macrofossils isprecedented (4). We can make no conclusion about the source of the DNA without further evidence. However, if rapid transport occurred—such as might be expected, for instance, using boats associated with pioneer groups on the western coast of France—then the source may even have been flour rather than grain.

REFERENCES AND NOTES

1. K. D. Bennett, *Science* **349**, 247 (2015).
2. O. Smith et al., *Science* **347**, 998–1001 (2015).
3. G. Momber et al., in *Mesolithic Occupation at Bouldnor Cliff and the Submerged Prehistoric Landscapes of the Solent*, G. Momber, D. Tomalin, R. Scaife, J. Satchell, J. Gillespie, Eds., Council for British Archaeology Research Report 164 (CBA, York, 2011), pp. 66–93.
4. J. Haile et al., *Proc. Natl. Acad. Sci. U.S.A.* **106**, 22352–22357 (2009).

ACKNOWLEDGMENTS

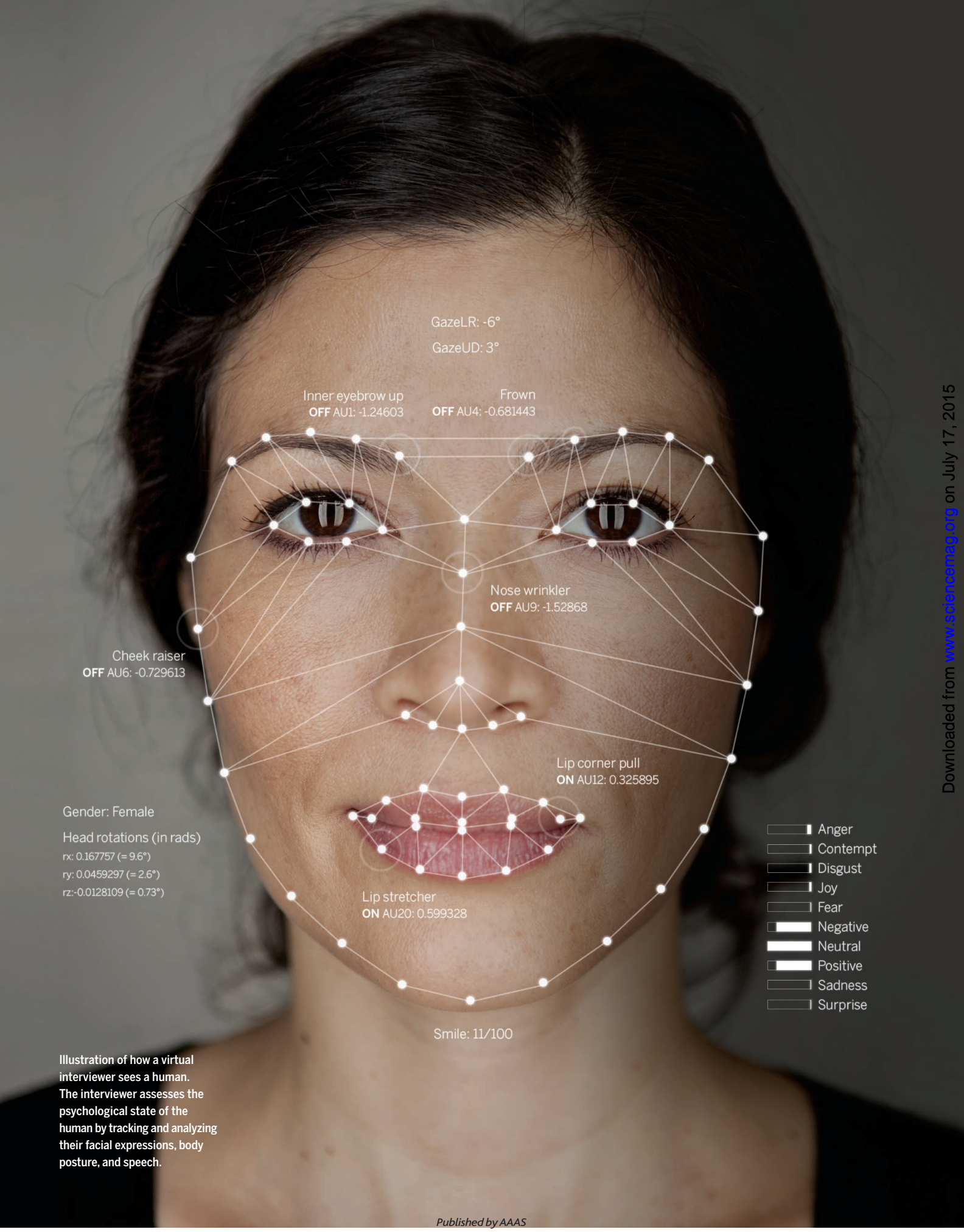
Funding support for O.S. was by the Natural Environment Research Council (NE/L006847/1).

7 April 2015; accepted 28 May 2015
10.1126/science.aab2062

¹School of Life Sciences, University of Warwick, Coventry CV4 7AL, UK. ²Maritime Archaeology Trust, Room W1/95, National Oceanography Centre, Empress Dock, Southampton SO14 3ZH, UK. ³Department of Earth Sciences University of St. Andrews, St. Andrews, Fife, KY16 9AL, Scotland.

⁴Department of Classics, Ancient History and Archaeology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK. ⁵School of History and Cultures, University of Birmingham, IBM VISTA ERI building, Pritchatts Road, Birmingham, B15 2TT, UK. ⁶Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK. ⁷Division of Archaeological, Geographical, and Environmental Sciences, University of Bradford, Bradford, West Yorkshire BD7 1DP, UK.

*These authors contributed equally to this work. †Corresponding author. E-mail: r.g.allaby@warwick.ac.uk



GazeLR: -6°

GazeUD: 3°

Inner eyebrow up
OFF AU1: -1.24603

Frown
OFF AU4: -0.681443

Nose wrinkler
OFF AU9: -1.52868

Cheek raiser
OFF AU6: -0.729613

Lip corner pull
ON AU12: 0.325895

Gender: Female

Head rotations (in rads)

rx: 0.167757 (= 9.6°)

ry: 0.0459297 (= 2.6°)

rz: -0.0128109 (= 0.73°)

Lip stretcher
ON AU20: 0.599328

Smile: 11/100











-  Anger
-  Contempt
-  Disgust
-  Joy
-  Fear
-  Negative
-  Neutral
-  Positive
-  Sadness
-  Surprise

Illustration of how a virtual interviewer sees a human. The interviewer assesses the psychological state of the human by tracking and analyzing their facial expressions, body posture, and speech.

RISE OF THE MACHINES

By Jelena Stajic, Richard Stone, Gilbert Chin, and Brad Wible

Although most would agree that the average person is smarter than the average cat, comparing humans and machines is not as straightforward. A computer may not excel at abstract reasoning, but it can process vast amounts of data in the blink of an eye. In recent years, researchers in artificial intelligence (AI) have used this computational firepower on the scads of data accumulating online, in academic research, in financial records, and in virtually all walks of life. The algorithms they develop help machines learn from data and apply that knowledge in new situations, much like humans do. The ability of computers to extract personal information from seemingly innocuous data raises privacy concerns. Yet many AI systems indisputably improve our lives; for example, by making communication easier through machine translation, by helping diagnose illness, and by providing modern comforts, such as your smartphone acting as your personal assistant. This special issue presents a survey of the remarkable progress made in AI and outlines challenges lying ahead.

Many AI systems are designed for narrow applications, such as playing chess, flying a jet, or trading stocks. AI researchers also have a grander aspiration: to create a well-rounded and thus more humanlike intelligent agent. Scaling that research peak is daunting. But triumphs in the field of AI are bringing to the fore questions that, until recently, seemed better left to science fiction than to science: How will we ensure that the rise of the machines is entirely under human control? And what will the world be like if truly intelligent computers come to coexist with humankind?

The editors gratefully acknowledge the advice of Eric Horvitz (Microsoft) on the Reviews in this special issue.

INSIDE

NEWS

The synthetic therapist *p.* 250

Fears of an AI pioneer *p.* 252

POLICY FORUM

Data, privacy, and the greater good
p. 253

REVIEWS

Machine learning: Trends, perspectives, and prospects *p.* 255

Advances in natural language processing *p.* 261

Economic reasoning and artificial intelligence *p.* 267

Computational rationality: A converging paradigm for intelligence in brains, minds, and machines *p.* 273

RELATED ITEM

► BOOKS ET AL. *p.* 243



Ellie, a virtual health agent, monitors patients' expressions, gestures, and voice.

The synthetic therapist

Some people prefer to bare their souls to computers rather than to fellow humans

By John Bohannon

People have always noticed Yrsa Sverrisdottir. First it was ballet, which she performed intensively while growing up in Iceland. Then it was science, which she excelled at and which brought her to the stage at conferences. And starting in 2010, when she moved to the University of Oxford in the United Kingdom to study the neurophysiology of the heart, it was her appearance. With her Nordic features framed by radiant blonde hair, “I just stand out here,” she says. “I can’t help it.”

After she arrived in the United Kingdom, she found that she no longer enjoyed the attention. She began to feel uncomfortable in crowds. Her relationships suffered. There had been some obvious stressors, such as the deaths of both of her parents. But the unease didn’t let up. By 2012, she says, “I felt

like I was losing control.” Then she met Fjola Helgadóttir, one of the few other Icelanders in town. Helgadóttir, a clinical psychology researcher at Oxford, had created a computer program to help people identify and manage psychological problems on their own. Sverrisdottir decided to give it a try.

The program, based on a technique called cognitive behavioral therapy (CBT) and dubbed CBTpsych, begins with several days of interactive questioning. “It was exhausting,” Sverrisdottir says. The interrogation started easily enough with basic personal details, but then began to probe more deeply. Months of back and forth followed as the program forced her to examine her anxieties and to identify distressing thoughts. CBTpsych diagnosed her as having social anxiety, and the insight rang true. Deep down, Sverrisdottir

realized, “I didn’t want people to see me.”

Then the program assumed the role of full-fledged therapist, guiding her through a regimen of real-world exercises for taking control. It sounds like a typical success story for clinical psychology. But no human psychologist was involved.

CBTpsych is far from the only computerized psychotherapy tool available, nor the most sophisticated. Ellie, a system built at the University of Southern California (USC) in Los Angeles, uses artificial intelligence (AI) and virtual reality to break down barriers between computers and humans. Originally funded by the U.S. military, its focus is on diagnosing and treating psychological trauma. Because patients interact with a digital system, the project is generating a rare trove of data about psychotherapy itself. The aim, says Albert “Skip” Rizzo, the USC

psychologist who leads the effort, is nothing short of “dragging clinical psychology kicking and screaming into the 21st century.”

A 19 June editorial in *The New York Times* deemed computerized psychotherapy “effective against an astonishing variety of disorders.” The penetration of the Internet into far-flung communities could also bring mental health treatment to vast numbers of people who otherwise have no access.

But whether clinical psychologists will accept AI into their practice is uncertain. Nor is it clear that the tools of AI can carry computerized psychotherapy beyond its so far limited capacity, says Selmer Bringsjord, a cognitive scientist and AI researcher at Rensselaer Polytechnic Institute in Troy, New York. “It is incredibly ambitious.”

ALL OF TODAY’S VIRTUAL psychologists trace their origins to ELIZA, a computer program created half a century ago. Named after the young woman in *Pygmalion* who rapidly acquires sophisticated language, ELIZA was nothing more than a few thousand lines of code written by Joseph Weizenbaum and other computer scientists at the Massachusetts Institute of Technology (MIT) in the early 1960s to study human-computer interaction.

ELIZA followed rules that determined how to respond during a dialogue. The most convincing results came from a rule set called DOCTOR that simulated a psychotherapist: By turning patients’ statements around as questions, the program coaxed them to do most of the talking. For instance, in response to a patient saying, “I feel helpless,” the computer might respond, “Why do you think you feel that way?” (You can talk to ELIZA yourself at <http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm>.)

People engaged readily with ELIZA, perhaps more for its novelty than its conversational skills, but AI researchers were unimpressed. “The idea that you could make a convincing AI system that didn’t really have any intelligence was seen as cheating,” says Terry Winograd, a computer scientist at Stanford University in Palo Alto, California, who was a Ph.D. student down the hall from Weizenbaum. This was a wildly optimistic time for the field, with many researchers anticipating computers with human-level general intelligence right around the corner.

But work on artificial general intelligence didn’t pan out, and funding and interest dried up in what has come to be known as the “AI winter.” It wasn’t until the turn of the new millennium that mainstream interest in AI resurged, driven by advances in “narrow AI,” focusing on specific problems such as voice recognition and machine vision.

Conversational “chatbots” such as ELIZA are still viewed as a parlor trick by most computer scientists (*Science*, 9 January, p. 116). But the chatbots are finding a new niche in clinical psychology. Their success may hinge on the very thing that AI researchers eschew: the ability of an unintelligent computer to trick people into believing that they are talking to an intelligent, empathetic person.

THAT ISN’T EASY, as Rizzo is keenly aware. What most often breaks the spell for a patient conversing with Ellie isn’t the content of the conversation, because the computer hews closely to a script that Rizzo’s team based on traditional clinical therapy sessions. “The problem is entrainment,” he says, referring to the way that humans subconsciously track and mirror each other’s emotions during a conversation.

For example, a patient might say to Ellie, “Today was not the best day,” but the voice recognition software misses the “not.” So Ellie smiles and exclaims, “That’s great!” For an AI system striving to bond with a human patient and earn trust, Rizzo says, “that’s a disaster.”

The goal is “dragging clinical psychology kicking and screaming into the 21st century.”

Albert “Skip” Rizzo, University of Southern California

To improve entrainment, a camera tracks a patient’s psychological signals: facial expression, posture, hand movement, and voice dynamics. Ellie crunches those data in an attempt to gauge emotional state.

The patterns can be subtle, says Louis-Philippe Morency, a computer scientist at USC who has led the development of the AI that underlies Ellie. For instance, he says, a person’s voice may shift “from breathy to tense.” The team devised algorithms to match patterns to a likely emotional state. It’s imperfect, he says, but “our experiments showed strong correlation with [a patient’s] psychological distress level.”

Other patterns unfold over multiple sessions. For instance, the team’s work with U.S. veterans suffering from post-traumatic stress disorder (PTSD) revealed that “smile dynamics” are a strong predictor of depression. The pattern is so subtle that it took a computer to detect it: Smiling frequency remained the same in depressed patients, on average, but the duration and intensity of their smiles was reduced.

Even if Ellie were to achieve perfect entrainment, Rizzo says, it “is really just an enhanced ELIZA.” The AI under the hood can only sustain about a 20-minute conversation

before the spell breaks, which limits the system’s usefulness for diagnosis and treatment of most psychological problems. Without sophisticated natural language processing and semantic knowledge, Ellie will never fool people into believing that they are talking to a human. But that’s okay, Rizzo says: Becoming too humanlike might backfire. One counterintuitive finding from Rizzo’s lab came from telling some patients that Ellie is a puppet controlled by a human while telling others she is fully autonomous. The patients told there was a puppeteer were less engaged and less willing to open up during therapy.

That’s no surprise to AI researchers like Winograd. “This goes right back to ELIZA,” he says. “If you don’t feel judged, you open up.”

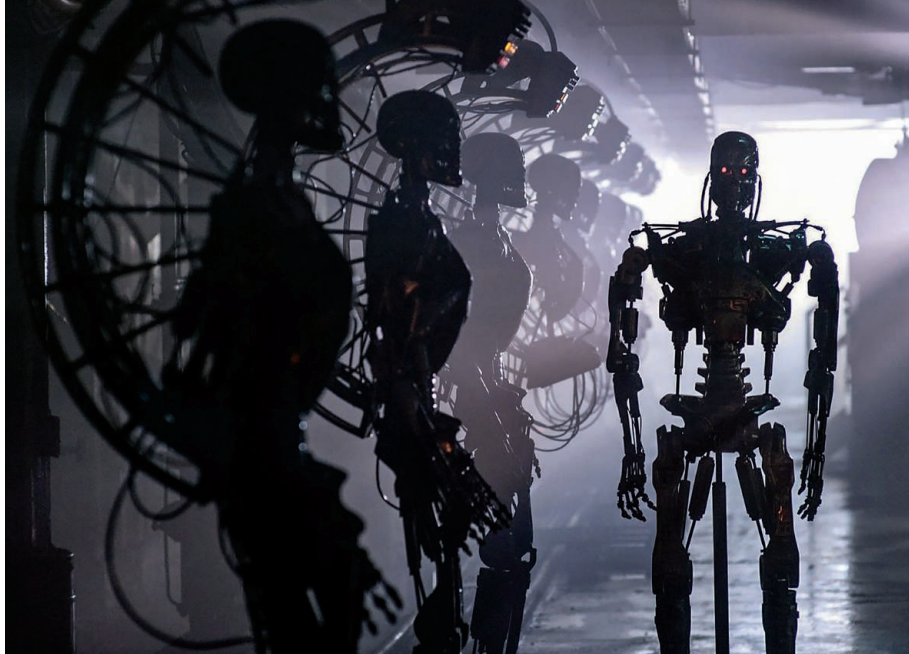
Ethical and privacy issues may loom if AI therapy goes mainstream. Winograd worries that online services may not be forthcoming about whether there is a human in the loop. “There is a place for deceiving people for their own good, such as using placebos in medicine,” he says. But when it comes to AI psychology, “you have to make it clear to people that they are talking to a machine and not a human.”

If patients readily open up to a machine, will clinicians be needed at all? Rizzo is adamant that a human must always be involved because machines cannot genuinely empathize with patients. And Ellie, he points out, has a long way to go before being ready for prime time: The program does not yet have the ability to learn from individual patients. Rizzo envisions AI systems as a way to gather baseline data, providing psychologists with the equivalent of a standard battery of blood tests. “The goal isn’t to replace people,” he says, “but to create tools for human caregivers.”

Helgadottir has a bolder vision. Although computers are not going to replace therapists anytime soon, she says, “I do believe that in some circumstances computerized therapy can be successful with no human intervention ... in many ways people are not well suited to be therapists.” A computer may be more probing and objective.

Sverrisdottir’s experience suggests that CBTpsych, at least, can make a difference. Under the program’s tutelage, she says, “very slowly, I started to analyze myself when I’m amongst other people.” She identified a pattern of “negative thoughts about people judging me.”

She might have got there with a human therapist, she says. But in the years since she first started talking to a computer about the trouble swirling in her mind, Sverrisdottir says, “I have been able to change it.” ■



Fears of an AI pioneer

Stuart Russell argues that AI is as dangerous as nuclear weapons *By John Bohannon*

From the enraged robots in the 1920 play *R.U.R.* to the homicidal computer H.A.L. in *2001: A Space Odyssey*, science fiction writers have embraced the dark side of artificial intelligence (AI) ever since the concept entered our collective imagination. Sluggish progress in AI research, especially during the “AI winter” of the 1970s and 1980s, made such worries seem far-fetched. But recent breakthroughs in machine learning and vast improvements in computational power have brought a flood of research funding—and fresh concerns about where AI may lead us.



One researcher now speaking up is Stuart Russell, a computer scientist at the University of California, Berkeley, who with Peter Norvig, director of research at Google, wrote the premier AI textbook, *Artificial Intelligence: A Modern Approach*, now in its third edition. Last year, Russell joined the Centre for the Study of Existential Risk at Cambridge University in the United Kingdom as an AI expert focusing on “risks that could lead to human extinction.” Among his chief concerns, which he aired at an April meeting in Geneva, Switzerland, run by the United Nations, is the danger of putting military drones and weaponry under the

full control of AI systems. This interview has been edited for clarity and brevity.

Q: What do you see as a likely path from AI to disaster?

A: The basic scenario is explicit or implicit value misalignment: AI systems [that are] given objectives that don’t take into account all the elements that humans care about. The routes could be varied and complex—corporations seeking a super-technological advantage, countries trying to build [AI systems] before their enemies, or a slow-boiled frog kind of evolution leading to dependency and enfeeblement not unlike E. M. Forster’s *The Machine Stops*.

Q: You’ve grappled with this issue for a long time.

A: My textbook has a section “What If We Do Succeed?” devoted to the question of whether human-level AI or superintelligent systems would be a good idea. [More recent causes for concern are] the rapid developments in AI capabilities such as the legged locomotion of Big Dog [the autonomous robot created by Boston Dynamics, recently acquired by Google] and progress in computer vision.

Q: What needs to be done to prevent an AI catastrophe?

A: First, research into the precise nature of the potential risk and development of

technical approaches to eliminate the risk. Second, modification of the goals of AI and the training of students so that alignment of AI systems with human objectives is central to the field, just as containment is central to the goals of fusion research.

Q: But by the time we were developing nuclear fusion, we already had atomic bombs, whereas the AI threat seems speculative. Are we really at the “fusion” stage of AI research?

A: Here’s what Leo Szilard wrote in 1939 after demonstrating a [nuclear] chain reaction: “We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief.” To those who say, well, we may never get to human-level or superintelligent AI, I would reply: It’s like driving straight toward a cliff and saying, “Let’s hope I run out of gas soon!”

Q: The intention with fission was to create a weapon. The intention with AI is to create a tool: intelligence on tap. Does that explain the reluctance to regulate AI?

A: From the beginning, the primary interest in nuclear technology was the “inexhaustible supply of energy.” The possibility of weapons was also obvious. I think there is a reasonable analogy between unlimited amounts of energy and unlimited amounts of intelligence. Both seem wonderful until one thinks of the possible risks. In neither case will anyone regulate the mathematics. The regulation of nuclear weapons deals with objects and materials, whereas with AI it will be a bewildering variety of software that we cannot yet describe. I’m not aware of any large movement calling for regulation either inside or outside AI, because we don’t know how to write such regulation.

Q: Should we start tracking AI research as we track fissile material? Who should do the policing?

A: I think the right approach is to build the issue directly into how practitioners define what they do. No one in civil engineering talks about “building bridges that don’t fall down.” They just call it “building bridges.” Essentially all fusion researchers work on containment as a matter of course; uncontained fusion reactions just aren’t useful. Right now we have to say “AI that is probably beneficial,” but eventually that will just be called “AI.” [We must] redirect the field away from its current goal of building pure intelligence for its own sake, regardless of the associated objectives and their consequences. ■

POLICY FORUM

Data, privacy, and the greater good

Eric Horvitz^{1*} and Deirdre Mulligan^{2*}

Large-scale aggregate analyses of anonymized data can yield valuable results and insights that address public health challenges and provide new avenues for scientific discovery. These methods can extend our knowledge and provide new tools for enhancing health and wellbeing. However, they raise questions about how to best address potential threats to privacy while reaping benefits for individuals and to society as a whole. The use of machine learning to make leaps across informational and social contexts to infer health conditions and risks from nonmedical data provides representative scenarios for reflections on directions with balancing innovation and regulation.

What if analyzing Twitter tweets or Facebook posts could identify new mothers at risk for postpartum depression (PPD)? Despite PPD's serious consequences, early identification and prevention remain difficult. Absent a history of depression, detection is largely dependent on new mothers' self-reports. But researchers found that shifts in sets of activities and language usage on Facebook are predictors of PPD (1) (see the photo). This is but one example of promising research that uses machine learning to derive and leverage health-related inferences from the massive flows of data about individuals and populations generated through social media and other digital data streams. At the same time, machine learning presents new challenges for protecting individual privacy and ensuring fair use of data. We need to strike a new balance between controls on collecting information and controls on how it is used, as well as pursue auditable and accountable technologies and systems that facilitate greater use-based privacy protections.

Researchers have coined terms, such as digital disease detection (2) and infodemiology (3), to define the new science of harnessing diverse streams of digital information to inform public health and policy, e.g., earlier identification of epidemics, (4) modeling communicability and flow of illness (5), and stratifying individuals at risk for illness (6). This new form of health research can also inform and extend understandings drawn from traditional health records and human subjects research. For example, the detection of adverse drug reactions could be improved by jointly leveraging data from the U.S. Food and Drug Administration's Adverse Event Reporting System and anonymized search logs (7). Search logs can serve as a large-scale sensing system that can be used for drug safety surveillance—pharmacovigilance.

Infodemiology studies are typically large-scale aggregate analyses of anonymized data—publicly disclosed or privately held—that yield results and insights on public health questions across populations. However, some methods and models can be aimed at making inferences about unique individuals that could drive actions, such as alerting or providing digital nudges, to improve individual or public health outcomes.

Although digital nudging shows promise, a recent flare-up in the United Kingdom highlights the privacy concerns it can ignite. A Twitter suicide-prevention application called Good Samaritan monitored individuals' tweets for words and phrases indicating a potential mental health crisis. The app notified the person's followers so they



Machine learning can make “category-jumping” inferences about health. New mother's activities and language usage on social media are predictors of postpartum depression.

could intervene to avert a potential suicide. But the app was shuttered after public outcry drew regulator concern (8). Critics worried the app would encourage online stalkers and bullies to target vulnerable individuals and collected 1200 signatures on a petition arguing that the app breached users' privacy by collecting, processing, and sharing sensitive information. Despite the developers' laudable goal of preventing suicide, the nonprofit was chastised for playing fast and loose with the privacy and mental health of those it was seeking to save (9).

Machine learning can facilitate leaps across informational and social contexts, making “category-jumping” inferences about health conditions or propensities from nonmedical data generated far outside the medical context. The implications for privacy are profound. Category-jumping inferences may reveal attributes or conditions an individual has specifically withheld from others. To protect against such violations, the United States heavily regulates health care privacy. But, although information about health conditions garnered from health care treatment and payment must be handled in a manner that respects patient privacy, machine learning and inference can sidestep many of the existing protections.

Even when not category-jumping, machine learning can be used to draw powerful and compromising inferences from self-disclosed, seemingly benign data or readily observed behavior. These inferences can undermine a basic goal of many privacy laws—to allow individuals to control who knows what about them. Machine learning and inference makes it increasingly difficult for individuals to understand what others can know about them based on what they have explicitly or implicitly shared. And these computer-generated channels of information about health conditions join other technically created fissures in existing legal protections for health privacy. In particular, it is difficult to reliably deidentify publicly shared data sets, given the enormous amount and variety of ancillary data that can be used to reidentify individuals.

The capacities of machine learning expose the fundamental limitations of existing U.S. privacy rules that tie the privacy protection of an individual's health status to specific contexts or specific types of information a priori identified as health information. Health privacy regulations and privacy laws in the United States generally are based on the assumption that the semantics of data are relatively fixed and knowable and reside in isolated contexts. Machine learning techniques can instead be used to infer new meaning within and across contexts and is generally unencumbered by privacy rules in the United States. Using publicly available Twitter posts to infer risk of PPD, for example, does not run afoul of existing privacy law. This might be unsurprising, and seem unproblematic, given that the posts were publicly shared, but there are troubling consequences.

Current privacy laws often do double duty. At a basic level, they limit who has access to information about a person. This implicitly limits the extent to which that information influences decision-making and thus doubles as a limit on the opportunities for information to fuel discrimination. Because of the heightened privacy sensitivities and concerns with health-related discrimination, we have additional laws that regulate the use of health information outside the health care context. U.S. laws specifically limit the use of some health information in ways considered unfair. For example, credit-reporting agencies are generally prohibited from providing medical information to make decisions about employment, credit, or housing. The Americans with Disabilities Act (ADA) prohibits discrimination on the basis of substantial physical or mental disabilities—or even a mistaken belief that an individual suffers from such a disability. If machine learning is used to infer that an individual suffers from a physical or mental impairment, an employer who bases a hiring decision on it, even if the inference is wrong, would violate the law.

But the ADA does not prohibit discrimination based on predispositions for such disabilities (10). Machine learning might discover those, too. In theory, the Genetic Information Non-Discrimination Act (GINA) should fill this gap by protecting people genetically predisposed to a disease. But again, machine learning exposes cracks in this protection. Although GINA prohibits discrimination based on information derived from genetic tests or a family history of a disease (11), it does not limit the use of information

¹Microsoft Research, Redmond, WA 98052, USA. ²University of California, Berkeley, Berkeley, CA 94720, USA.

*Corresponding author. E-mail: horvitz@microsoft.com (E.H.); dmulligan@berkeley.edu (D.M.)

about such a disposition—even if it is grounded in genetics—inferred through machine learning techniques that mine other sorts of data. In other words, machine learning that predicts future health status from nongenetic information—including health status changes due to genetic predisposition—would circumvent existing legal protections (12).

Just as machine learning can expose secrets, it facilitates social sorting—placing individuals into categories for differential treatment—with good or bad intent and positive or negative outcomes. The methods used to classify individuals as part of beneficial public health programs and nudges can just as easily be used for more nefarious purposes, such as discrimination to protect organizational profits.

Policy-makers in the United States and elsewhere are just beginning to address the challenges that machine learning and inference pose to commitments to privacy and equal treatment. Although not specifically focused on health information, reports issued by the White House—discuss the potential for large-scale data analyses to result in discrimination (13)—and the Federal Trade Commission (FTC) have suggested new efforts to protect privacy, regulate harmful uses of information, and increase transparency.

The FTC is the key agency policing unfair and deceptive practices in the commercial marketplace, including those that touch on the privacy and security of personal information. Its proposed privacy framework encourages companies to combine technical and policy mechanisms to protect against reidentification. The FTC's proposed rules would work to ensure that data are both “not reasonably identifiable” and accompanied by public company commitments not to reidentify it. The same privacy rules should apply to downstream users of the data (14). This approach is promising for machine learning and other areas of artificial intelligence that rely on data-centric analyses. It allows learning from large data sets—and sharing them—by encouraging companies to reduce the risks that data pools and data sharing pose for individual privacy.

The FTC proposal grows, in part, from recent agency actions focused on inferences that we have deemed “context-jumping.” In one high-profile case, Netflix publicly released data sets to support a competition to improve their recommendation algorithm. When outside researchers used ancillary data to reidentify and infer sensitive attributes about individuals from the Netflix data sets, the FTC worked with the company to limit future public disclosures—setting out the limits discussed above. In a similar vein, the FTC objected to a change in Facebook's defaults that exposed individuals' group affiliations from which sensitive information, such as political views and sexual orientation, could be inferred (15).

Additionally, the FTC has made efforts to ensure that individuals can control tracking in the online and mobile environments. These are in part due to the nonobvious inferences that can be drawn from vast collections of data (16–18) and the subsequent risks to consumers, who may be placed in classifications that single them out for specific treatment in the marketplace (19, 20). In a related context, the FTC recommended that Congress require data brokers—companies that collect consumers' personal

information and resell or share that information with others—to clearly disclose to consumers information about the data they collect, as well as the fact that they derive inferences from it (21). Here, too, the FTC appears concerned with not just the raw data, but inferences from its analysis.

The Obama Administration's Big Data Initiative has also considered the risks to privacy posed by machine learning and the potential downsides of using machine inferences in the commercial marketplace (22, 23), concluding that we need to update our privacy rules, increase technical expertise in consumer protection and civil rights agencies to address novel discrimination issues arising from big data, provide individuals with privacy preserving tools that allow them to control the collection and manage the use of personal information, as well as increase transparency into how companies use and trade data. The Administration is also concerned with the use of machine learning in policing and national security. The White House report called for increased technical expertise to help civil rights and consumer protection agencies identify, investigate, and resolve uses of big data analytics that have a discriminatory impact on protected classes (24).

Note that reports and proposals from the Administration distinctly emphasize policies and regulations focused on data use rather than collection. While acknowledging the need for tools that allow consumers to control when and how their data is collected, the Administration recommendations focus on empowering individuals to participate in decisions about future uses and disclosures of collected data (25). A separate report by the President's Council of Advisors on Science and Technology (PCAST) concluded that this was a more fruitful direction for technical protections. Both reports suggest that use-based protections better address the latent meaning of data—inferences drawn from data using machine learning—and can adapt to the scale of the data-rich and connected environment of the future (26). The Administration called for collaborative efforts to ensure that regulations in the health context will allow society to reap the benefits and mitigate the risks posed by machine learning and inferences. Use-based approaches are often favored by industry, as well, which tends to view data as akin to a natural resource to be mined for commercial and public benefit, and industry is resistant to efforts to constrain data collection.

Although incomplete and unlikely to be acted upon by the current gridlocked Congress, adoption of these recommendations would increase transparency about data's collection, use, and consequences. Along with efforts to identify and constrain discriminatory or unfair uses of data and inferences, they are promising steps. They also align with aspects of existing European privacy laws concerned with the transparency and fairness of data processing, particularly the risks to individuals of purely automated decision-making.

Current European Union (EU) law requires entities to provide individuals with access to the data on which decisions are rendered, as well as information about decision criteria [see Articles 12 and 15 of (27)]. Although currently governed by a Europe-wide directive, both provisions are a matter of na-

tional law. What exactly individuals receive when they request access to their data and to processing logic varies by country, as does the implementation of the limitation on “purely automated” processing. The EU is expected to adopt a data privacy regulation that will supplant local law, with a single national standard. Although the current draft includes parallel provisions, their final form is not yet known nor is how they will ultimately be interpreted (27). In theory, a new EU requirement to disclose the logic of processing could apply quite broadly, with implications for public access to data analytics and algorithms. In the interim, a decision expected this summer in a case before the European Court of Justice may provide some detail as to what level of access to both data and the logic of processing is currently required under the EU Directive (28).

Improving the transparency of data processing to data subjects is both important and challenging. Although the goal may be to promote actual understanding of the workings or likely outputs of machine learning and reasoning methods, the workflows and dynamism of algorithms and decision criteria may be difficult to characterize and explain. For example, popular convolutional neural-network learning procedures (commonly referred to as “deep learning”) automatically induce rich, multilayered representations that their developers themselves may not understand with clarity. Although high-level descriptions of procedures and representations might be provided, even an accomplished programmer with access to the source code would be unable to describe the precise operation of such a system or predict the output of a given set of inputs.

Data's meaning has become a moving target. Data sets can be easily combined to reidentify data sets thought deidentified, and sensitive knowledge can be inferred from benign data that are routinely and promiscuously shared. These pose difficulties for current U.S. legal approaches to privacy protection that regulate data on the basis of its identifiability and express meaning.

Use-based approaches are driven, in part, by the realization that focusing solely on limiting data collection is inadequate. In a way, this presupposes that data are an unalloyed good that should be collected on principle, whenever and wherever possible. Whereas we are not ready to abandon limits on data collection, we agree that use-based regulations, although challenging to implement, are an important part of the future legal landscape—and will help to advance privacy, equality, and the public good. To advance transparency and to balance the constraints they impose, use-based approaches would need to emphasize access, accuracy, and correction rights for individuals.

The evolution of regulations for health information, although incomplete, provides a useful map for thinking about the challenges and opportunities we face today and frames potential solutions. In health care, privacy rules were joined by nondiscrimination rules and always were accompanied by special provisions to support research. Today, they are being joined by collective governance models designed to encourage pooling of data in biobanks that support research on health conditions while protecting collective interests in privacy.

Despite practical challenges, we are hopeful that informed discussions among policy-makers and the public about data and the capabilities of machine learning, will lead to insightful designs of programs and policies that can balance the goals of protecting privacy and ensuring fairness with those of reaping the benefits to scientific research and to individual and public health. Our commitments to privacy and fairness are evergreen, but our policy choices must adapt to advance them, and support new techniques for deepening our knowledge.

REFERENCES AND NOTES

1. M. De Choudhury, S. Counts, E. Horvitz, A. Hoff, in *Proceedings of International Conference on Weblogs and Social Media* [Association for the Advancement of Artificial Intelligence (AAAI), Palo Alto, CA, 2014].
2. J. S. Brownstein, C. C. Freifeld, L. C. Madoff, *N. Engl. J. Med.* **360**, 2153–2155 (2009).
3. G. Eysenbach, *J. Med. Internet Res.* **11**, e11 (2009).
4. D. A. Broniatowski, M. J. Paul, M. Dredze, *PLOS ONE* **8**, e83672 (2013).
5. A. Sadilek, H. Kautz, V. Silenzio, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, CA, 2012).
6. M. De Choudhury, S. Counts, E. Horvitz, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, 2013), pp. 3267–3276.
7. R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, E. Horvitz, *Clin. Pharmacol. Ther.* **96**, 239–246 (2014).
8. Samaritans Radar; www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar.
9. Shut down Samaritans Radar; <http://bit.ly/Samaritans-after>.
10. U.S. Equal Employment Opportunity Commission (EEOC), 29 Code of Federal Regulations (C.F.R.), 1630.2 (g) (2013).
11. EEOC, 29 CFR 1635.3 (c) (2013).
12. M. A. Rothstein, *J. Law Med. Ethics* **36**, 837–840 (2008).
13. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); <http://1.usa.gov/1TS0hiG>.
14. Letter from Maneesha Mithal, FTC, to Reed Freeman, Morrison, & Foerster LLP, Counsel for Netflix, 2 [closing letter] (2010); <http://1.usa.gov/1GCFyXR>.
15. In re Facebook, Complaint, FTC File No. 092 3184 (2012).
16. FTC Staff Report, *Mobile Privacy Disclosures: Building Trust Through Transparency* (FTC, Washington, DC, 2013); <http://1.usa.gov/1eNz8zr>.
17. FTC, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (FTC, Washington, DC, 2012).
18. Directive 95/46/ec of the European Parliament and of The Council of Europe, 24 October 1995.
19. L. Sweeney, Online ads roll the dice [blog]; <http://1.usa.gov/1KgEcYg>.
20. FTC, “Big data: A tool for inclusion or exclusion?” (workshop, FTC, Washington, DC, 2014); <http://1.usa.gov/1SR65cv>.
21. FTC, *Data Brokers: A Call for Transparency and Accountability* (FTC, Washington, DC, 2014); <http://1.usa.gov/1GCF0j5>.
22. J. Podesta, “Big data and privacy: 1 year out” [blog]; <http://bit.ly/WHsePrivacy>.
23. White House Council of Economic Advisers, *Big Data and Differential Pricing* (White House, Washington, DC, 2015).
24. Executive Office of the President, *Big Data and Differential Processing* (White House, Washington, DC, 2015); <http://1.usa.gov/1eN7qR>.
25. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); <http://1.usa.gov/1TS0hiG>.
26. President’s Council of Advisors on Science and Technology (PCAST), *Big Data and Privacy: A Technological Perspective* (White House, Washington, DC, 2014); <http://1.usa.gov/1C5ewNv>.
27. European Commission, Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), COM(2012) 11 final (2012); <http://bit.ly/1Lu5POV>.
28. *M. Schrems v. Facebook Ireland Limited*, *§1*. Unlawful data transmission to the U.S.A. (“PRISM”), ¶166 and 167 (2013); www.europe-v-facebook.org/sk/sk_en.pdf.

REVIEW

Machine learning: Trends, perspectives, and prospects

M. I. Jordan^{1*} and T. M. Mitchell^{2*}

Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today’s most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.

Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software it has produced and fielded across many applications.

Machine learning has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use. Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. Many developers of AI systems now recognize that, for many applications, it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs. The effect of machine learning has also been felt broadly across computer science and across a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistics chains. There has been a similarly broad range of effects across empirical sciences, from biology to cosmology to social science, as machine-learning methods have been developed to analyze high-throughput experimental data in novel ways. See Fig. 1 for a depiction of some recent areas of application of machine learning.

A learning problem can be defined as the problem of improving some measure of perform-

ance when executing some task, through some type of training experience. For example, in learning to detect credit-card fraud, the task is to assign a label of “fraud” or “not fraud” to any given credit-card transaction. The performance metric to be improved might be the accuracy of this fraud classifier, and the training experience might consist of a collection of historical credit-card transactions, each labeled in retrospect as fraudulent or not. Alternatively, one might define a different performance metric that assigns a higher penalty when “fraud” is labeled “not fraud” than when “not fraud” is incorrectly labeled “fraud.” One might also define a different type of training experience—for example, by including unlabeled credit-card transactions along with labeled examples.

A diverse array of machine-learning algorithms has been developed to cover the wide variety of data and problem types exhibited across different machine-learning problems (1, 2). Conceptually, machine-learning algorithms can be viewed as searching through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric. Machine-learning algorithms vary greatly, in part by the way in which they represent candidate programs (e.g., decision trees, mathematical functions, and general programming languages) and in part by the way in which they search through this space of programs (e.g., optimization algorithms with well-understood convergence guarantees and evolutionary search methods that evaluate successive generations of randomly mutated programs). Here, we focus on approaches that have been particularly successful to date.

Many algorithms focus on function approximation problems, where the task is embodied in a function (e.g., given an input transaction, output a “fraud” or “not fraud” label), and the learning problem is to improve the accuracy of that function, with experience consisting of a sample of known input-output pairs of the function. In some cases, the function is represented explicitly as a parameterized functional form; in other cases, the function is implicit and obtained via a search process, a factorization, an optimization

¹Department of Electrical Engineering and Computer Sciences, Department of Statistics, University of California, Berkeley, CA, USA. ²Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA.

*Corresponding author. E-mail: jordan@cs.berkeley.edu (M.I.J.); tom.mitchell@cs.cmu.edu (T.M.M.)

Despite practical challenges, we are hopeful that informed discussions among policy-makers and the public about data and the capabilities of machine learning, will lead to insightful designs of programs and policies that can balance the goals of protecting privacy and ensuring fairness with those of reaping the benefits to scientific research and to individual and public health. Our commitments to privacy and fairness are evergreen, but our policy choices must adapt to advance them, and support new techniques for deepening our knowledge.

REFERENCES AND NOTES

1. M. De Choudhury, S. Counts, E. Horvitz, A. Hoff, in *Proceedings of International Conference on Weblogs and Social Media* [Association for the Advancement of Artificial Intelligence (AAAI), Palo Alto, CA, 2014].
2. J. S. Brownstein, C. C. Freifeld, L. C. Madoff, *N. Engl. J. Med.* **360**, 2153–2155 (2009).
3. G. Eysenbach, *J. Med. Internet Res.* **11**, e11 (2009).
4. D. A. Broniatowski, M. J. Paul, M. Dredze, *PLOS ONE* **8**, e83672 (2013).
5. A. Sadilek, H. Kautz, V. Silenzio, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, CA, 2012).
6. M. De Choudhury, S. Counts, E. Horvitz, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, 2013), pp. 3267–3276.
7. R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, E. Horvitz, *Clin. Pharmacol. Ther.* **96**, 239–246 (2014).
8. Samaritans Radar; www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar.
9. Shut down Samaritans Radar; <http://bit.ly/Samaritans-after>.
10. U.S. Equal Employment Opportunity Commission (EEOC), 29 Code of Federal Regulations (C.F.R.), 1630.2 (g) (2013).
11. EEOC, 29 CFR 1635.3 (c) (2013).
12. M. A. Rothstein, *J. Law Med. Ethics* **36**, 837–840 (2008).
13. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); <http://1.usa.gov/1TSoHiG>.
14. Letter from Maneesha Mithal, FTC, to Reed Freeman, Morrison, & Foerster LLP, Counsel for Netflix, 2 [closing letter] (2010); <http://1.usa.gov/1GCFyXR>.
15. In re Facebook, Complaint, FTC File No. 092 3184 (2012).
16. FTC Staff Report, *Mobile Privacy Disclosures: Building Trust Through Transparency* (FTC, Washington, DC, 2013); <http://1.usa.gov/1eNz8zr>.
17. FTC, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (FTC, Washington, DC, 2012).
18. Directive 95/46/ec of the European Parliament and of The Council of Europe, 24 October 1995.
19. L. Sweeney, Online ads roll the dice [blog]; <http://1.usa.gov/1KgEcYg>.
20. FTC, “Big data: A tool for inclusion or exclusion?” (workshop, FTC, Washington, DC, 2014); <http://1.usa.gov/1SR65cv>.
21. FTC, *Data Brokers: A Call for Transparency and Accountability* (FTC, Washington, DC, 2014); <http://1.usa.gov/1GCFoJ5>.
22. J. Podesta, “Big data and privacy: 1 year out” [blog]; <http://bit.ly/WHsePrivacy>.
23. White House Council of Economic Advisers, *Big Data and Differential Pricing* (White House, Washington, DC, 2015).
24. Executive Office of the President, *Big Data and Differential Processing* (White House, Washington, DC, 2015); <http://1.usa.gov/1eNz7qR>.
25. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); <http://1.usa.gov/1TSoHiG>.
26. President’s Council of Advisors on Science and Technology (PCAST), *Big Data and Privacy: A Technological Perspective* (White House, Washington, DC, 2014); <http://1.usa.gov/1C5ewNv>.
27. European Commission, Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), COM(2012) 11 final (2012); <http://bit.ly/1Lu5POV>.
28. *M. Schrems v. Facebook Ireland Limited*, *S.J.* Unlawful data transmission to the U.S.A. (“PRISM”), ¶166 and 167 (2013); www.europe-v-facebook.org/sk/sk_en.pdf.

10.1126/science.aac4520

REVIEW

Machine learning: Trends, perspectives, and prospects

M. I. Jordan^{1*} and T. M. Mitchell^{2*}

Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today’s most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.

Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software it has produced and fielded across many applications.

Machine learning has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use. Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. Many developers of AI systems now recognize that, for many applications, it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs. The effect of machine learning has also been felt broadly across computer science and across a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistics chains. There has been a similarly broad range of effects across empirical sciences, from biology to cosmology to social science, as machine-learning methods have been developed to analyze high-throughput experimental data in novel ways. See Fig. 1 for a depiction of some recent areas of application of machine learning.

A learning problem can be defined as the problem of improving some measure of perform-

ance when executing some task, through some type of training experience. For example, in learning to detect credit-card fraud, the task is to assign a label of “fraud” or “not fraud” to any given credit-card transaction. The performance metric to be improved might be the accuracy of this fraud classifier, and the training experience might consist of a collection of historical credit-card transactions, each labeled in retrospect as fraudulent or not. Alternatively, one might define a different performance metric that assigns a higher penalty when “fraud” is labeled “not fraud” than when “not fraud” is incorrectly labeled “fraud.” One might also define a different type of training experience—for example, by including unlabeled credit-card transactions along with labeled examples.

A diverse array of machine-learning algorithms has been developed to cover the wide variety of data and problem types exhibited across different machine-learning problems (1, 2). Conceptually, machine-learning algorithms can be viewed as searching through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric. Machine-learning algorithms vary greatly, in part by the way in which they represent candidate programs (e.g., decision trees, mathematical functions, and general programming languages) and in part by the way in which they search through this space of programs (e.g., optimization algorithms with well-understood convergence guarantees and evolutionary search methods that evaluate successive generations of randomly mutated programs). Here, we focus on approaches that have been particularly successful to date.

Many algorithms focus on function approximation problems, where the task is embodied in a function (e.g., given an input transaction, output a “fraud” or “not fraud” label), and the learning problem is to improve the accuracy of that function, with experience consisting of a sample of known input-output pairs of the function. In some cases, the function is represented explicitly as a parameterized functional form; in other cases, the function is implicit and obtained via a search process, a factorization, an optimization

¹Department of Electrical Engineering and Computer Sciences, Department of Statistics, University of California, Berkeley, CA, USA. ²Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA.

*Corresponding author. E-mail: jordan@cs.berkeley.edu (M.I.J.); tom.mitchell@cs.cmu.edu (T.M.M.)

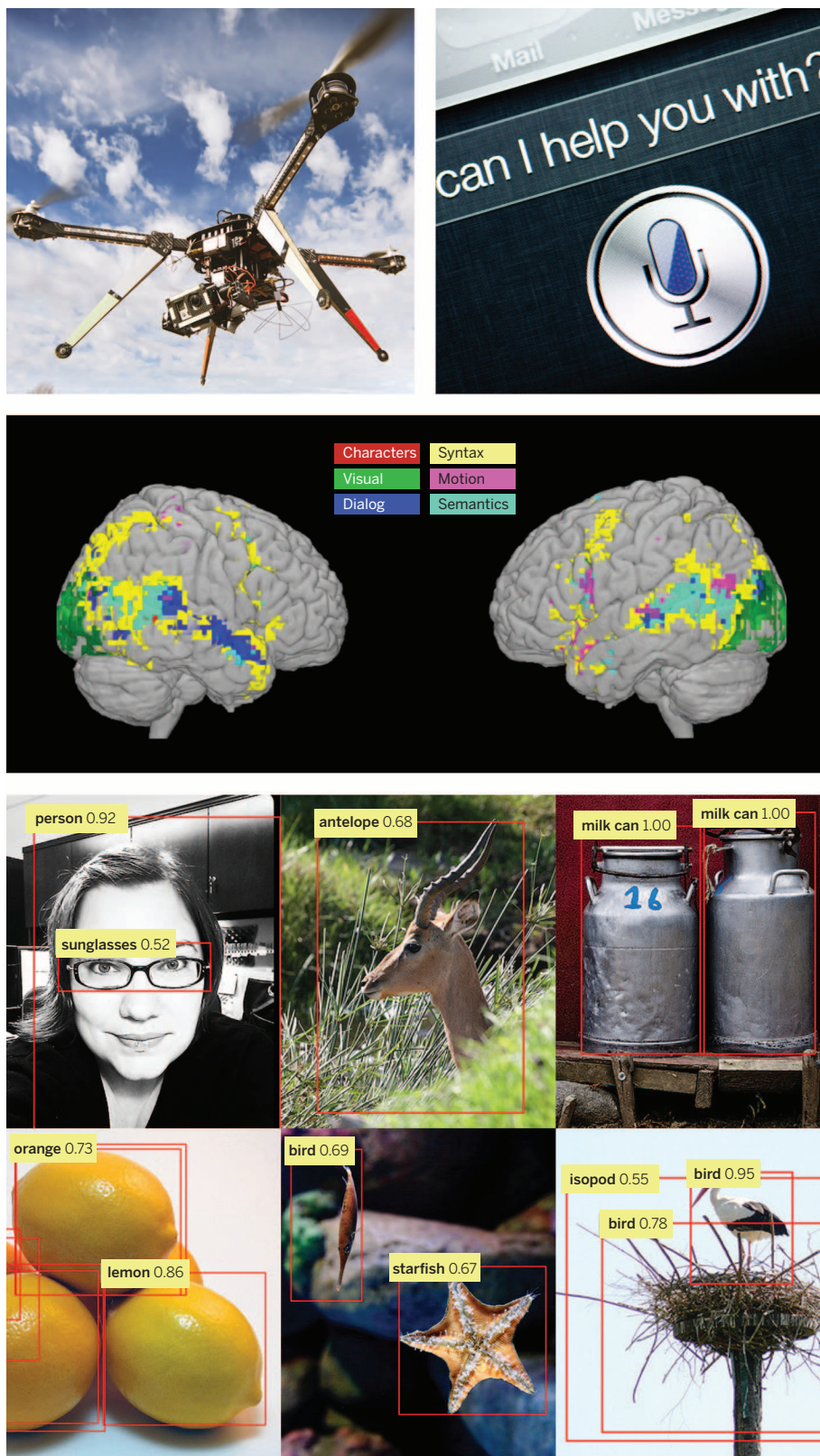


Fig. 1. Applications of machine learning. Machine learning is having a substantial effect on many areas of technology and science; examples of recent applied success stories include robotics and autonomous vehicle control (top left), speech processing and natural language processing (top right), neuroscience research (middle), and applications in computer vision (bottom). [The middle panel is adapted from (29). The images in the bottom panel are from the ImageNet database; object recognition annotation is by R. Girshick.]

procedure, or a simulation-based procedure. Even when implicit, the function generally depends on parameters or other tunable degrees of freedom, and training corresponds to finding values for these parameters that optimize the performance metric.

Whatever the learning algorithm, a key scientific and practical goal is to theoretically characterize the capabilities of specific learning algorithms and the inherent difficulty of any given learning problem: How accurately can the algorithm learn from a particular type and volume of training data? How robust is the algorithm to errors in its modeling assumptions or to errors in the training data? Given a learning problem with a given volume of training data, is it possible to design a successful algorithm or is this learning problem fundamentally intractable? Such theoretical characterizations of machine-learning algorithms and problems typically make use of the familiar frameworks of statistical decision theory and computational complexity theory. In fact, attempts to characterize machine-learning algorithms theoretically have led to blends of statistical and computational theory in which the goal is to simultaneously characterize the sample complexity (how much data are required to learn accurately) and the computational complexity (how much computation is required) and to specify how these depend on features of the learning algorithm such as the representation it uses for what it learns (3–6). A specific form of computational analysis that has proved particularly useful in recent years has been that of optimization theory, with upper and lower bounds on rates of convergence of optimization procedures merging well with the formulation of machine-learning problems as the optimization of a performance metric (7, 8).

As a field of study, machine learning sits at the crossroads of computer science, statistics and a variety of other disciplines concerned with automatic improvement over time, and inference and decision-making under uncertainty. Related disciplines include the psychological study of human learning, the study of evolution, adaptive control theory, the study of educational practices, neuroscience, organizational behavior, and economics. Although the past decade has seen increased cross-talk with these other fields, we are just beginning to tap the potential synergies and the diversity of formalisms and experimental methods used across these multiple fields for studying systems that improve with experience.

Drivers of machine-learning progress

The past decade has seen rapid growth in the ability of networked and mobile computing systems to gather and transport vast amounts of data, a phenomenon often referred to as “Big Data.” The scientists and engineers who collect such data have often turned to machine learning for solutions to the problem of obtaining useful insights, predictions, and decisions from such data sets. Indeed, the sheer size of the data makes it essential to develop scalable procedures that blend computational and statistical

considerations, but the issue is more than the mere size of modern data sets; it is the granular, personalized nature of much of these data. Mobile devices and embedded computing permit large amounts of data to be gathered about individual humans, and machine-learning algorithms can learn from these data to customize their services to the needs and circumstances of each individual. Moreover, these personalized services can be connected, so that an overall service emerges that takes advantage of the wealth and diversity of data from many individuals while still customizing to the needs and circumstances of each. Instances of this trend toward capturing and mining large quantities of data to improve services and productivity can be found across many fields of commerce, science, and government. Historical medical records are used to discover which patients will respond best to which treatments; historical traffic data are used to improve traffic control and reduce congestion; historical crime data are used to help allocate local police to specific locations at specific times; and large experimental data sets are captured and curated to accelerate progress in biology, astronomy, neuroscience, and other data-intensive empirical sciences. We appear to be at the beginning of a decades-long trend toward increasingly data-intensive, evidence-based decision-making across many aspects of science, commerce, and government.

With the increasing prominence of large-scale data in all areas of human endeavor has come a wave of new demands on the underlying machine-learning algorithms. For example, huge data sets require computationally tractable algorithms, highly personal data raise the need for algorithms that minimize privacy effects, and the availability of huge quantities of unlabeled data raises the challenge of designing learning algorithms to take advantage of it. The next sections survey some of the effects of these demands on recent

work in machine-learning algorithms, theory, and practice.

Core methods and recent progress

The most widely used machine-learning methods are supervised learning methods (1). Supervised learning systems, including spam classifiers of e-mail, face recognizers over images, and medical diagnosis systems for patients, all exemplify the function approximation problem discussed earlier, where the training data take the form of a collection of (x, y) pairs and the goal is to produce a prediction y^* in response to a query x^* . The inputs x may be classical vectors or they may be more complex objects such as documents, images, DNA sequences, or graphs. Similarly, many different kinds of output y have been studied. Much progress has been made by focusing on the simple binary classification problem in which y takes on one of two values (for example, “spam” or “not spam”), but there has also been abundant research on problems such as multiclass classification (where y takes on one of K labels), multilabel classification (where y is labeled simultaneously by several of the K labels), ranking problems (where y provides a partial order on some set), and general structured prediction problems (where y is a combinatorial object such as a graph, whose components may be required to satisfy some set of constraints). An example of the latter problem is part-of-speech tagging, where the goal is to simultaneously label every word in an input sentence x as being a noun, verb, or some other part of speech. Supervised learning also includes cases in which y has real-valued components or a mixture of discrete and real-valued components.

Supervised learning systems generally form their predictions via a learned mapping $f(x)$, which produces an output y for each input x (or a probability distribution over y given x). Many different forms of mapping f exist, including

decision trees, decision forests, logistic regression, support vector machines, neural networks, kernel machines, and Bayesian classifiers (1). A variety of learning algorithms has been proposed to estimate these different types of mappings, and there are also generic procedures such as boosting and multiple kernel learning that combine the outputs of multiple learning algorithms. Procedures for learning f from data often make use of ideas from optimization theory or numerical analysis, with the specific form of machine-learning problems (e.g., that the objective function or function to be integrated is often the sum over a large number of terms) driving innovations. This diversity of learning architectures and algorithms reflects the diverse needs of applications, with different architectures capturing different kinds of mathematical structures, offering different levels of amenability to post-hoc visualization and explanation, and providing varying trade-offs between computational complexity, the amount of data, and performance.

One high-impact area of progress in supervised learning in recent years involves deep networks, which are multilayer networks of threshold units, each of which computes some simple parameterized function of its inputs (9, 10). Deep learning systems make use of gradient-based optimization algorithms to adjust parameters throughout such a multilayered network based on errors at its output. Exploiting modern parallel computing architectures, such as graphics processing units originally developed for video gaming, it has been possible to build deep learning systems that contain billions of parameters and that can be trained on the very large collections of images, videos, and speech samples available on the Internet. Such large-scale deep learning systems have had a major effect in recent years in computer vision (11) and speech recognition (12), where they have yielded major improvements in performance over previous approaches

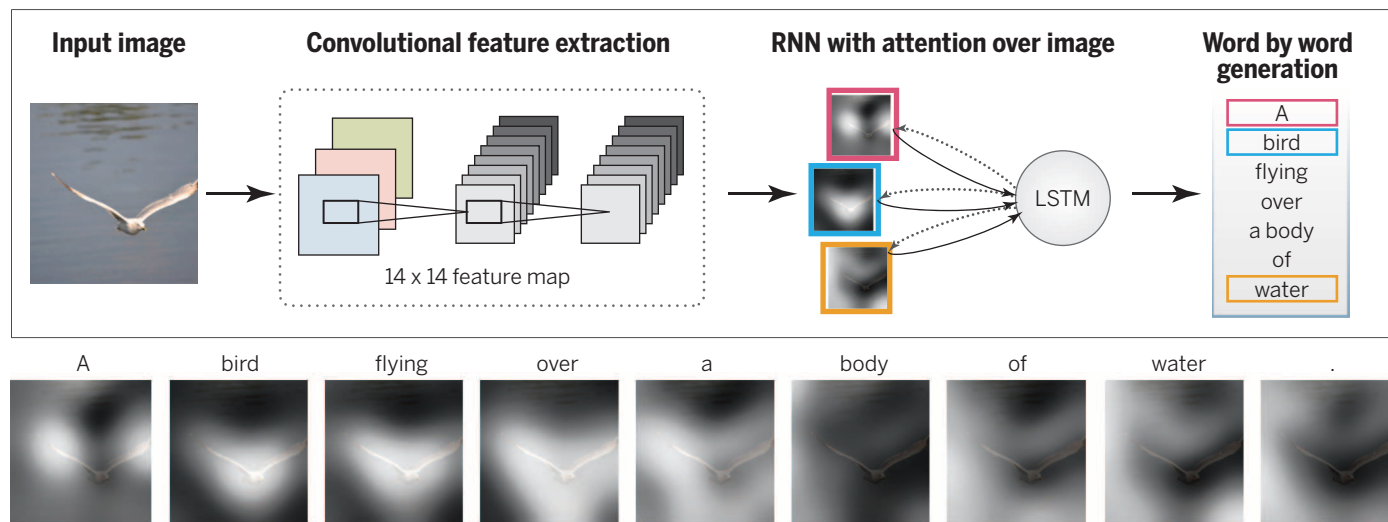


Fig. 2. Automatic generation of text captions for images with deep networks. A convolutional neural network is trained to interpret images, and its output is then used by a recurrent neural network trained to generate a text caption (top). The sequence at the bottom shows the word-by-word focus of the network on different parts of input image while it generates the caption word-by-word. [Adapted with permission from (30)]

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

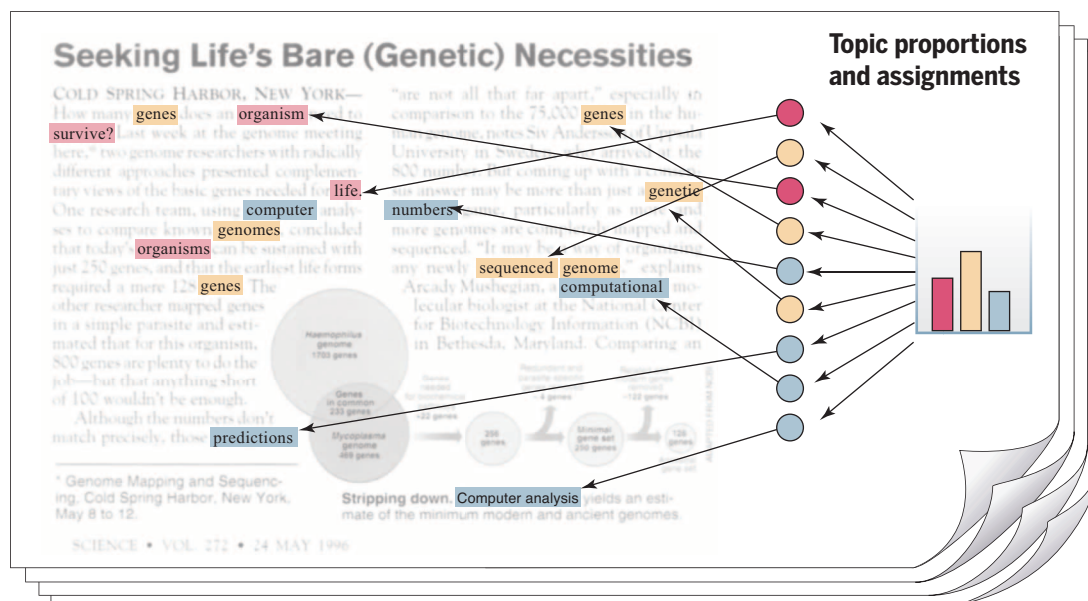


Fig. 3. Topic models. Topic modeling is a methodology for analyzing documents, where a document is viewed as a collection of words, and the words in the document are viewed as being generated by an underlying set of topics (denoted by the colors in the figure). Topics are probability distributions across words (leftmost column), and each document is characterized by a probability distribution across topics (histogram). These distributions are inferred based on the analysis of a collection of documents and can be viewed to classify, index, and summarize the content of documents. [From (31). Copyright 2012, Association for Computing Machinery, Inc. Reprinted with permission]

(see Fig. 2). Deep network methods are being actively pursued in a variety of additional applications from natural language translation to collaborative filtering.

The internal layers of deep networks can be viewed as providing learned representations of the input data. While much of the practical success in deep learning has come from supervised learning methods for discovering such representations, efforts have also been made to develop deep learning algorithms that discover useful representations of the input without the need for labeled training data (13). The general problem is referred to as unsupervised learning, a second paradigm in machine-learning research (2).

Broadly, unsupervised learning generally involves the analysis of unlabeled data under assumptions about structural properties of the data (e.g., algebraic, combinatorial, or probabilistic). For example, one can assume that data lie on a low-dimensional manifold and aim to identify that manifold explicitly from data. Dimension reduction methods—including principal components analysis, manifold learning, factor analysis, random projections, and autoencoders (1, 2)—make different specific assumptions regarding the underlying manifold (e.g., that it is a linear subspace, a smooth nonlinear manifold, or a collection of submanifolds). Another example of dimension reduction is the topic modeling framework depicted in Fig. 3. A criterion function is defined that embodies these assumptions—often making use of general statistical principles such as maximum likelihood, the method of moments, or Bayesian integration—and optimization or sampling algo-

ritms are developed to optimize the criterion. As another example, clustering is the problem of finding a partition of the observed data (and a rule for predicting future data) in the absence of explicit labels indicating a desired partition. A wide range of clustering procedures has been developed, all based on specific assumptions regarding the nature of a “cluster.” In both clustering and dimension reduction, the concern with computational complexity is paramount, given that the goal is to exploit the particularly large data sets that are available if one dispenses with supervised labels.

A third major machine-learning paradigm is reinforcement learning (14, 15). Here, the information available in the training data is intermediate between supervised and unsupervised learning. Instead of training examples that indicate the correct output for a given input, the training data in reinforcement learning are assumed to provide only an indication as to whether an action is correct or not; if an action is incorrect, there remains the problem of finding the correct action. More generally, in the setting of sequences of inputs, it is assumed that reward signals refer to the entire sequence; the assignment of credit or blame to individual actions in the sequence is not directly provided. Indeed, although simplified versions of reinforcement learning known as bandit problems are studied, where it is assumed that rewards are provided after each action, reinforcement learning problems typically involve a general control-theoretic setting in which the learning task is to learn a control strategy (a “policy”) for an agent acting in an unknown dynamical environment, where that learned strat-

egy is trained to choose actions for any given state, with the objective of maximizing its expected reward over time. The ties to research in control theory and operations research have increased over the years, with formulations such as Markov decision processes and partially observed Markov decision processes providing points of contact (15, 16). Reinforcement-learning algorithms generally make use of ideas that are familiar from the control-theory literature, such as policy iteration, value iteration, rollouts, and variance reduction, with innovations arising to address the specific needs of machine learning (e.g., large-scale problems, few assumptions about the unknown dynamical environment, and the use of supervised learning architectures to represent policies). It is also worth noting the strong ties between reinforcement learning and many decades of work on learning in psychology and neuroscience, one notable example being the use of reinforcement learning algorithms to predict the response of dopaminergic neurons in monkeys learning to associate a stimulus light with subsequent sugar reward (17).

Although these three learning paradigms help to organize ideas, much current research involves blends across these categories. For example, semi-supervised learning makes use of unlabeled data to augment labeled data in a supervised learning context, and discriminative training blends architectures developed for unsupervised learning with optimization formulations that make use of labels. Model selection is the broad activity of using training data not only to fit a model but also to select from a family of models, and the fact that training data do not directly indicate

which model to use leads to the use of algorithms developed for bandit problems and to Bayesian optimization procedures. Active learning arises when the learner is allowed to choose data points and query the trainer to request targeted information, such as the label of an otherwise unlabeled example. Causal modeling is the effort to go beyond simply discovering predictive relations among variables, to distinguish which variables causally influence others (e.g., a high white-blood-cell count can predict the existence of an infection, but it is the infection that causes the high white-cell count). Many issues influence the design of learning algorithms across all of these paradigms, including whether data are available in batches or arrive sequentially over time, how data have been sampled, requirements that learned models be interpretable by users, and robustness issues that arise when data do not fit prior modeling assumptions.

Emerging trends

The field of machine learning is sufficiently young that it is still rapidly expanding, often by inventing new formalizations of machine-learning problems driven by practical applications. (An example is the development of recommendation systems, as described in Fig. 4.) One major trend driving this expansion is a growing concern with the environment in which a machine-learning algorithm operates. The word “environment” here refers in part to the computing architecture; whereas a classical machine-learning system involved a single program running on a single machine, it is now common for machine-learning systems to be deployed in architectures that include many thousands or ten of thousands of processors, such that communication constraints and issues of parallelism and distributed processing take center stage. Indeed, as depicted in Fig. 5, machine-learning systems are increasingly taking the form of complex collections of software that run on large-scale parallel and distributed computing platforms and provide a range of algorithms and services to data analysts.

The word “environment” also refers to the source of the data, which ranges from a set of people who may have privacy or ownership concerns, to the analyst or decision-maker who may have certain requirements on a machine-learning system (for example, that its output be visualizable), and to the social, legal, or political framework surrounding the deployment of a system. The environment also may include other machine-learning systems or other agents, and the overall collection of systems may be cooperative or adversarial. Broadly speaking, environments provide various resources to a learning algorithm and place constraints on those resources. Increasingly, machine-learning researchers are formalizing these relationships, aiming to design algorithms that are provably effective in various environments and explicitly allow users to express and control trade-offs among resources.

As an example of resource constraints, let us suppose that the data are provided by a set of individuals who wish to retain a degree of pri-

vacy. Privacy can be formalized via the notion of “differential privacy,” which defines a probabilistic channel between the data and the outside world such that an observer of the output of the channel cannot infer reliably whether particular individuals have supplied data or not (18). Classical applications of differential privacy have involved insuring that queries (e.g., “what is the maximum balance across a set of accounts?”) to a privatized database return an answer that is close to that returned on the nonprivate data. Recent research has brought differential privacy into contact with machine learning, where queries involve predictions or other inferential assertions (e.g., “given the data I’ve seen so far, what is the probability that a new transaction is fraudulent?”) (19, 20). Placing the overall design of a privacy-enhancing machine-learning system within a decision-theoretic framework provides users with a tuning knob whereby they can choose a desired level of privacy that takes into account the kinds of questions that will be asked of the data and their own personal utility for the answers. For example, a person may be willing to

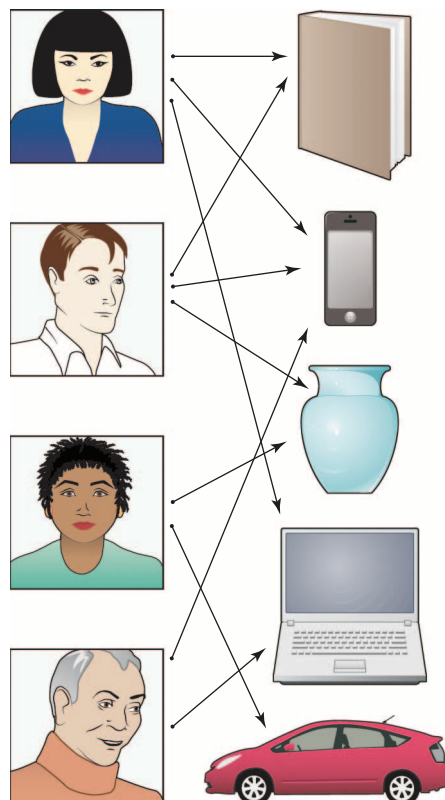


Fig. 4. Recommendation systems. A recommendation system is a machine-learning system that is based on data that indicate links between a set of a users (e.g., people) and a set of items (e.g., products). A link between a user and a product means that the user has indicated an interest in the product in some fashion (perhaps by purchasing that item in the past). The machine-learning problem is to suggest other items to a given user that he or she may also be interested in, based on the data across all users.

reveal most of their genome in the context of research on a disease that runs in their family but may ask for more stringent protection if information about their genome is being used to set insurance rates.

Communication is another resource that needs to be managed within the overall context of a distributed learning system. For example, data may be distributed across distinct physical locations because their size does not allow them to be aggregated at a single site or because of administrative boundaries. In such a setting, we may wish to impose a bit-rate communication constraint on the machine-learning algorithm. Solving the design problem under such a constraint will generally show how the performance of the learning system degrades under decrease in communication bandwidth, but it can also reveal how the performance improves as the number of distributed sites (e.g., machines or processors) increases, trading off these quantities against the amount of data (21, 22). Much as in classical information theory, this line of research aims at fundamental lower bounds on achievable performance and specific algorithms that achieve those lower bounds.

A major goal of this general line of research is to bring the kinds of statistical resources studied in machine learning (e.g., number of data points, dimension of a parameter, and complexity of a hypothesis class) into contact with the classical computational resources of time and space. Such a bridge is present in the “probably approximately correct” (PAC) learning framework, which studies the effect of adding a polynomial-time computation constraint on this relationship among error rates, training data size, and other parameters of the learning algorithm (3). Recent advances in this line of research include various lower bounds that establish fundamental gaps in performance achievable in certain machine-learning problems (e.g., sparse regression and sparse principal components analysis) via polynomial-time and exponential-time algorithms (23). The core of the problem, however, involves time-data trade-offs that are far from the polynomial/exponential boundary. The large data sets that are increasingly the norm require algorithms whose time and space requirements are linear or sublinear in the problem size (number of data points or number of dimensions). Recent research focuses on methods such as subsampling, random projections, and algorithm weakening to achieve scalability while retaining statistical control (24, 25). The ultimate goal is to be able to supply time and space budgets to machine-learning systems in addition to accuracy requirements, with the system finding an operating point that allows such requirements to be realized.

Opportunities and challenges

Despite its practical and commercial successes, machine learning remains a young field with many underexplored research opportunities. Some of these opportunities can be seen by contrasting current machine-learning approaches to the types of learning we observe in naturally

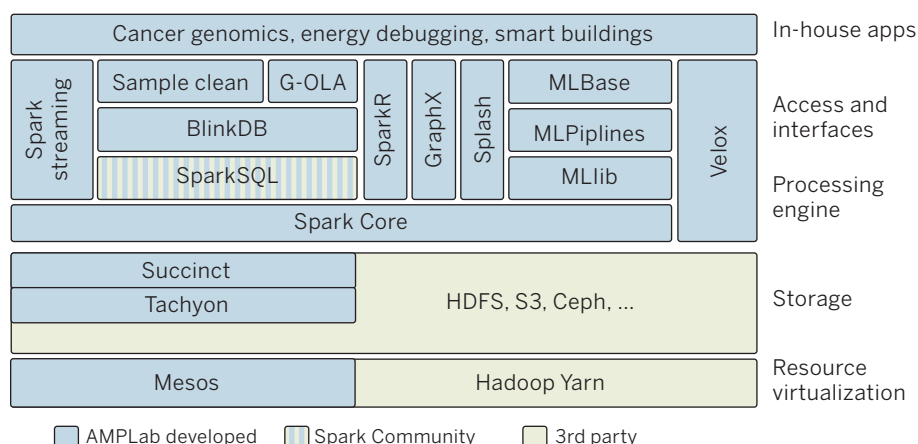


Fig. 5. Data analytics stack. Scalable machine-learning systems are layered architectures that are built on parallel and distributed computing platforms. The architecture depicted here—an open-source data analysis stack developed in the Algorithms, Machines and People (AMP) Laboratory at the University of California, Berkeley—includes layers that interface to underlying operating systems; layers that provide distributed storage, data management, and processing; and layers that provide core machine-learning competencies such as streaming, subsampling, pipelines, graph processing, and model serving.

occurring systems such as humans and other animals, organizations, economies, and biological evolution. For example, whereas most machine-learning algorithms are targeted to learn one specific function or data model from one single data source, humans clearly learn many different skills and types of knowledge, from years of diverse training experience, supervised and unsupervised, in a simple-to-more-difficult sequence (e.g., learning to crawl, then walk, then run). This has led some researchers to begin exploring the question of how to construct computer lifelong or never-ending learners that operate nonstop for years, learning thousands of interrelated skills or functions within an overall architecture that allows the system to improve its ability to learn one skill based on having learned another (26–28). Another aspect of the analogy to natural learning systems suggests the idea of team-based, mixed-initiative learning. For example, whereas current machine-learning systems typically operate in isolation to analyze the given data, people often work in teams to collect and analyze data (e.g., biologists have worked as teams to collect and analyze genomic data, bringing together diverse experiments and perspectives to make progress on this difficult problem). New machine-learning methods capable of working collaboratively with humans to jointly analyze complex data sets might bring together the abilities of machines to tease out subtle statistical regularities from massive data sets with the abilities of humans to draw on diverse background knowledge to generate plausible explanations and suggest new hypotheses. Many theoretical results in machine learning apply to all learning systems, whether they are computer algorithms, animals, organizations, or natural evolution. As the field progresses, we may see machine-learning theory and algorithms increasingly providing models for understanding learning in neural systems,

organizations, and biological evolution and see machine learning benefit from ongoing studies of these other types of learning systems.

As with any powerful technology, machine learning raises questions about which of its potential uses society should encourage and discourage. The push in recent years to collect new kinds of personal data, motivated by its economic value, leads to obvious privacy issues, as mentioned above. The increasing value of data also raises a second ethical issue: Who will have access to, and ownership of, online data, and who will reap its benefits? Currently, much data are collected by corporations for specific uses leading to improved profits, with little or no motive for data sharing. However, the potential benefits that society could realize, even from existing online data, would be considerable if those data were to be made available for public good.

To illustrate, consider one simple example of how society could benefit from data that is already online today by using this data to decrease the risk of global pandemic spread from infectious diseases. By combining location data from online sources (e.g., location data from cell phones, from credit-card transactions at retail outlets, and from security cameras in public places and private buildings) with online medical data (e.g., emergency room admissions), it would be feasible today to implement a simple system to telephone individuals immediately if a person they were in close contact with yesterday was just admitted to the emergency room with an infectious disease, alerting them to the symptoms they should watch for and precautions they should take. Here, there is clearly a tension and trade-off between personal privacy and public health, and society at large needs to make the decision on how to make this trade-off. The larger point of this example, however, is that, although the data are already online, we do not currently have the laws, customs, culture, or mechanisms to enable

society to benefit from them, if it wishes to do so. In fact, much of these data are privately held and owned, even though they are data about each of us. Considerations such as these suggest that machine learning is likely to be one of the most transformative technologies of the 21st century. Although it is impossible to predict the future, it appears essential that society begin now to consider how to maximize its benefits.

REFERENCES

1. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2011).
2. K. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, 2012).
3. L. Valiant, *Commun. ACM* **27**, 1134–1142 (1984).
4. V. Chandrasekaran, M. I. Jordan, *Proc. Natl. Acad. Sci. U.S.A.* **110**, E1181–E1190 (2013).
5. S. Decatur, O. Goldreich, D. Ron, *SIAM J. Comput.* **29**, 854–879 (2000).
6. S. Shalev-Shwartz, O. Shamir, E. Tromer, Using more data to speed up training time, *Proceedings of the Fifteenth Conference on Artificial Intelligence and Statistics*, Canary Islands, Spain, 21 to 23 April, 2012.
7. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, in *Foundations and Trends in Machine Learning* 3 (Now Publishers, Boston, 2011), pp. 1–122.
8. S. Sra, S. Nowozin, S. Wright, *Optimization for Machine Learning* (MIT Press, Cambridge, MA, 2011).
9. J. Schmidhuber, *Neural Netw.* **61**, 85–117 (2015).
10. Y. Bengio, in *Foundations and Trends in Machine Learning* 2 (Now Publishers, Boston, 2009), pp. 1–127.
11. A. Krizhevsky, I. Sutskever, G. Hinton, *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2015).
12. G. Hinton et al., *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
13. G. E. Hinton, R. R. Salakhutdinov, *Science* **313**, 504–507 (2006).
14. V. Mnih et al., *Nature* **518**, 529–533 (2015).
15. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
16. E. Yajlali, J. S. Ivy, Partially observable MDPs (POMDPs): Introduction and examples, *Encyclopedia of Operations Research and Management Science* (John Wiley, New York, 2011).
17. W. Schultz, P. Dayan, P. R. Montague, *Science* **275**, 1593–1599 (1997).
18. C. Dwork, F. McSherry, K. Nissim, A. Smith, in *Proceedings of the Third Theory of Cryptography Conference*, New York, 4 to 7 March 2006, pp. 265–284.
19. A. Blum, K. Ligett, A. Roth, *J. ACM* **20**, (2013).
20. J. Duchi, M. I. Jordan, J. Wainwright, *J. ACM* **61**, 1–57 (2014).
21. M.-F. Balcan, A. Blum, S. Fine, Y. Mansour, Distributed learning, communication complexity and privacy, *Proceedings of the 29th Conference on Computational Learning Theory*, Edinburgh, UK, 26 June to 1 July 2012.
22. Y. Zhang, J. Duchi, M. Jordan, M. Wainwright, in *Advances in Neural Information Processing Systems* 26, L. Bottou, C. Burges, Z. Ghahramani, M. Welling, Eds. (Curran Associates, Red Hook, NY, 2014), pp. 1–23.
23. Q. Berthet, P. Rigollet, *Ann. Stat.* **41**, 1780–1815 (2013).
24. A. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan, *J. R. Stat. Soc., B* **76**, 795–816 (2014).
25. M. Mahoney, *Found. Trends Machine Learn.* **3**, 123–224 (2011).
26. T. Mitchell et al., *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*, 25 to 30 January 2015, Austin, TX.
27. M. Taylor, P. Stone, *J. Mach. Learn. Res.* **10**, 1633–1685 (2009).
28. S. Thrun, L. Pratt, *Learning To Learn* (Kluwer Academic Press, Boston, 1998).
29. L. Wehbe et al., *PLOS ONE* **9**, e112575 (2014).
30. K. Xu et al., *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, Lille, France, 6 to 11 July 2015, pp. 2048–2057.
31. D. Blei, *Commun. ACM* **55**, 77–84 (2012).

10.1126/science.aaa8415

Advances in natural language processing

Julia Hirschberg^{1,*} and Christopher D. Manning^{2,3}

Natural language processing employs computational techniques for the purpose of learning, understanding, and producing human language content. Early computational approaches to language research focused on automating the analysis of the linguistic structure of language and developing basic technologies such as machine translation, speech recognition, and speech synthesis. Today's researchers refine and make use of such tools in real-world applications, creating spoken dialogue systems and speech-to-speech translation engines, mining social media for information about health or finance, and identifying sentiment and emotion toward products and services. We describe successes and challenges in this rapidly advancing area.

Over the past 20 years, computational linguistics has grown into both an exciting area of scientific research and a practical technology that is increasingly being incorporated into consumer products (for example, in applications such as Apple's Siri and Skype Translator). Four key factors enabled these developments: (i) a vast increase in computing power, (ii) the availability of very large amounts of linguistic data, (iii) the development of highly successful machine learning (ML) methods, and (iv) a much richer understanding of the structure of human language and its deployment in social contexts. In this Review, we describe some current application areas of interest in language research. These efforts illustrate computational approaches to big data, based on current cutting-edge methodologies that combine statistical analysis and ML with knowledge of language.

Computational linguistics, also known as natural language processing (NLP), is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content. Computational linguistic systems can have multiple purposes: The goal can be aiding human-human communication, such as in machine translation (MT); aiding human-machine communication, such as with conversational agents; or benefiting both humans and machines by analyzing and learning from the enormous quantity of human language content that is now available online.

During the first several decades of work in computational linguistics, scientists attempted to write down for computers the vocabularies and rules of human languages. This proved a difficult task, owing to the variability, ambiguity, and context-dependent interpretation of human languages. For instance, a star can be either an astronomical object or a person, and "star" can be a noun or a verb. In another example, two interpretations are possible for the headline "Teacher

strikes idle kids," depending on the noun, verb, and adjective assignments of the words in the sentence, as well as grammatical structure. Beginning in the 1980s, but more widely in the 1990s, NLP was transformed by researchers starting to build models over large quantities of empirical language data. Statistical or corpus ("body of words")-based NLP was one of the first notable successes of the use of big data, long before the power of ML was more generally recognized or the term "big data" even introduced.

A central finding of this statistical approach to NLP has been that simple methods using words, part-of-speech (POS) sequences (such as whether a word is a noun, verb, or preposition), or simple templates can often achieve notable results when trained on large quantities of data. Many text and sentiment classifiers are still based solely on the different sets of words ("bag of words") that documents contain, without regard to sentence and discourse structure or meaning. Achieving improvements over these simple baselines can be quite difficult. Nevertheless, the best-performing systems now use sophisticated ML approaches and a rich understanding of linguistic structure. High-performance tools that identify syntactic and semantic information as well as information about discourse context are now available. One example is Stanford CoreNLP (*1*), which provides a standard NLP preprocessing pipeline that includes POS tagging (with tags such as noun, verb, and preposition); identification of named entities, such as people, places, and organizations; parsing of sentences into their grammatical structures; and identifying co-references between noun phrase mentions (Fig. 1).

Historically, two developments enabled the initial transformation of NLP into a big data field. The first was the early availability to researchers of linguistic data in digital form, particularly through the Linguistic Data Consortium (LDC) (*2*), established in 1992. Today, large amounts of digital text can easily be downloaded from the Web. Available as linguistically annotated data are large speech and text corpora annotated with POS tags, syntactic parses, semantic labels, annotations of named entities (persons, places, organizations), dialogue acts (statement,

question, request), emotions and positive or negative sentiment, and discourse structure (topic or rhetorical structure). Second, performance improvements in NLP were spurred on by shared task competitions. Originally, these competitions were largely funded and organized by the U.S. Department of Defense, but they were later organized by the research community itself, such as the CoNLL Shared Tasks (*3*). These tasks were a precursor of modern ML predictive modeling and analytics competitions, such as on Kaggle (*4*), in which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

A major limitation of NLP today is the fact that most NLP resources and systems are available only for high-resource languages (HRLs), such as English, French, Spanish, German, and Chinese. In contrast, many low-resource languages (LRLs)—such as Bengali, Indonesian, Punjabi, Cebuano, and Swahili—spoken and written by millions of people have no such resources or systems available. A future challenge for the language community is how to develop resources and tools for hundreds or thousands of languages, not just a few.

Machine translation

Proficiency in languages was traditionally a hallmark of a learned person. Although the social standing of this human skill has declined in the modern age of science and machines, translation between human languages remains crucially important, and MT is perhaps the most substantial way in which computers could aid human-human communication. Moreover, the ability of computers to translate between human languages remains a consummate test of machine intelligence: Correct translation requires not only the ability to analyze and generate sentences in human languages but also a humanlike understanding of world knowledge and context, despite the ambiguities of languages. For example, the French word "bordel" straightforwardly means "brothel"; but if someone says "My room is un bordel," then a translating machine has to know enough to suspect that this person is probably not running a brothel in his or her room but rather is saying "My room is a complete mess."

Machine translation was one of the first non-numeric applications of computers and was studied intensively starting in the late 1950s. However, the hand-built grammar-based systems of early decades achieved very limited success. The field was transformed in the early 1990s when researchers at IBM acquired a large quantity of English and French sentences that were translations of each other (known as parallel text), produced as the proceedings of the bilingual Canadian Parliament. These data allowed them to collect statistics of word translations and word sequences and to build a probabilistic model of MT (*5*).

Following a quiet period in the late 1990s, the new millennium brought the potent combination of ample online text, including considerable quantities of parallel text, much more abundant and inexpensive computing, and a new idea for building statistical phrase-based MT systems

¹Department of Computer Science, Columbia University, New York, NY 10027, USA. ²Department of Linguistics, Stanford University, Stanford, CA 94305-2150, USA. ³Department of Computer Science, Stanford University, Stanford, CA 94305-9020, USA.

*Corresponding author. E-mail: julia@cs.columbia.edu

(6). Rather than translating word by word, the key advance is to notice that small word groups often have distinctive translations. The Japanese 水色 “mizu iro” is literally the sequence of two words (“water color”), but this is not the correct meaning (nor does it mean a type of painting); rather, it indicates a light, sky-blue color. Such phrase-based MT was used by Franz Och in the development of Google Translate.

This technology enabled the services we have today, which allow free and instant translation between many language pairs, but it still produces translations that are only just serviceable for determining the gist of a passage. However, very promising work continues to push MT forward. Much subsequent research has aimed to better exploit the structure of human language sentences (i.e., their syntax) in translation systems (7, 8), and researchers are actively building deeper meaning representations of language (9) to enable a new level of semantic MT.

Finally, just in the past year, we have seen the development of an extremely promising approach to MT through the use of deep-learning-based sequence models. The central idea of deep learning is that if we can train a model with several representational levels to optimize a final objective, such as translation quality, then the model can itself learn intermediate representations that are useful for the task at hand. This idea has been explored particularly for neural network models in which information is stored in real-valued vectors, with the mapping between vectors consisting of a matrix multiplication followed by a nonlinearity, such as a sigmoid function that maps the output values of the matrix multiplication onto $[-1, 1]$. Building large models

of this form is much more practical with the massive parallel computation that is now economically available via graphics processing units. For translation, research has focused on a particular version of recurrent neural networks, with enhanced “long short-term memory” computational units that can better maintain contextual information from early until late in a sentence (10, 11) (Fig. 2). The distributed representations of neural networks are often very effective for capturing subtle semantic similarities, and neural MT systems have already produced some state-of-the-art results (12, 13).

A still-underexplored area in MT is getting machines to have more of a sense of discourse, so that a sequence of sentences translates naturally—although work in the area has begun (14). Finally, MT is not necessarily a task for machines to do alone. Rather it can be reconceptualized as an opportunity for computer-supported cooperative work that also exploits human skills (15). In such a system, machine intelligence is aimed at human-computer interface capabilities of giving effective suggestions and reacting productively to human input, rather than wholly replacing the skills and knowledge of a human translator.

Spoken dialogue systems and conversational agents

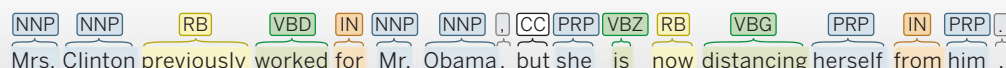
Dialogue has been a popular topic in NLP research since the 1980s. However, early work on text-based dialogue has now expanded to include spoken dialogue on mobile devices (e.g., Apple’s Siri, Amtrak’s Julie, Google Now, and Microsoft’s Cortana) for information access and task-based apps. Spoken dialogue systems (SDSs) also allow robots to help people with simple manual tasks [e.g., Manuela Veloso’s CoBots (16)] or provide

therapy for less-abled persons [e.g., Maja Mataric’s socially assistive robots (17)]. They also enable avatars to tutor people in interview or negotiation strategies or to help with health care decisions (18, 19).

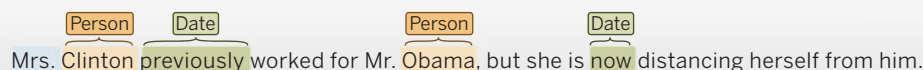
The creation of SDSs, whether between humans or between humans and artificial agents, requires tools for automatic speech recognition (ASR), to identify what a human says; dialogue management (DM), to determine what that human wants; actions to obtain the information or perform the activity requested; and text-to-speech (TTS) synthesis, to convey that information back to the human in spoken form. (Fig. 3). In addition, SDSs need to be ready to interact with users when an error in speech recognition occurs; to decide what words might be incorrectly recognized; and to determine what the user actually said, either automatically or via dialogue with the user. In speech-to-speech translation systems, MT components are also needed to facilitate dialogue between speakers of different languages and the system, to identify potential mistranslations before they occur, and to clarify these with the speaker.

Practical SDSs have been enabled by breakthroughs in speech recognition accuracy, mainly coming from replacing traditional acoustic feature-modeling pipelines with deep-learning models that map sound signals to sequences of human language sounds and words (20). Although SDSs now work fairly well in limited domains, where the topics of the interaction are known in advance and where the words people are likely to use can be predetermined, they are not yet very successful in open-domain interaction, where users may talk about anything at all. Chatbots following in the tradition of ELIZA (21) handle open-domain interaction by cleverly repeating variations of the human input;

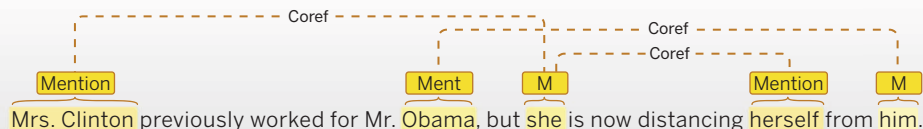
Part of speech:



Named entity recognition:



Co-reference:



Basic dependencies:

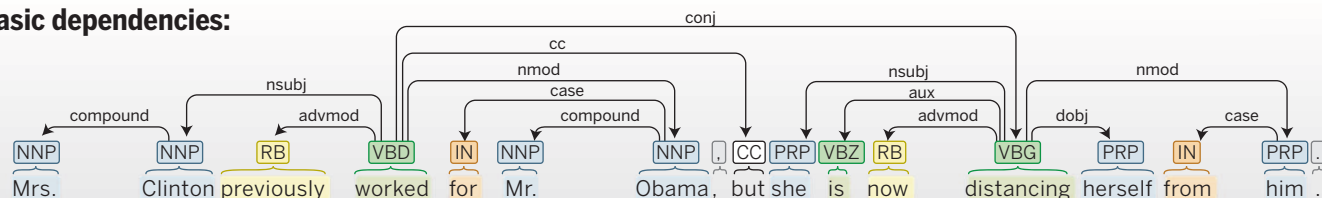


Fig. 1. Many language technology tools start by doing linguistic structure analysis. Here we show output from Stanford CoreNLP. As shown from top to bottom, this tool determines the parts of speech of each word, tags various words or phrases as semantic named entities of various sorts, determines which entity mentions co-refer to the same person or organization, and then works out the syntactic structure of each sentence, using a dependency grammar analysis.

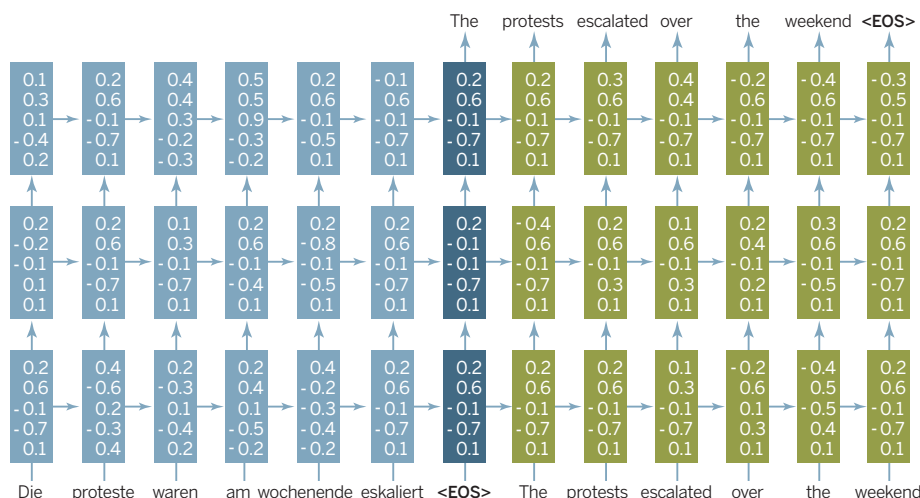


Fig. 2. A deep, recurrent neural MT system (10). Initially trained on parallel sentences that translate each other, the model learns a representation of each word as a real-valued vector and internal parameter matrices so as to optimize translation quality. The trained network can then translate new sentences. Each arrow represents a computation unit of a matrix multiplication followed by a nonlinear transformation; the small vectors shown in the illustration might really be 1000-dimensional. The recurrent network first encodes the meaning of the source sentence (left side, blue). It maintains an internal state representing a partial sentence, which is updated after each new word is read (horizontal arrows). Having the upper network layers [mapped to by additional (upper vertical) computation arrows] makes this a deep recurrent network. Adding depth improves the ability of the model to learn, generalize, and remember. Once the end of the sentence (denoted by <EOS>) is reached (middle, dark blue), the network additionally starts to produce a word of translated output at each step from its internal state (using a multiclass logistic regression-style model). During translation generation (right side, green), the last generated word is fed in as the input at each step. From the stored hidden state and this input, the model calculates the next word of the translation. This process repeats until <EOS> is generated.

this approach is also being attempted in spoken-chat systems designed to provide a sense of companionship for target audiences such as the elderly or individuals with dementia (22). In spoken dialogue, information about the speaker's mental state inferred from multimodal information can be used to supplement the system's knowledge of what the user is saying.

There are many challenges in building SDSs, in addition to the primary challenge of improving the accuracy of the basic ASR, DM, and TTS building blocks and extending their use into less-restricted domains. These include basic problems of recognizing and producing normal human conversational behaviors, such as turn-taking and coordination. Humans interpret subtle cues in speakers' voices and facial and body gestures (where available) to determine when the speaker is ready to give up the turn versus simply pausing. These cues, such as a filled pause (e.g., "um" or "uh"), are also used to establish when some feedback from the listener is desirable, to indicate that he or she is listening or working on a request, as well as to provide "grounding" (i.e., information about the current state of the conversation). Non-humanlike latency often makes SDS burdensome, as users must wait seconds to receive a system response. To address this, researchers are exploring incremental processing of ASR, MT, and TTS modules, so that systems can respond more quickly to users by beginning these recognition, translation, and generation processes while the user is still speaking. Hu-

mans can also disambiguate words such as "yeah" and "okay," which may have diverse meanings—including agreement, topic shift, and even disagreement—when spoken in different ways. In successful and cooperative conversations, humans also tend to entrain to their conversational partners, becoming more similar to each other in pronunciation, word choice, acoustic and prosodic features, facial expressions, and gestures. This tendency has long been used to subtly induce SDS users to employ terms that the system can more easily recognize. Currently, researchers are beginning to believe that systems (particularly embodied agents) should entrain to their users in these different modalities, and some experimental results have shown that users prefer such systems (23) and even think they are more intelligent (24). Open issues for DM have long been the determination of how to architect the appropriate dialogue flow for particular applications, where existing experimental data may be sparse and some aspects of the dialogue state may not yet have been observed or even be observable from the data. Currently, the most widely used approach is the POMDP (partially observable Markov decision process), which attempts to identify an optimal system policy by maintaining a probability distribution over possible SDS states and updating this distribution as the system observes additional dialogue behavior (25). This approach may make use of the identification of dialogue acts, such as whether the user input represents a question, statement, or indication of agreement, for example.

Machine reading

The printed word has great power to enlighten. Machine reading is the idea that machines could become intelligent, and could usefully integrate and summarize information for humans, by reading and understanding the vast quantities of text that are available.

In the early decades of artificial intelligence, many researchers focused on the approach of trying to enable intelligent machines by manually building large structured knowledge bases in a formal logical language and developing automated reasoning methods for deriving further facts from this knowledge. However, with the emergence of the modern online world, what we mainly have instead is huge repositories of online information coded in human languages. One place where this is true is in the scientific literature, where findings are still reported almost entirely in human language text (with accompanying tables and diagrams). However, it is equally true for more general knowledge, where we now have huge repositories of information such as Wikipedia (26). The quantity of scientific literature is growing rapidly: For example, the size of the U.S. National Library of Medicine's Medline index has grown exponentially (27). At such a scale, scientists are unable to keep up with the literature, even in their narrow domains of expertise. Thus, there is an increased need for machine reading for the purposes of comprehending and summarizing the literature, as well as extracting facts and hypotheses from this material.

An initial goal is to extract basic facts, most commonly a relation between two entities, such as "child of" (for instance, Bill Clinton, Chelsea Clinton). This is referred to as relation extraction. For particular domain-specific relations, many such systems have been successfully built. One technique is to use handwritten patterns that match the linguistic expression of relations (e.g., <PERSON>'s daughter, <PERSON>). Better results can be obtained through the use of

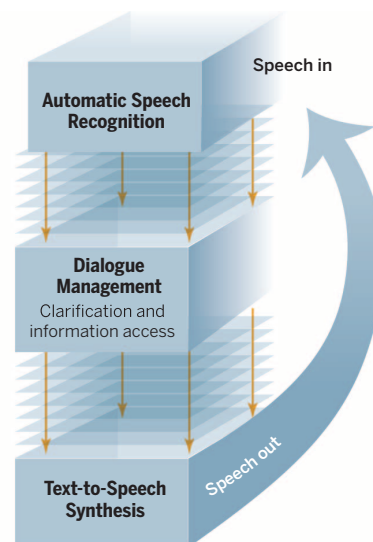


Fig. 3. A spoken dialogue system. The three main components are represented by rectangles; arrows denote the flow of information.

ML. A structured prediction classifier proposes instances of such relations based on extracted features from the sequence of words and grammatical structure of a sentence (28, 29). Such systems are the mainstay of literature fact-extraction tools in fields such as biomedicine (30, 31).

In many scientific fields, there have been major efforts to build databases of structured information based on the textual scientific record, such as the Gene Ontology database (32) in biomedicine or the PaleoBiology Database for fossil records (33). This has generally been done manually, via concerted work by trained professionals. Using artificial intelligence software to extract these databases, as well as to perform subsequent reasoning and hypothesis generation, has become a major research goal. One subfield where these questions have been actively pursued is pharmacogenomics (34). For example, Percha *et al.* (35) trained a model of drug-drug interactions based on drug-gene interactions extracted from the literature and were able to use it to predict novel drug-drug interactions.

If a partial knowledge base—for instance, Freebase (36), dbpedia (37), Wikidata (38) (related to Wikipedia), or the Gene Ontology database (32)—has already been extracted from biomedical research articles, then there is an opportunity to automatically align known facts from the knowledge base with putative expressions of those facts in text. The type labels from this mapping can then be used as if they were supervised data for ML information-extraction systems (Fig. 4). This is referred to as distantly supervised relation extraction. Early systems aligned entity mentions and then made the naïve assumption that sentences containing a pair of entities expressed every known relation between the two entities in the database (39). More recent systems have used increasingly sophisticated probabilistic inference to discern which textual clauses map to which facts in the knowledge base, or to something else entirely (40, 41). A dramatic recent application of this approach has been the DeepDive system (42), which aims to automate the construction of such systems by providing efficient large-scale learning and inference so a user can simply focus on good features for their domain. PaleoDeepDive, its application to the fossil record, has recently been shown to do a better job at fact extraction from journal articles than the scientist volunteers who maintain the PaleoBiology Database (43).

The relation-extraction task is made general, if less semantically precise, by aiming to extract all relations from any piece of text, a task normally referred to as open information extraction (Open IE) (44). Early work emphasized the development of simple but highly scalable fact-extraction techniques that do not require any kind of hand-labeled data (45). With ever-growing computational power, a second generation of work increasingly emphasized careful use of linguistic structure, which can reliably be extracted with the use of detailed NLP (46).

Currently, a number of avenues are being explored to further extend the ability of computers to build and use knowledge bases starting from textual information. An exciting unification

is the proposal for universal schemas (47), which allow simultaneous inference and knowledge-base completion over both the open set of textual relations (such as “born in”) found in Open IE and the more exact schema of databases (such as `per:city_of_birth`). Even with all of our text-extraction techniques, any knowledge base will only be partial and incomplete; some recent work explores how it can be probabilistically completed to deliver a form of common-sense reasoning (48). Finally, we hope to move beyond simply extracting relations, events, and facts to be able to understand the relations between events (such as causation) and complex multistep procedures and processes. In (49), Berant *et al.* explore how this can be done for understanding the steps in biological processes, showing that extracting explicit process structures can improve the accuracy of question answering. The flip side of machine reading is to provide question-answering systems, by which humans can get answers from constructed knowledge bases. There has recently been dramatic progress in building such systems by learning semantic parsers (50).

Mining social media

The development of social media has revolutionized the amount and types of information available today to NLP researchers. Data available from sources such as Twitter, Facebook, YouTube, blogs, and discussion forums make it possible to examine relations between demographic information, language use, and social interaction (51). Researchers use Web-scraping techniques, often via application program interfaces provided by websites, to download previously unimaginable amounts and categories of data. Using statistical and ML techniques, they learn to identify demographic information (such as age and gender) from language, track trending topics and popular sentiment, identify opinions and beliefs about products and politicians, predict disease spreading (for instance, with Google Flu Trends: www.google.org/flutrends/) from symptoms mentioned in tweets or food-related illnesses (52), recognize deception in fake reviews (53), and identify social networks of people who interact together online.

In this era of big data, the availability of social media has revolutionized the ways advertisers, journalists, businesses, politicians, and medical experts acquire their data and the ways in which those data can be put to practical use. Product reviews can be mined to predict pricing trends and assess advertising campaigns. Political forums can be searched to predict candidate appeal and performance in elections. Social networks can be examined to find indicators of power and influence among different groups. Medical forums can be studied to discover common questions and misconceptions about sufferers from particular medical conditions so that website information can be improved.

Social media also provide very large and rich sources of conversational data in Web forums that can provide “found” data for the study of language phenomena such as code-switching (mixed

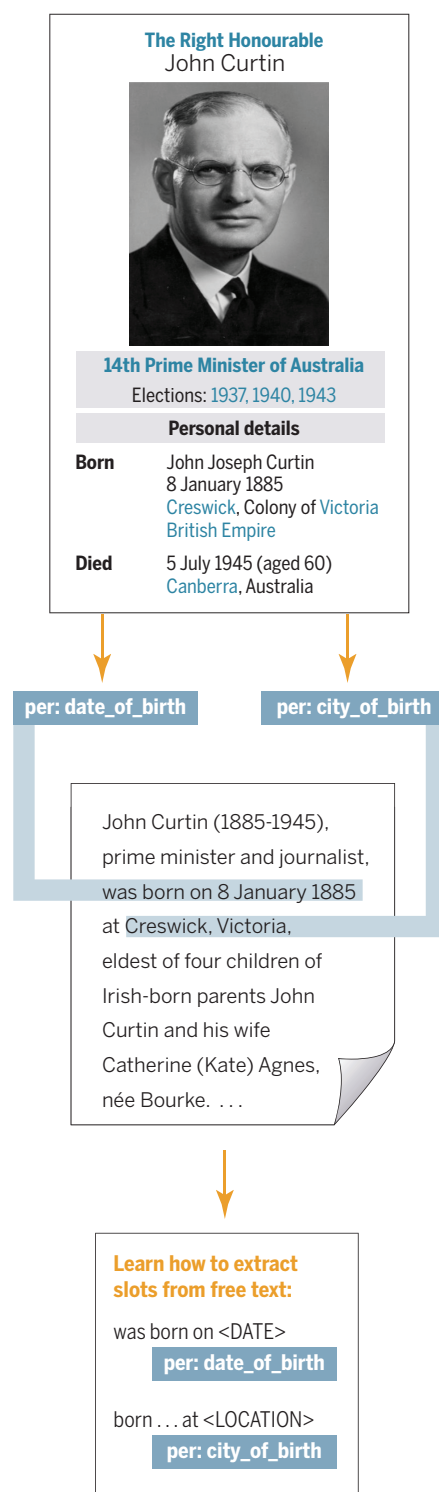


Fig. 4. Distantly supervised learning. In this approach, facts in a structured knowledge representation are projected onto pieces of text that mention the people, places, dates, etc., that appear in knowledge-base entries. This projection is noisy, but when done over large quantities of text, it provides enough signal to successfully learn good classifiers for extracting relations from text. [Photo source: National Library of Australia, <http://nla.gov.au/nla.pic-an12267621>]

language in bilingual speech), hedging behavior (words and phrases indicating lack of commitment to a proposition such as “sort of”), and hate speech or bullying behavior. Social media exist in a wide variety of languages, including both HRLs and LRLs. These data can be invaluable for enriching ASR language models and developing TTS synthesizers without the need to create costly special-purpose corpora. In turn, these technologies can be useful in producing SDSs in LRL areas. Such systems can provide millions of people with the ability to obtain information over their cell phones (which are ubiquitous, even among populations with low literacy rates or whose languages or dialects have no standard written form), similar to the residents of HRL countries. The development of tools for LRLs from found LRL data, by adapting HRL tools, is another important way to use found text data. A particular application of data mining in LRLs is the mining of data collected from Twitter or blogs to provide valuable information for disaster relief organizations, identifying the most serious problems, where they occur, and who is experiencing them.

There are also some drawbacks to social media data mining. There is an increasing concern for privacy issues, particularly for an individual’s control over their own data versus researchers’ desire to mine it. Sites such as Twitter severely limit a researcher’s ability to download data, which impedes speedy corpus collection. There is also a major issue with discovering “ground truth” in online postings, because there is no clear way of validating an individual’s demographic information; the validity of posts concerning events; and most reviews of hotels, restaurants, and products. Aggregating information from multiple sources at similar times can address some validity issues, and sites do attempt to identify spurious reviews, but this issue remains perhaps the most difficult one for those working with social media data.

Analysis and generation of speaker state

Speaker states (54), also termed “private states” (55), include opinions, speculations, beliefs, emotions, and any other evaluative views that are personally held by the speaker or writer of a language. Much of the work in NLP has focused on sentiment analysis (identification of positive or negative orientation of textual language) and identification of belief states (committed belief,

uncommitted belief, or neutrality of a sentence) on the basis of lexical and syntactic information. Both sentiment and belief constitute attitudes toward events and propositions, although sentiment can also concern attitudes toward objects such as people, organizations, and abstract concepts. Detection of sentiment and emotion in text requires lexical and sentence-level information. Sentiment can be signaled by words conveying positive or negative orientation: For example, “sad,” “worried,” “difficult,” and “weak” are all words with negative orientation, whereas “comfortable,” “important,” “successful,” and “interesting” convey a positive sentiment. Online sentiment dictionaries, such as Whissel’s Dictionary of Affect (56), and systems created from subject-ranked terms, such as Tausczik and Pennebaker’s LIWC (Linguistic Inquiry and Word Count) (57), can be used to assess positive and negative sentiment in a text. More sophisticated approaches to sentiment analysis also seek to identify the holder (source) as well as the object of the sentiment: for instance, who is positive about what person, country, activity, or concept (55).

The speech community has also studied positive and negative attitudes by focusing more generally on the identification of positive and negative emotions, primarily using acoustic and prosodic information. However, more work is currently being done to identify particular emotions, such as Ekman’s classic six basic emotions (anger, disgust, fear, happiness, sadness, surprise), which may be reactions to events, propositions, or objects. There has also been considerable research using features that have proven important in recognizing classic emotions to identify other speaker states (such as deception), medical conditions (such as autism and Parkinson’s disease), speaker characteristics (such as age, gender, likeability, pathology, and personality), and speaker conditions (such as cognitive load, drunkenness, sleepiness, interest, and trust). Corpora collected for such studies have been used in the Interspeech Paralinguistic Challenges, which have been conducted since 2009. Emotion generation has proven a more difficult challenge for TTS synthesis. Although there are some systems (e.g., MARY) that attempt to generate emotions such as depression, aggression, or cheerfulness (58), the best synthesized emotion still comes from corpora recorded for particular emotions by voice talent imitating those emotions.

Sentiment classification is widely used in opinion identification (positive or negative views of people, institutions, or ideas) in many languages and genres. Particular applications abound, such as identifying positive and negative movie or product reviews (59, 60) and predicting votes from congressional records (61) or Supreme Court decisions from court proceedings. Figure 5 illustrates a typical restaurant review, annotated for positive, negative, and neutral sentiment, as well as basic emotions.

Mining social media for sentiment or classic emotions has been a particularly popular topic for the purposes of assessing the “public mood” from Twitter, predicting stock market trends, or simply evaluating a community’s mental state (62). Social media such as Twitter, blog posts, and forums also provide researchers with very large amounts of data to use in assessing the role of sentiment and emotion in identifying other linguistic or social phenomena [e.g., sarcasm (63), power relationships, and social influence (64)], as well as mental health issues [e.g., depression (65)].

Conclusion and outlook

Many times during the past 50 years, enthusiastic researchers have had high hopes that the language-understanding ability of robots in science fiction movies was just around the corner. However, in reality, speech and language understanding did not work well enough at that time to power mainstream applications. The situation has been changing dramatically over the past five years. Huge improvements in speech recognition have made talking to your phone a commonplace activity, especially for young people. Web search engines are increasingly successful in understanding complex queries, and MT can at least yield the gist of material in another language, even if it cannot yet produce human-quality translations. Computer systems trade stocks and futures automatically, based on the sentiment of reports about companies. As a result, there is now great commercial interest in the deployment of human language technology, especially because natural language represents such a natural interface when interacting with mobile phones. In the short term, we feel confident that more data and computation, in addition to recent advances in ML and deep learning, will lead to further substantial progress in NLP. However, the truly difficult problems of semantics, context, and

Breakfast on Broadway is a new place focusing on, you guessed it, breakfast/brunch. Went there last Sunday around 1. **The food was not bad but the service was pretty terrible. We had to wait 15 minutes just to get menus and another 30 to get something to eat. And there were only a few tables occupied! If you don't mind the wait though, the price is right. I'll probably give it another try.** Maybe they need time to get their act together.

Breakfast on Broadway is a new place focusing on, you guessed it, breakfast/brunch. Went there last Sunday around 1. The food was not bad but **[Anger: the service was pretty terrible]. [Disgust: We had to wait 15 minutes just to get menus and another 30 to get something to eat. And there were only a few tables occupied!]** If you don't mind the wait though, the price is right. I'll probably give it another try. **[Uncertainty: Maybe they need time to get their act together.]**

Fig. 5. Manually annotated text analysis on a sample restaurant review. Sentiment analysis is shown on the left (blue, positive sentiments; red, negative; gray, neutral). In the emotion analysis on the right, emotions are shown in bold type and delineated by square brackets. Note in particular the importance of going beyond simple keyword analysis; for example, “not” has scope over “bad,” which might mislead simple systems. Also, the presence of “hedge” words and phrases, which muddle the intended meaning (e.g., “pretty,” which has a positive connotation, modifying the negative word “terrible”), somewhat decreases the negative score of the next clause.

knowledge will probably require new discoveries in linguistics and inference. From this perspective, it is worth noting that the development of probabilistic approaches to language is not simply about solving engineering problems: Probabilistic models of language have also been reflected back into linguistic science, where researchers are finding important new applications in describing phonology (66), understanding human language processing (67), and modeling linguistic semantics and pragmatics (68). Many areas of linguistics are themselves becoming more empirical and more quantitative in their approaches.

REFERENCES AND NOTES

- C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations* (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 55–60.
- Linguistic Data Consortium, www ldc.upenn.edu/.
- CoNLL Shared Tasks, <http://ifarnl.nl/signl/conll/>.
- Kaggle, www.kaggle.com.
- P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer, *Comput. Linguist.* **19**, 263–311 (1993).
- P. Koehn, F. J. Och, D. Marcu, "Statistical phrase-based translation," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, 2003), pp. 48–54.
- D. Chiang, "A hierarchical phrase-based model for statistical machine translation," *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, 2005), pp. 263–270.
- M. Galley, M. Hopkins, K. Knight, D. Marcu, "What's in a translation rule?" in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2004)* (Association for Computational Linguistics, Stroudsburg, PA, 2004).
- B. Jones, J. Andreas, D. Bauer, K. M. Hermann, K. Knight, "Semantics-based machine translation with hyperedge replacement grammars," in *Proceedings of COLING 2012* (Technical Papers, The COLING 2012 Organizing Committee, Mumbai, India, 2012), pp. 1359–1376.
- I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, Eds. (Curran Associates, Red Hook, NY, 2014), pp. 3104–3112.
- D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," <http://arxiv.org/abs/1409.0473> (2015).
- M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, W. Zaremba, "Addressing the rare word problem in neural machine translation," <http://arxiv.org/abs/1410.8206> (2015).
- S. Jean, K. Cho, R. Memisevic, Y. Bengio, "On using very large target vocabulary for neural machine translation," <http://arxiv.org/abs/1412.2007> (2015).
- S. Stymer, C. Hardmeier, J. Tiedemann, J. Nivre, "Feature weight optimization for discourse-level SMT," in *Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT)* (Association for Computational Linguistics, Stroudsburg, PA, 2013), pp. 60–69.
- S. Green, J. Chuang, J. Heer, C. D. Manning, "Predictive translation memory: A mixed-initiative system for human language translation," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, Honolulu, HI, 5 to 8 October 2014 (Association for Computing Machinery, New York, 2014), pp. 177–187.
- S. Rosenthal, J. Biswas, M. Veloso, "An effective personal mobile robot agent through symbiotic human-robot interaction," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, Toronto, Canada, 10 to 14 May 2010 (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2010), pp. 915–922.
- J. Fasola, M. J. Mataric, *J. Human-Robot Interact.* **2**, 3–32 (2013).
- M. Core, H. C. Lane, D. Traum, "Intelligent tutoring support for learners interacting with virtual humans," in *Design Recommendations for Intelligent Tutoring Systems* (U.S. Army Research Laboratory, Orlando, FL, 2014), vol. 2, pp. 249–257.
- D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratos, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, L.-P. Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, Paris, France, 5 to 9 May 2014 (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2014), pp. 1061–1068; <http://aamas2014.lip6.fr/proceedings/aamas/pl061.pdf>.
- G. Hinton *et al.*, *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
- J. Weizenbaum, *Commun. ACM* **9**, 36–45 (1966).
- Y. Nonaka, Y. Sakai, K. Yasuda, Y. Nakano, "Towards assessing the communication responsiveness of people with dementia," in *12th International Conference on Intelligent Virtual Agents (IVA'12)* (Springer, Berlin, 2012), pp. 496–498.
- C. Nass, Y. Moon, B. J. Fogg, B. Reeves, D. C. Dryer, *Int. J. Hum. Comput. Stud.* **43**, 223–239 (1995).
- H. Giles, A. Mulac, J. J. Bradac, P. Johnson, "Speech accommodation theory: The next decade and beyond," in *Communication Yearbook* (Sage, Newbury Park, CA, 1987), vol. 10, pp. 13–48.
- S. Young, M. Gasic, B. Thomson, J. Williams, *Proc. IEEE* **101**, 1160–1179 (2013).
- Wikipedia, www.wikipedia.org/.
- L. Hunter, K. B. Cohen, *Mol. Cell* **21**, 589–594 (2006).
- A. Culotta, J. Sorensen, "Dependency tree kernels for relation extraction," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, 2004), pp. 423–429.
- K. Fundel, R. Küffner, R. Zimmer, *Bioinformatics* **23**, 365–371 (2007).
- J. Björne *et al.*, *Comput. Intell.* **27**, 541–557 (2011).
- S. Van Landeghem *et al.*, *PLOS ONE* **8**, e55814 (2013).
- M. Ashburner *et al.* The Gene Ontology Consortium, *Nat. Genet.* **25**, 25–29 (2000).
- PaleoBiology Database, <https://paleobiodb.org/>.
- A. Coulet, K. B. Cohen, R. B. Altman, *J. Biomed. Inform.* **45**, 825–826 (2012).
- B. Percha, Y. Garten, R. B. Altman, *Pac. Symp. Biocomput.* **2012**, 410–421 (2012).
- Freebase, www.freebase.com/.
- dbpedia, <http://dbpedia.org/>.
- Wikidata, www.wikidata.org/.
- M. Mintz, S. Bills, R. Snow, D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Association for Computational Linguistics, Stroudsburg, PA, 2009), vol. 2, pp. 1003–1011.
- M. Surdeanu, J. Tibshirani, R. Nallapati, C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju Island, South Korea, 12 to 14 July 2012 (Association for Computational Linguistics, Stroudsburg, PA, 2012), pp. 455–465.
- B. Min, R. Grishman, L. Wan, C. Wang, D. Gondek, "Distant supervision for relation extraction with an incomplete knowledge base," in *Proceedings of NAACL-HLT 2013*, Atlanta, GA, 9 to 14 June 2013 (Association for Computational Linguistics, Stroudsburg, PA, 2013), pp. 777–782.
- DeepDive, <http://deeplive.stanford.edu/>.
- S. E. Peters, C. Zhang, M. Lin, C. Ré, *PLOS ONE* **9**, e113523 (2014).
- E. Etzioni, M. Banko, M. J. Cafarella, "Machine reading," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, Boston, MA, 16 to 20 July 2006 (AAAI Press, Menlo Park, CA, 2006), vol. 2, pp. 1517–1519.
- M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, "Open information extraction from the web," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)* (Morgan Kaufmann, San Francisco, 2007), pp. 2670–2676.
- O. Etzioni, A. Fader, J. Christensen, S. Soderland, Mausam, "Open information extraction: The second generation," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 16 to 22 July 2011 (AAAI Press, Menlo Park, CA, 2011), pp. 3–10.
- S. Riedel, L. Yao, A. McCallum, B. M. Marlin, "Relation extraction with matrix factorization and universal schemas," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (HLT NAACL 2013)* (Stroudsburg, PA, 2013), pp. 74–84.
- G. Angeli, C. D. Manning, "NaturalL: Natural logic inference for common sense reasoning," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 25 to 29 October 2014 (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 534–545.
- J. Berant, V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, C. D. Manning, "Modeling biological processes for reading comprehension," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 25 to 29 October 2014 (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 1499–1510.
- A. Fader, L. Zettlemoyer, O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)* (Association for Computing Machinery, New York, 2014), pp. 1156–1165.
- M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More* (O'Reilly Media, Sebastopol, CA, ed. 2, 2013).
- N. Elhadad, L. Gravano, D. Hsu, S. Balter, V. Reddy, H. Waechter, "Information extraction from social media for public health," in *KDD at Bloomberg Workshop, Data Frameworks Track (KDD 2014)* (Association for Computing Machinery, New York, 2014).
- M. Ott, C. Cardie, J. T. Hancock, "Estimating the prevalence of deception in online review communities," in *Proceedings of the 21st International Conference on World Wide Web Conference*, Lyon, France, 16 to 20 April 2012 (Association for Computing Machinery, New York, 2012), pp. 201–210.
- J. Liscombe, thesis, Columbia University (2007).
- J. Wiebe, T. Wilson, C. Cardie, *Lang. Resour. Eval.* **39**, 165–210 (2005).
- C. Whissell, "The dictionary of affect in language," in *Emotion: Theory, Research and Experience*, R. Plutchik, H. Kellerman, Eds. (Academic Press, London, 1989).
- Y. R. Tausczik, J. W. Pennebaker, *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
- O. Türk, M. Schröder, *IEEE Trans. Audio Speech Lang. Proc.* **18**, 965–973 (2010).
- B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, July 2002 (Association for Computational Linguistics, Stroudsburg, PA, 2002), vol. 10, pp. 79–86.
- H. Wang, M. Ester, "A sentiment-aligned topic model for product aspect rating prediction," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 25 to 29 October 2014 (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 1192–1202.
- M. Thomas, Bo Pang, L. Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 22 to 23 July 2006 (Association for Computational Linguistics, Stroudsburg, PA, 2006), pp. 327–335.
- J. Bollen, H. Mao, A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 17 to 21 July 2011 (AAAI Press, Menlo Park, 2011), pp. 450–453.
- R. Gonzalez-Ibanez, S. Muresan, N. Wacholder, "Identifying sarcasm in Twitter: A closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, 19 to 24 June 2011 (Association for Computational Linguistics, Stroudsburg, PA, 2011), pp. 581–586.
- O. Biran, S. Rosenthal, J. Andreas, K. McKeown, O. Rambow, "Detecting influencers in written online conversations," in *Proceedings of the 2012 Workshop on Language in Social Media*, Montreal, Canada, 7 June 2012 (Association for Computational Linguistics, Stroudsburg, PA, 2012), pp. 37–45.
- L.-C. Yu, C.-Y. Ho, "Identifying emotion labels from psychiatric social texts using independent component analysis," in *Proceedings of COLING 2014* (Technical Papers, Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 837–847.
- B. Hayes, Z. Londe, *Phonology* **23**, 59–104 (2006).
- R. Levy, *Cognition* **106**, 1126–1177 (2008).
- N. D. Goodman, D. Lassiter, "Probabilistic semantics and pragmatics: Uncertainty in language and thought," in *Handbook of Contemporary Semantics*, C. Fox, S. Lappin, Eds. (Blackwell, Hoboken, NJ, ed. 2, 2015).

ACKNOWLEDGMENTS

C.D.M. holds equity in Google and serves in an advising role to Idbion, Liit, Lex Machina, and Xseed.

10.1126/science.aaa8685

Economic reasoning and artificial intelligence

David C. Parkes^{1*} and Michael P. Wellman^{2*}

The field of artificial intelligence (AI) strives to build rational agents capable of perceiving the world around them and taking actions to advance specified goals. Put another way, AI researchers aim to construct a synthetic *homo economicus*, the mythical perfectly rational agent of neoclassical economics. We review progress toward creating this new species of machine, *machina economicus*, and discuss some challenges in designing AIs that can reason effectively in economic contexts. Supposing that AI succeeds in this quest, or at least comes close enough that it is useful to think about AIs in rationalistic terms, we ask how to design the rules of interaction in multi-agent systems that come to represent an economy of AIs. Theories of normative design from economics may prove more relevant for artificial agents than human agents, with AIs that better respect idealized assumptions of rationality than people, interacting through novel rules and incentive systems quite distinct from those tailored for people.

Economics models the behavior of people, firms, and other decision-makers as a means to understand how these decisions shape the pattern of activities that produce value and ultimately satisfy (or fail to satisfy) human needs and desires. In this enterprise, the field classically starts from an assumption that actors behave rationally—that is, their decisions are the best possible given their available actions, their preferences, and their beliefs about the outcomes of these actions. Economics is drawn to rational decision models because they directly connect choices and values in a mathematically precise manner. Critics argue that the field studies a mythical species, *homo economicus* (“economic man”) and produces theories with limited applicability to how real humans behave. Defenders acknowledge that rationality is an idealization but counter that the abstraction supports powerful analysis, which is often quite predictive of people’s behavior (as individuals or in aggregate). Even if not perfectly accurate representations, rational models also allow preferences to be estimated from observed actions and build understanding that can usefully inform policy.

Artificial intelligence (AI) research is likewise drawn to rationality concepts, because they provide an ideal for the computational artifacts it seeks to create. Core to the modern conception of AI is the idea of designing agents: entities that perceive the world and act in it (1). The quality of an AI design is judged by how well the agent’s actions advance specified goals, conditioned on the perceptions observed. This coherence among perceptions, actions, and goals is the essence of rationality. If we represent goals in terms of preference over outcomes, and conceive perception and action within the framework of decision-

making under uncertainty, then the AI agent’s situation aligns squarely with the standard economic paradigm of rational choice. Thus, the AI designer’s task is to build rational agents, or agents that best approximate rationality given the limits of their computational resources (2–4). In other words, AI strives to construct—out of silicon (or whatever) and information—a synthetic *homo economicus*, perhaps more accurately termed *machina economicus*.

The shared rationality abstraction provides a strong foundation for research that spans AI and economics. We start this review by describing progress on the question of how to operationalize rationality and how to construct AI agents that are able to reason about other AIs. Supposing that AI research succeeds in developing an agent that can be usefully modeled as rational (perhaps

more so than human agents), we turn to research on the design of systems populated by multiple AIs. These multi-agent systems will function as AI economies, with AIs engaged in transactions with other AIs as well as with firms and people. This prospect has spawned interest in expanding theories of normative design from economics, optimizing rules of encounter (5) to guide multi-agent interactions. Systems populated by AIs may exhibit new economic phenomena and thus require a new science with which to understand the way they function and to guide their design. For example, although human cognitive constraints limit the design of current markets, systems designed for AIs may admit more complex interfaces, impose greater calculation burdens, and demand more stamina of attention.

At the same time, the ways in which the behavior of AIs deviate from the behavior of people can present new challenges. We can already glimpse the future of economic AIs, with simple AI bots pricing books for sale on Amazon and scanning for restaurant tables on OpenTable for resale at a profit (6). Such AIs may introduce some efficiencies, but their lack of common sense and their designer’s failure to anticipate interactions can also lead to books priced at \$23 million (7). More sophisticated AI strategies, presumably more carefully vetted, exert a large influence on financial markets, with automated trading algorithms estimated to be responsible for more than 70% of trades on U.S. stock markets (8). Given the consequences, it is important to understand the effect of ubiquitous automated agents on the performance of economic systems. As reasoning is shifted from people to AIs—designed to learn our preferences, overcome our decision biases, and make complex cost-benefit trade-offs—how too should the economic institutions that mediate everyday transactions change?

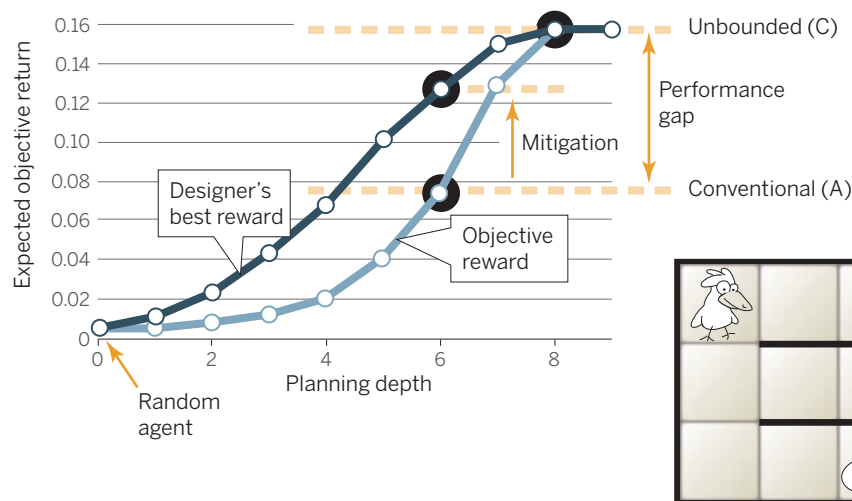
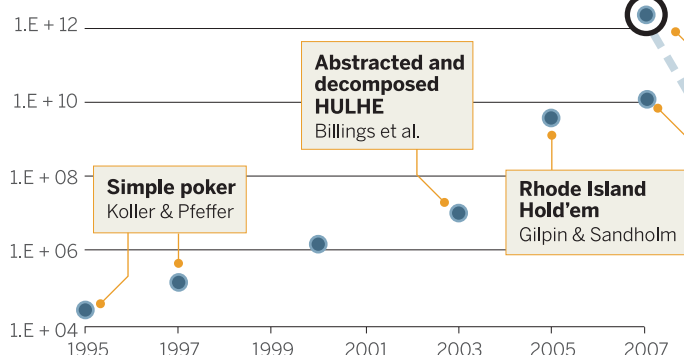


Fig. 1. A bounded reinforcement learning agent performs better by pursuing a designed reward function different from the objective reward: its actual fitness evaluation. Results (left) from a gridworld foraging domain (right), for various limits on the agent’s planning horizon (84). Unless the agent is perfectly rational (i.e., no horizon limit)—not typically feasible in realistic applications—the designer can often achieve better fitness by directing the agent to optimize an alternative measure.

¹Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge MA 02138, USA. ²Computer Science and Engineering, University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109, USA. *Corresponding author. E-mail: parkes@eecs.harvard.edu (D.C.P.); wellman@umich.edu (M.P.W.)

Game tree size

(nodes)



Game tree size

(information sets)

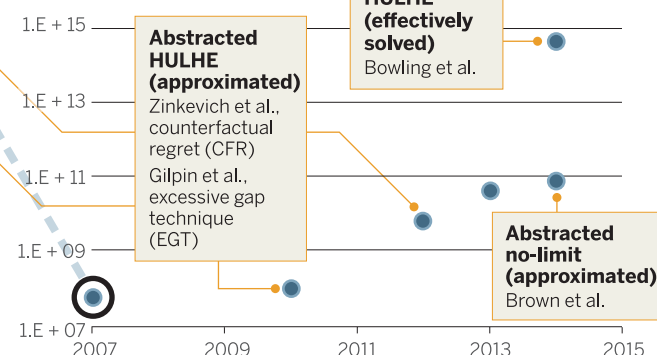


Fig. 2. Researchers produced steady exponential progress on solving games of imperfect information from 1995 to the present. Up to 2007 (left), game size was generally reported in terms of nodes in the game tree. Based on methods introduced around that time, it became more meaningful (right) to report size in terms of the number of information sets (each many nodes), which represent distinct situations as perceived from the perspective of a player. The circled data points correspond to the same milestone; combining the two graphs thus demonstrates the continual exponential improvement. Data are from (23, 35, 85–90).

We focus here on some of the research directions we consider most salient for a future synthesis of economics and AI engendered by the emergence of *machina economicus*. Interesting as they are, we only briefly mention here the many exciting applications of AI to problems in economics such as matching (9), market clearing (10), and preference modeling for smart grids (11). Nor will we showcase the many ways in which economic theory is finding application today within AI—for example, game-theoretic approaches to multi-agent learning (12) and voting procedures to combine the opinions of AIs (13).

Building *machina economicus*

Constructing a rational AI raises a host of technical challenges not previously addressed in the long tradition of rationalistic modeling in the social sciences. For economics, the agent attitudes (e.g., beliefs and preferences) underlying rationality are conceptual abstractions. Economists need not explain how capabilities and preferences, for example, are encoded, nor the algorithm by which an agent plans what actions to take conditional on its perceptions. Computation is abstracted away in the standard economic model and is precisely what the AI scientist must account for to operationalize rationality in a realized agent.

This does not mean that an AI design needs to incorporate data structures corresponding directly to rationality constructs, although many AI architectures do feature direct representations for propositions, goals, and the like. Such representations may simplify the analysis of AI systems—for example, we can ask whether an inference algorithm operating on logical expressions possesses desirable properties such as soundness: that all conclusions follow from the premises. Similarly, if an AI's beliefs are encoded as probability distributions, we can ask whether it updates its beliefs from observations in proper accord with Bayesian theory. However, care must be taken in under-

standing an agent's attitudes solely in terms of its internal data structures. Imperfections in decision-making may mean that the beliefs held and objectives pursued by a computational agent, in effect, vary systematically from those directly encoded.

As an example illustrating this distinction, machine-learning researchers adapted from animal learning the concept of reward shaping (14). In reinforcement learning, the agent derives a policy (mapping from perception sequences to actions) based on rewards representing instantaneous value associated with a state and action. A designer specifying the input reward can often train the agent more efficiently by shaping the reward signal over the learning process to facilitate convergence to behavior optimizing the designer's objective. The framework of optimal rewards (15) provides a general treatment distinguishing reward specifications and designer goals. As shown in Fig. 1, the optimal reward input to the agent does not generally correspond to the designer's ideal reward. This perspective helps explain the role of intrinsic motivations (e.g., curiosity) in a flexible learning agent.

Although the mantle of designing *machina economicus* may not be adopted (particularly in such explicit terms) by all AI researchers, many AI advances over the past few decades can be characterized as progress in operationalizing rationality. For instance, probabilistic reasoning was largely eschewed by AI 30 years ago but now pervades the field, thanks to developments in representation and inference using Bayesian networks and related graphical formalisms. Expressing uncertainty about general relationships, beyond mere propositions, is routinely supported in probabilistic modeling languages (16). Statistical approaches now dominate machine learning and natural language processing (17, 18). Likewise, preference handling (including methods for eliciting preferences from the designer of an AI agent, compactly representing preferences over com-

plex domains, and enabling inference about preferences) is regarded as a necessary AI facility. Planning, the AI subfield concerned with action over time, now conventionally frames its problem as one of optimization, subject to resource constraints, multiple objectives, and probabilistic effects of actions.

Will AI succeed in developing the ideal rational agent? As much as we strive to create *machina economicus*, absolutely perfect rationality is unachievable with finite computational resources. A more salient question is whether AI agents will be sufficiently close to the ideal as to merit thinking about them and interacting with them in rationalistic terms. Such is already the case, at least in a limited sense. Whenever we anthropomorphize our machines, we are essentially treating them as rational beings, responding to them in terms of our models of their knowledge, goals, and intentions. A more refined version of the question is whether our formal rationality theories will fit well the behavior of AI agents in absolute terms or compared to how well the theories work for people. Without offering any judgment on the question of how well rationality theories capture essential human behavior, we note the irony in the prospect that social science theories may turn out to apply with greater fidelity to nonhuman agent behavior.

Reasoning about other agents

The issue of agent theorizing is not merely academic. If we can build one AI agent, then we can build many, and these AIs will need to reason about each other as well as about people. For AIs designed to approximate *machina economicus*, it stands to reason that they should treat each other as rational, at least as a baseline assumption. These AIs would adopt a game-theoretic view of the world, where agents rationally respond to each others' behavior, presumed (recursively) to be rational as well. A consequence is that agents would expect their joint decisions to be in some form of equilibrium, as in standard economic thinking.

That AIs (or AI-human combinations) are reasonably modeled as approximately rational is the premise of a growing body of AI research applying economic equilibrium models to scenarios involving multiple agents (19). The approach has achieved notable successes, providing evidence for the premise, at least in particular circumstances. Just as single-agent rationality does not require literal expected-utility calculations, applicability of an equilibrium model does not require that agents themselves be explicitly engaged in equilibrium reasoning. For example, the literature on learning in games (20) has identified numerous conditions in which simple adaptive strategies converge to strategic equilibria. We can evaluate the effectiveness of economic modeling by examining agents built by AI designers for specified tasks. For instance, in a study of AI trading agents competing in a shopping game (21), an agent using standard price equilibrium models from economics (specifically, Walrasian equilibrium) achieved comparable prediction accuracy to sophisticated machine-learning approaches without using any data, even though none of the other agents employed equilibrium reasoning.

A Prisoner's dilemma

	Cooperate	Defect
Cooperate	4,4	0,6
Defect	6,0	1,1

Fig. 3. Each entry gives the utility to (row player, column player). (A) Prisoner's dilemma. The dominant strategy equilibrium is (Defect, Defect). **(B)** Mediated prisoner's dilemma. The dominant strategy equilibrium is (Mediator, Mediator).

In the rest of this section, we describe further examples in which economic modeling, in the form of game-theoretic algorithms, has provided an effective way for AIs to reason about other agents. The first example is computer poker. Although poker is an artificial game, many humans have invested a great deal of time and money to develop their playing skills. More important, poker's uncertainty and complexity have made it a compelling challenge problem for AI techniques. Early approaches aimed to capture the knowledge of expert human players (22), but over the past decade, game-theoretic algorithms have predominated. Technically, poker is a game of imperfect information, where each player knows elements of history (cards dealt to them) that are secret from others. As uncertainty gets partially resolved over time, through card turns and betting, players must update their beliefs about both card outcomes and the beliefs of others.

A major milestone in computer poker was achieved in 2014 with the effective solution of "heads up limit hold'em" (HULHE), which is a standard two-player version of the most popular

poker game (23). HULHE is the largest game of imperfect information ever solved (with more than 10^{13} information sets after removing symmetries) and the first imperfect-information game widely played by humans to be solved. The solution was the culmination of two decades of effort by a series of researchers (see Fig. 2), beginning with the exact solution of simplified poker games, and proceeding to the approximate solution of abstracted versions of the full game (24). Computing the approximate Nash equilibrium of the full game required massive computation and new methods for equilibrium search based on regret-matching techniques from machine learning. The result is a strategy against which even a perfect opponent cannot earn a detectable profit.

In general, the optimal strategy against perfect opponents may not be the ideal strategy against the more typical fallible kind. Despite considerable effort, however, researchers have not found poker algorithms that perform considerably better than game-theoretic solutions, even against natural distributions of opponents. It has also turned out that game-theoretic approaches have been more successful than alternatives, even for

B Mediated Prisoner's dilemma

	Mediator	Cooperate	Defect
Mediator	4,4	6,0	1,1
Cooperate	0,6	4,4	0,6
Defect	1,1	6,0	1,1

poker variants that are far from being exactly solved, such as no-limit (where bets are unrestricted) (25), or games with three or more players (26).

Much of the interest in game-theoretic reasoning for AI is driven by its applicability to real-world problems. The most prominent area of application in recent years, and our second example, is that of security games, based on a pioneering series of systems developed by Tambe *et al.* (27). In these systems, an agent decides how to defend facilities (e.g., airport security through placement of checkpoints) by solving a game where an attacker is presumed to rationally plan in response to the defender's decision. This approach has been successfully deployed in a variety of domains, including airport and airline security and coast guard patrols.

As for any game-theoretic approach, the recommendations from these systems are sensitive to assumptions made about the other agents (here, attackers): their respective preferences, beliefs, capabilities, and level of rationality. Representational approaches from AI provide flexibility, allowing the assumptions made in the strict versions typically employed by game theorists to

be relaxed (28). The field of behavioral game theory has developed detailed predictive models based on how humans have been observed to deviate from game-theoretic rationality (29). Such predictive models can be readily incorporated in existing game-theoretic reasoning algorithms, as has been demonstrated in the context of modeling attackers in security games (30). An interesting open question is whether the kinds of behavioral models that best explain human decision-making [see Wright and Leyton-Brown (31) for a meta-study] will also prove effective in capturing the bounded rationality of computational agents.

Designing multi-agent systems

At the multi-agent level, a designer cannot directly program behavior of the AIs but instead defines the rules and incentives that govern interactions among AIs. The idea is to change the "rules of the game" (e.g., rewards associated with actions and outcomes) to effect change in agent behavior and achieve system-wide goals. System goals might include, for instance, promoting an allocation of resources to maximize total value, coordinating behavior to complete a project on time, or pooling decentralized information to form an accurate prediction about a future event. The power to change the interaction environment is special and distinguishes this level of design from the standard AI design problem of performing well in the world as given.

An interesting middle ground is to take the world as given but employ reliable entities—mediators—that can interact with AIs and perform actions on their behalf (32). Introducing mediating entities is relatively straightforward in the new AI economy. To see how this can be powerful, consider a mediated extension of the classic prisoner's dilemma game (Fig. 3). If both AIs grant the mediator the authority to play on their behalf (i.e., proxy right), it performs Cooperate on behalf of both agents. However, if only one AI grants the mediator proxy, it performs Defect on behalf of that agent. In equilibrium, both AIs grant proxy, and the effect is to change the outcome from (Defect, Defect) to (Cooperate, Cooperate), increasing utility to both participants.

For the more general specification of rules of interaction for rational agents, economics has a well-developed mathematical theory of mechanism design (33). The framework of mechanism design has been fruitfully applied, for example, to the design of matching markets (34) and auctions (35). Mechanism design is a kind of inverse game theory, with the rules inducing a game and the quality of the system evaluated in an equilibrium. In the standard model, design goals are specified in terms of agent preferences on outcomes, but these preferences are private and the agents are self-interested. A mechanism is a trusted entity, able to receive messages from agents that make claims (perhaps untruthfully) about preferences and select an outcome (e.g., an allocation of resources or a plan of behavior) on the basis of these messages. The challenge is to align incentives and promote truthful reports.

Varian (36) has argued that the theory of mechanism design may actually prove more relevant for artificial agents than for human agents, because AIs may better respect the idealized assumptions of rationality made in this framework. For example, one desirable property of a mechanism is incentive compatibility, which stipulates that truthful reports constitute an equilibrium. Sometimes it is even possible to make truthful reporting a dominant strategy (optimal whatever others do), achieving the strong property of strategy-proofness (37). It seems, however, that people do not reliably understand this property; evidence from medical matching markets, and also from laboratory experiments, suggests that some participants in strategy-proof matching mechanisms try to misrepresent their preferences even though it provides no advantage (38, 39).

For artificial systems, in comparison, we might expect AIs to be truthful where this is optimal and to avoid spending computation reasoning about the behavior of others where this is not useful (5). More generally, mechanism designs for AI systems need not be simple because they need not be understandable to people. On the contrary, AI techniques such as preference representation, preference elicitation, and search algorithms can be used to turn the mathematical formalisms of mechanism design into concrete computational methods (40–42). The design problem itself can also be usefully formulated as a computational problem, with optimization and machine learning used to find solutions to design problems for which analytical solutions are unavailable (43–46).

The prospect of an economy of AIs has also inspired expansions to new mechanism design settings. Researchers have developed incentive-compatible multiperiod mechanisms, considering such factors as uncertainty about the future and changes to agent preferences because of changes in local context (47–49). Another direction considers new kinds of private inputs beyond preference information (50, 51). For example, in a team formation setting, each AI might misreport information about the capabilities of other AIs in order to get itself selected for the team (52). Similarly, AIs seeking to maximize task assignments might provide false reports of experience in task performance in order to mislead a learning mechanism constructing an automatic task classifier (53). Systems of AIs can also create new challenges for mechanism design. One such challenge is false-name bidding, where an AI exploits its ability to manage multiple identities. For example, it may gain resources more cheaply by dividing a request into a set of smaller requests, each placed from a different identity under its control. In response, researchers have developed mechanisms that are robust to this new kind of attack (54).

The important role of mechanism design in an economy of AIs can be observed in practice. Search engines run auctions to allocate ads to positions alongside search queries. Advertisers bid for their ads to appear in response to specific queries (e.g., “personal injury lawyer”). Ads are ranked according

to bid amount (as well as other factors, such as ad quality), with higher-ranked ads receiving a higher position on the search results page. Early auction mechanisms employed first-price rules, charging an advertiser its bid amount when its ad receives a click. Recognizing this, advertisers employed AIs to monitor queries of interest, ordered to bid as little as possible to hold onto the current position. This practice led to cascades of responses in the form of bidding wars, amounting to a waste of computation and market inefficiency (55). To combat this, search engines introduced second-price auction mechanisms (37), which charge advertisers based on the next-highest bid price rather than their own price. This approach (a standard idea of mechanism design) removed the need to continually monitor the bidding to get the best price for position, thereby ending bidding wars (56).

In recent years, search engine auctions have supported richer, goal-based bidding languages.

The tangle between automated agents and the design of rules of interaction also features prominently in today's financial markets, where the dominance of computerized traders has, by most accounts, qualitatively shaped the behavior of these markets. Although details of implementation are closely held secrets, it is well understood that techniques from AI and machine learning are widely employed in the design and analysis of algorithmic traders (66). Algorithmic trading has enabled the deployment of strategies that exploit speed advantages and has led in turn to a costly arms race of measures to respond to market information with minimum latency. A proposed design response would replace continuous-time auctions with periodic auctions that clear on the order of once per second, thus negating the advantage of tiny speed improvements (67, 68).

We describe two additional examples of the design of multi-agent systems for an economy of AIs. The first example system aggregates

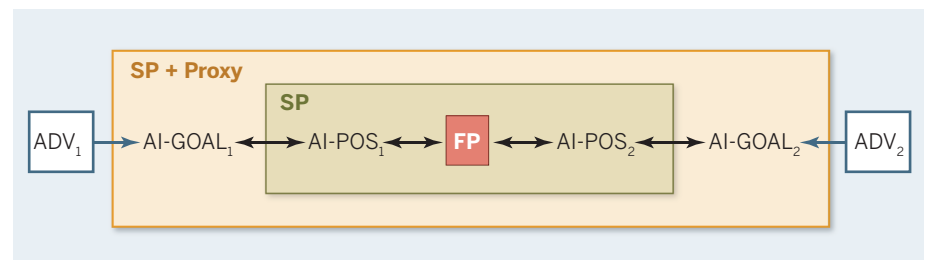


Fig. 4. Two generations of sponsored search mechanisms. Early designs were first price (FP), and advertisers (ADV) used AIs (AI-POS) to maintain a position on the list of search results at the lowest possible price. Second-price (SP) auction mechanisms were introduced, designed to replace the combination of FP and AI-POS. Advertisers adopted new AIs (AI-GOAL) to achieve higher-level goals such as to maximize profit or to maximize the number of clicks. The second price auction was extended to include proxy agents (SP+Proxy), designed to replace the combination of SP and AI-GOAL.

For example, an advertiser can ask to maximize clicks over a weighted set of queries subject to a budget constraint (57, 58). Search engines provide proxy agents that then bid on behalf of advertisers to achieve the stated goal (59). This introduction of proxy agents and the earlier switch from first price to second price can be interpreted as a computational application of a fundamental concept in mechanism design—the revelation principle (60–62). Briefly, this states that if the rules of a mechanism and the equilibrium strategies in that mechanism are replaced by a new mechanism that is functionally equivalent to the composition of these rules and strategies, then the new mechanism will be incentive compatible. Although neither redesign provides incentive compatibility in a formal sense, both second-pricing and proxy bidding can be interpreted as accomplishing on behalf of advertisers what they were doing (through AIs) in an earlier design (see Fig. 4). Still other ad platform designs are using a strategy-proof mechanism [the Vickrey-Clarke-Groves mechanism (37, 63, 64)] to make decisions about the space to allocate to ads, which ads to allocate, and which (nonsponsored) content to display to a user (65).

information held by multiple AIs. The rules of a system that achieves this goal can be engineered purposefully through the design of a prediction market (69). Popular versions of prediction markets feature questions such as who will be elected U.S. president (e.g., Betfair offers many such markets). The basic idea of a prediction market is to facilitate trade in securities contracts (e.g., a possible contract will pay \$1 if Hilary Clinton is elected). The price that balances supply and demand is then interpreted as a market prediction (e.g., price \$0.60 reflects probability 0.6 for the payoff event).

Consider a domain with a large number of interrelated random variables—for example, “flight BA214 delayed by more than 1 hour,” “snowstorm in Boston,” “de-icing machine fail,” “incoming flight BA215 delayed by more than 1 hour,” and “security alert in London.” In a combinatorial prediction market (70), a large bet on the contract “de-icing machine fail” would affect the price of “flight BA214 delayed by more than 1 hour” and all other connected events. A challenge is that the number of conceivable events is exponential in the number of random variables. Among other properties, a good market design should allow bets

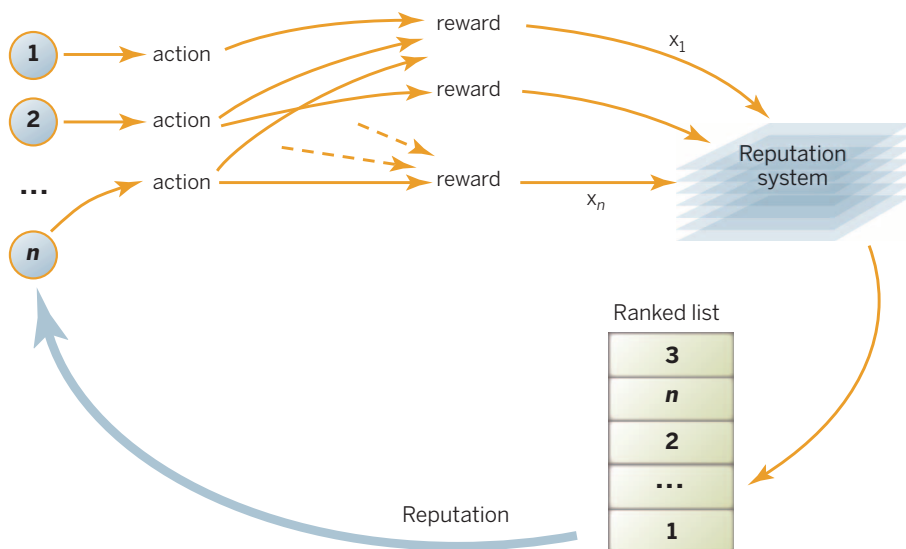


Fig. 5. In a reputation system for a multi-agent AI, each agent chooses an action, and the combined effect of these actions generates rewards (i.e., utility). Based on the actions taken and the rewards received, agent i can submit a report, x_i , to the reputation system. The reputation system aggregates this feedback—for example, providing a ranked list to reflect the estimated trustworthiness of agents. Each agent observes this ranked list, and this information may influence future actions.

on all events about which AIs have information (e.g., “de-icing machine fail AND all subsequent flights from Boston delayed by more than 1 hour”). A good design should also align incentives—for example, making it utility-maximizing to trade immediately on current information until the market price reflects an agent’s belief. Progress in scaling up combinatorial markets has been made by relating the problem of pricing bets to well-understood problems in statistical inference and convex optimization (71, 72). Related research advances are being made by allowing AIs to transact in hypotheses that are acquired through machine learning as well as trade directly in information signals rather than beliefs (73–75).

The second example is the management of information concerning the trustworthiness of agents within an economy of AIs. Trust that a counterparty will complete a transaction or invest effort or resources is crucial for any well-functioning economic system. A standard approach is to associate participants with a reputation, which can serve to align incentives in the present under the threat of a damaged reputation and lost opportunities in the future. In addition to this problem of moral hazard (i.e., will agents behave cooperatively when completing economic transactions), reputation systems can address the problem of adverse selection (i.e., will high-quality agents choose to enter a market in the first place) (76, 77).

A special challenge in an economy of AIs arises because of the fluidity of identity and the ease with which agents can be replaced. This raises, for example, the specter of whitewashing attacks, where an AI repeatedly runs down its reputation before reentering with a different identity. Without the possibility of enforcing strong identities that cannot be changed, this suggests a social cost of fluid identities, where it becomes neces-

sary to impose a penalty on all new participants and make them build up reputations from an assumption of being untrustworthy (78).

We should also consider that *machina economicus* will be strategic in sharing feedback on other AIs. For example, in eBay’s original reputation system, buyers were often reluctant to leave negative feedback about deadbeat sellers, because the sellers could retaliate with negative feedback about the buyer. In response, eBay introduced an additional feedback mechanism that was one-directional from the buyer to the seller and could not be easily traced to a particular buyer. The change resulted in a greater amount of negative feedback (79).

The economy of AIs also offers positive opportunities for promoting trust through bookkeeping, collecting feedback, and tracking the provenance of feedback in novel reputation mechanisms (see Fig. 5). AI researchers are designing reputation systems that align incentives with making truthful reports, while provably satisfying axiomatic properties such as symmetry: Two agents that are in an equivalent position from the perspective of reports made and received should have the same trust score (80, 81). Another example is the design of accounting systems that elicit truthful reports about the resources contributed or work performed by other AIs and enable the design of systems to mitigate free-riding and promote fair contributions to an economic system (82). Still, the extent to which effective, multi-agent AIs can be developed entirely through computational infrastructure such as reputation mechanisms and without recourse to legal systems remains an interesting open question.

Closing comments

Whatever one’s thoughts about when or whether AI will transcend human-level performance, the

rapidly advancing capabilities of AI are fueling considerable optimism and investment in AI research. AI has surpassed or will likely soon surpass humans in narrow domains such as playing chess, controlling a jumbo jet during cruise, making product recommendations, pricing millions of products on an eCommerce platform, reasoning about whether a patient is likely to be re-admitted to a hospital, and detecting signals from a massive volume of financial news stories.

Certainly, many fundamental challenges remain, including how to design reasoning and inference methods that effectively balance the benefit of additional computation with the costs that may arise from additional delay to acting in the world and how to design AI systems that can learn and generalize from reward signals in unconstrained domains. Given that decision problems related to economic transactions are often relatively well structured, however, it seems likely to us that AI will continue to make especially rapid inroads in economically important applications. This in turn will ensure continued effort on methods for rational, economic reasoning toward the broader goal of developing *machina economicus*.

We should not leave the impression that AI researchers unanimously embrace economic perspectives on single- or multi-agent AI. For some, multi-agent economic models are still seen as a distraction. After all, a centralized perspective allows focusing on overall goals without worrying about the incentives of individual parts of the system. Others conduct research into multi-agent systems composed of agents under the control of the designer, so that they can be programmed in any way desired. Just as with centralized solutions, these so-called “cooperative” multi-agent systems allow design without concern for the self-interest of individual agents, albeit often with decomposition or communication constraints. But cooperative versus self-interested is really a difference in assumptions on the power of a system designer, rather than a technical dispute. The viewpoint that we ascribe to is that a large number of AI systems will, given the existing structure of human economic systems, be populated by AIs that are designed, deployed, owned, and operated by a myriad of different parties, each with possibly misaligned goals. Finally, some may object to the economic approach on the basis that AIs are and will remain far from perfectly rational, simply by virtue of physical and computational limits. More direct models of the AIs’ computational behavior, in terms of the automata they are, could in principle be more accurate. The analytical utility of a rationality abstraction for AIs is ultimately an empirical question to be resolved as AI progresses.

Among those adopting an economic approach, there persist some disagreements on specific techniques—for example, on the role of equilibrium reasoning. Even if agents can be viewed as rational, some question whether it is plausible that they reach equilibrium configurations, the

particularly in situations where multiple equilibria exist. As Shoham (83) argues, game theory lacks a well-accepted pragmatic account of how it should be deployed in concrete reasoning contexts. A positive view is that AI researchers, in their efforts to operationalize economic reasoning, are developing exactly this needed body of pragmatics.

Some may object that mechanism design is too idealized even for systems of AIs—for example, in its insistence on design under equilibrium behavior, its assumption that rules of interaction can be designed from scratch, and its lack of attention to the details of the human and legal contexts in which designed systems will operate. A positive view is that AI systems are precisely the kinds of environments where we can build *tabula rasa* new rules of interaction, because these rules will be realized through the Internet and as programs running on computer servers. That such rules of interaction can come into existence is as much a matter of science and engineering as it is of public policy.

As AI advances, we are confident that economic reasoning will continue to have an important role in the design of single-agent and multi-agent AIs, and we have argued that, as economies of AIs continue to emerge, there will need to be a new science to understand how to design these systems. These AIs will no doubt exert strong forces on the economy and broader society; understanding the effect and extent of this will shape the research agendas of both AI and economics in years to come.

REFERENCES AND NOTES

1. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, ed. 3, 2009).
2. J. Doyle, *Comput. Intell.* **8**, 376–409 (1992).
3. E. J. Horvitz, *Computation and action under bounded resources*, thesis, Stanford University (1990).
4. S. Russell, in *Fundamental Issues of Artificial Intelligence*, V. C. Muller, Ed. (Springer, Berlin, 2015).
5. J. S. Rosenschein, G. Zlotkin, *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers* (MIT Press, Cambridge, 1994).
6. M. McCue, Can't secure a reservation to a hot restaurant? Try using an algorithm, *Fortune*, 3 September 2014; <http://fortune.com/2014/09/03/opentable-dinematic-tablesweep-restaurant-reservation-service/>.
7. J. D. Sutter, CNN, 25 April 2011; <http://edition.cnn.com/2011/TECH/web/04/25/amazon.price.algorithm>.
8. T. Hendershott, C. M. Jones, A. J. Menkveld, *J. Finance* **66**, 1–33 (2011).
9. D. J. Abraham, A. Blum, T. Sandholm, 8th ACM Conference on Electronic Commerce (2007), pp. 295–304.
10. A. Frechette, N. Newman, K. Leyton-Brown, 24th International Joint Conference on Artificial Intelligence (2015).
11. S. D. Ramchurn, P. Vytelingum, A. Rogers, N. R. Jennings, *Commun. ACM* **55**, 86–97 (2012).
12. R. I. Brafman, M. Tennenholtz, *Artif. Intell.* **159**, 27–47 (2004).
13. A. X. Jiang et al., *Adv. Neural Inf. Process. Syst.* **27**, 2573–2581 (2014).
14. A. Y. Ng, D. Harada, S. J. Russell, 16th International Conference on Machine Learning (1999), pp. 278–287.
15. S. Singh, R. L. Lewis, A. G. Barto, J. Sorg, *IEEE Transactions on Autonomous Mental Development* **2**, 70–82 (2010).
16. L. Getoor, B. Taskar, Eds., *Introduction to Statistical Relational Learning* (MIT Press, Cambridge, 2007).
17. M. I. Jordan, T. M. Mitchell, *Science* **349**, 255–260 (2015).
18. J. Hirschberg, C. D. Manning, *Science* **349**, 261–266 (2015).
19. Y. Shoham, K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic and Logical Foundations* (Cambridge Univ. Press, Cambridge, 2009).
20. D. Fudenberg, D. K. Levine, *The Theory of Learning in Games* (MIT Press, Cambridge, 1998).
21. M. P. Wellman, D. M. Reeves, K. M. Lochner, Y. Vorobeychik, *J. Artif. Intell. Res.* **21**, 19–36 (2004).
22. D. Billings, A. Davidson, J. Schaeffer, D. Szafron, *Artif. Intell.* **134**, 201–240 (2002).
23. M. Bowling, N. Burch, M. Johanson, O. Tammelin, *Science* **347**, 145–149 (2015).
24. T. Sandholm, *AI Mag.* **31**, 13–32 (2010).
25. N. Brown, S. Ganzfried, T. Sandholm, 14th International Conference on Autonomous Agents and Multi-Agent Systems (2015), pp. 7–15.
26. R. Gibson, Regret minimization in games and the development of champion multiplayer computer poker-playing agents, Ph.D. thesis, University of Alberta (2014).
27. M. Tambe, *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned* (Cambridge Univ. Press, Cambridge, 2011).
28. Y. Gal, A. Pfeffer, *J. Artif. Intell. Res.* **33**, 109–147 (2008).
29. C. F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton Univ. Press, Princeton, 2003).
30. R. Yang, C. Kiekintveld, F. Ordóñez, M. Tambe, R. John, *Artif. Intell.* **195**, 440–469 (2013).
31. J. R. Wright, K. Leyton-Brown, Evaluating, understanding, and improving behavioral game theory models for predicting human behavior in unreplicated normal-form games, CoRR abs/1306.0918 (2013); <http://arxiv.org/abs/1306.0918>.
32. D. Monderer, M. Tennenholtz, *Artif. Intell.* **173**, 180–195 (2009).
33. L. Hurwicz, in *Mathematical Methods in the Social Sciences*, K. J. Arrow, S. Karlin, P. Suppes, Eds. (Stanford University Press, Stanford 1960), Ch. 3, pp. 27–46.
34. T. Sönmez, U. Ünver, in *Handbook of Social Economics*, A. Bisin, J. Benhabib, M. Jackson, Eds. (North-Holland, 2011), vol. 1A, pp. 781–852.
35. P. Milgrom, *Putting Auction Theory to Work* (Cambridge Univ. Press, Cambridge, 2004).
36. H. R. Varian, 1st USENIX Workshop on Electronic Commerce (1995), pp. 13–21.
37. W. Vickrey, *J. Finance* **16**, 8–37 (1961).
38. Y. Chen, T. Sönmez, *J. Econ. Theory* **127**, 202–231 (2006).
39. F. Echenique, A. J. Wilson, L. Yariv, Clearinghouses for two-sided matching: An experimental study, Working Papers 487, University of Pittsburgh, Department of Economics (2013).
40. N. Nisan, in *Combinatorial Auctions*, P. Cramton, Y. Shoham, R. Steinberg, Eds. (MIT Press, Cambridge, 2006), chap. 9.
41. T. Sandholm, C. Boutilier, in *Combinatorial Auctions*, P. Cramton, Y. Shoham, R. Steinberg, Eds. (MIT Press, Cambridge, 2006), chap. 10.
42. T. Sandholm, *Artif. Intell.* **135**, 1–54 (2002).
43. V. Conitzer, T. Sandholm, 18th Conference on Uncertainty in Artificial Intelligence (2002), pp. 103–110.
44. S. Alaei, H. Fu, N. Haghpour, J. D. Hartline, A. Malekian, 13th ACM Conference on Electronic Commerce (2012), p. 17.
45. Y. Cai, C. Daskalakis, S. M. Weinberg, 54th Annual IEEE Symposium on Foundations of Computer Science (2013), pp. 618–627.
46. P. Duetting et al., *ACM Transactions on Economics and Computation* **3**, 5:1–5:41 (2015).
47. D. C. Parkes, S. Singh, *Adv. Neural Inf. Process. Syst.* **16**, 791–798 (2003).
48. R. Cavallo, D. C. Parkes, S. Singh, 22nd Conference on Uncertainty in Artificial Intelligence (Cambridge, MA, 2006), pp. 55–62.
49. D. C. Parkes, in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani, Eds. (Cambridge Univ. Press, 2007), chap. 16, pp. 411–439.
50. Y. Shoham, M. Tennenholtz, *Theor. Comput. Sci.* **343**, 97–113 (2005).
51. J. Y. Halpern, V. Teague, 36th Annual ACM Symposium on Theory of Computing (2004), pp. 623–632.
52. N. Alon, F. Fischer, A. Procaccia, M. Tennenholtz, 13th Conference on Theoretical Aspects of Rationality and Knowledge (2011), pp. 101–110.
53. O. Dekel, F. A. Fischer, A. D. Procaccia, *J. Comput. Syst. Sci.* **76**, 759–777 (2010).
54. M. Yokoo, Y. Sakurai, S. Matsubara, *Games Econ. Behav.* **46**, 174–188 (2004).
55. B. Edelman, M. Ostrovsky, *Decis. Support Syst.* **43**, 192–198 (2007).
56. B. Edelman, M. Ostrovsky, M. Schwarz, *Am. Econ. Rev.* **97**, 242–259 (2007).
57. C. Borgs et al., 16th International Conference on World-Wide Web (2007), pp. 531–540.
58. J. Feldman, S. Muthukrishnan, M. Pál, C. Stein, 8th ACM Conference on Electronic Commerce (2007), pp. 40–49.
59. A. Z. Broder, E. Gabrilovich, V. Josifovski, G. Mavromatis, A. J. Smola, 4th International Conference on Web Search and Web Data Mining (2011), pp. 515–524.
60. L. Hurwicz, in *Decision and Organization*, R. Radner, C. B. McGuire, Eds. (North-Holland, Amsterdam, 1972), Ch. 14, pp. 297–336.
61. A. Gibbard, *Econometrica* **41**, 587–602 (1973).
62. R. Myerson, *Econometrica* **47**, 61–73 (1979).
63. E. Clarke, *Public Choice* **11**, 17–33 (1971).
64. T. Groves, *Econometrica* **41**, 617–631 (1973).
65. H. R. Varian, C. Harris, *Am. Econ. Rev.* **104**, 442–445 (2014).
66. M. Kearns, Y. Nemyyaka, in *High Frequency Trading: New Realities for Traders, Markets and Regulators*, D. Easley, M. Lopez de Prado, M. O'Hara, Eds. (Risk Books, London, 2013), Ch. 5, pp. 91–124.
67. E. Budish, P. Cramton, J. Shim, The high-frequency trading arms race: Frequent batch auctions as a market design response, Tech. Rep. 14-03, Booth School of Business, University of Chicago (2015).
68. E. Wah, M. P. Wellman, 14th ACM Conference on Electronic Commerce (2013), pp. 855–872.
69. R. Forsythe, F. Nelson, G. R. Neumann, J. Wright, in *Contemporary Laboratory Experiments in Political Economy*, T. R. Palfrey, Ed. (University of Michigan Press, Ann Arbor, 1991), pp. 69–111.
70. R. D. Hanson, *Inf. Syst. Front.* **5**, 107–119 (2003).
71. J. Abernethy, Y. Chen, J. W. Vaughan, *ACM Transactions on Economics and Computation* **1**, 12:1–12:39 (2013).
72. M. Dudik, S. Lahaie, D. M. Pennock, D. Rothschild, 14th ACM Conference on Electronic Commerce (2013), pp. 341–358.
73. J. Abernethy, R. Frongillo, *Adv. Neural Inf. Process. Syst.* **24**, 2600–2608 (2011).
74. J. Witkowski, D. C. Parkes, 13th ACM Conference on Electronic Commerce (2012), pp. 964–981.
75. J. Hu, A. J. Storkey, 31st International Conference on Machine Learning (2014), pp. 1773–1781.
76. P. Resnick, K. Kuwabara, R. Zeckhauser, E. J. Friedman, *Commun. ACM* **43**, 45–48 (2000).
77. C. Dellarocas, *Manage. Sci.* **49**, 1407–1424 (2003).
78. E. J. Friedman, P. Resnick, *J. Econ. Manage. Strategy* **10**, 173–199 (2001).
79. G. Bolton, B. Greiner, A. Ockenfels, *Manage. Sci.* **59**, 265–285 (2013).
80. A. Cheng, E. Friedman, ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems (2005), pp. 128–132.
81. A. Altman, M. Tennenholtz, *J. Artif. Intell. Res.* **31**, 473–495 (2008).
82. S. Seuken, J. Tang, D. C. Parkes, 24th AAAI Conference on Artificial Intelligence (2010), pp. 860–866.
83. Y. Shoham, *Commun. ACM* **51**, 74–79 (2008).
84. J. Sorg, S. Singh, R. Lewis, 27th International Conference on Machine Learning (2010), pp. 1007–1014.
85. D. Koller, A. Pfeffer, *Artif. Intell.* **94**, 167–215 (1997).
86. D. Billings et al., 18th International Joint Conference on Artificial Intelligence (2003), pp. 661–668.
87. A. Gilpin, T. Sandholm, 7th ACM Conference on Electronic Commerce (2006), pp. 160–169.
88. M. Zinkevich, M. Johanson, M. Bowling, C. Piccione, *Adv. Neural Inf. Process. Syst.* **20**, 905–912 (2007).
89. A. Gilpin, S. Hoda, J. Peña, T. Sandholm, 3rd International Workshop on Internet and Network Economics (2007), pp. 57–69.
90. S. Hoda, A. Gilpin, J. Peña, T. Sandholm, *Math. Oper. Res.* **35**, 494–512 (2010).

ACKNOWLEDGMENTS

Thanks to the anonymous referees and to M. Bowling, Y. Chen, K. Gal, S. Lahaie, D. Pennock, A. Procaccia, T. Sandholm, S. Seuken, Y. Shoham, A. Storkey, M. Tambe, and M. Tennenholtz for their thoughtful comments on an earlier draft.

10.1126/science.aaa8403

Computational rationality: A converging paradigm for intelligence in brains, minds, and machines

Samuel J. Gershman,^{1*} Eric J. Horvitz,^{2*} Joshua B. Tenenbaum^{3*}

After growing up together, and mostly growing apart in the second half of the 20th century, the fields of artificial intelligence (AI), cognitive science, and neuroscience are reconverging on a shared view of the computational foundations of intelligence that promotes valuable cross-disciplinary exchanges on questions, methods, and results. We chart advances over the past several decades that address challenges of perception and action under uncertainty through the lens of computation. Advances include the development of representations and inferential procedures for large-scale probabilistic inference and machinery for enabling reflection and decisions about tradeoffs in effort, precision, and timeliness of computations. These tools are deployed toward the goal of computational rationality: identifying decisions with highest expected utility, while taking into consideration the costs of computation in complex real-world problems in which most relevant calculations can only be approximated. We highlight key concepts with examples that show the potential for interchange between computer science, cognitive science, and neuroscience.

Imagine driving down the highway on your way to give an important presentation, when suddenly you see a traffic jam looming ahead. In the next few seconds, you have to decide whether to stay on your current route or take the upcoming exit—the last one for several miles—all while your head is swimming with thoughts about your forthcoming event. In one sense, this problem is simple: Choose the path with the highest probability of getting you to your event on time. However, at best you can implement this solution only approximately: Evaluating the full branching tree of possible futures with high uncertainty about what lies ahead is likely to be infeasible, and you may consider only a few of the vast space of possibilities, given the urgency of the decision and your divided attention. How best to make this calculation? Should you make a snap decision on the basis of what you see right now, or explicitly try to imagine the next several miles of each route? Perhaps you should stop thinking about your presentation to focus more on this choice, or maybe even pull over so you can think without having to worry about your driving? The decision about whether to exit has spawned a set of internal decision problems: how much to think, how far should you plan ahead, and even what to think about.

This example highlights several central themes in the study of intelligence. First, maximizing some measure of expected utility provides a general-

purpose ideal for decision-making under uncertainty. Second, maximizing expected utility is nontrivial for most real-world problems, necessitating the use of approximations. Third, the choice of how best to approximate may itself be a decision subject to the expected utility calculus—thinking is costly in time and other resources, and sometimes intelligence comes most in knowing how best to allocate these scarce resources.

The broad acceptance of guiding action with expected utility, the complexity of formulating and solving decision problems, and the rise of approximate methods for multiple aspects of decision-making under uncertainty has motivated artificial intelligence (AI) researchers to take a fresh look at probability through the lens of computation. This examination has led to the development of computational representations and procedures for performing large-scale probabilistic inference; methods for identifying best actions, given inferred probabilities; and machinery for enabling reflection and decision-making about tradeoffs in effort, precision, and timeliness of computations under bounded resources. Analogous ideas have come to be increasingly important in how cognitive scientists and neuroscientists think about intelligence in human minds and brains, often being explicitly influenced by AI researchers and sometimes influencing them back. In this Review, we chart this convergence of ideas around the view of intelligence as computational rationality: computing with representations, algorithms, and architectures designed to approximate decisions with the highest expected utility, while taking into account the costs of computation. We share our reflections about this perspective on intelligence, how it encompasses interdisciplinary goals and insights, and why we think it will be increasingly useful as a shared perspective.

Models of computational rationality are built on a base of inferential processes for perceiving, predicting, learning, and reasoning under uncertainty (1–3). Such inferential processes operate on representations that encode probabilistic dependencies among variables capturing the likelihoods of relevant states in the world. In light of incoming streams of perceptual data, Bayesian updating procedures or approximations are used to propagate information and to compute and revise probability distributions over states of variables. Beyond base processes for evaluating probabilities, models of computational rationality require mechanisms for reasoning about the feasibility and implications of actions. Deliberation about the best action to take hinges on an ability to make predictions about how different actions will influence likelihoods of outcomes and a consideration of the value or utilities of the outcomes (4). Learning procedures make changes to parameters of probabilistic models so as to better explain perceptual data and provide more accurate inferences about likelihoods to guide actions in the world.

Last, systems with bounded computational power must consider important tradeoffs in the precision and timeliness of action in the world. Thus, models of computational rationality may include policies or deliberative machinery that make inferences and decisions at the “metalevel” in order to regulate base-level inferences. These decisions rely on reflection about computational effort, accuracy, and delay associated with the invocation of different base-level algorithms in different settings. Such metalevel decision-making, or “metareasoning,” can be performed via real-time reflection or as policies computed during offline optimizations. Either way, the goal is to identify configurations and uses of base-level processes with the goal of maximizing the expected value of actions taken in the world. These computational considerations become increasingly important when we consider richer representations (graphs, grammars, and programs) that support signature features of human intelligence, such as recursion and compositionality (5).

Key advances in AI on computational machinery for performing inference, identifying ideal actions, and deliberating about the end-to-end operation of systems have synergies and resonances with human cognition. After a brief history of developments in AI, we consider links between computational rationality and findings in cognitive psychology and neuroscience.

Foundations and early history

AI research has its roots in the theory of computability developed in the 1930s. Efforts then highlighted the power of a basic computing system (the Turing Machine) to support the real-world mechanization of any feasible computation (6). The promise of such general computation and the fast-paced rise of electronic computers fueled the imagination of early computer scientists about the prospect of developing computing systems that might one day both explain and replicate aspects of human intelligence (7, 8).

¹Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA. ²Microsoft Research, Redmond, WA 98052, USA. ³Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*Corresponding author. E-mail: gershman@fas.harvard.edu (S.J.G.); horvitz@microsoft.com (E.J.H.); jbt@mit.edu (J.B.T.)

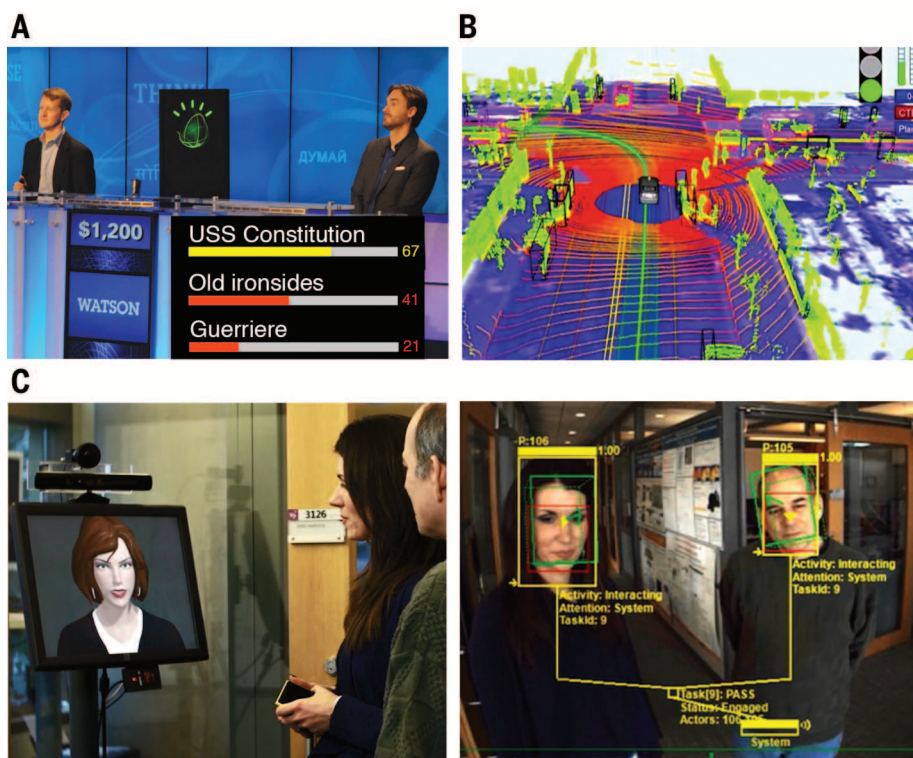


Fig. 1. Examples of modern AI systems that use approximate inference and decision-making.

These systems cannot rely on exhaustive enumeration of all relevant utilities and probabilities. Instead, they must allocate computational resources (including time and energy) to optimize approximations for inferring probabilities and identifying best actions. **(A)** The internal state of IBM Watson as it plays *Jeopardy!*, representing a few high-probability hypotheses. [Photo by permission of IBM News Room] **(B)** The internal state of the Google self-driving car, which represents those aspects of the world that are potentially most valuable or costly for the agent in the foreseeable future, such as the positions and velocities of the self-driving car, other cars and pedestrians, and the state of traffic signals. [Reprinted with permission from Google] **(C)** The Assistant (left), an interactive automated secretary fielded at Microsoft Research, recognizes multiple people in its proximity (right); deliberates about their current and future goals, attention, and utterances; and engages in natural dialog under uncertainty. [Permission from Microsoft]

Early pioneers in AI reflected about uses of probability and Bayesian updating in learning, reasoning, and action. Analyses by Cox, Jaynes, and others had provided foundational arguments for probability as a sufficient measure for assessing and revising the plausibility of events in light of perceptual data. In influential work, von Neumann and Morgenstern published results on utility theory that defined ideal, or “rational,” actions for a decision-making agent (4). They presented an axiomatic formalization of preferences and derived the “principle of maximum expected of utility” (MEU). Specifically, they showed that accepting a compact and compelling set of desiderata about preference orderings implies that ideal decisions are those actions that maximize an agent’s expected utility, which is computed for each action as the average utility of the action when considering the probability of states of the world.

The use of probability and MEU decision-making soon pervaded multiple disciplines, including some areas of AI research, such as projects in robotics. However, the methods did not gain

a large following in studies of AI until the late 1980s. For decades after the work by von Neumann and Morgenstern, probabilistic and decision-theoretic methods were deemed by many in the AI research community to be too inflexible, simplistic, and intractable for use in understanding and constructing sophisticated intelligent systems. Alternative models were explored, including logical theorem-proving and various heuristic procedures.

In the face of the combinatorial complexity of formulating and solving real-world decision-making, a school of research on heuristic models of bounded rationality blossomed in the later 1950s. Studies within this paradigm include the influential work of Simon and colleagues, who explored the value of informal, heuristic strategies that might be used by people—as well as by computer-based reasoning systems—to cut through the complexity of probabilistic inference and decision-making (9). The perspective of such heuristic notions of bounded rationality came to dominate a large swath of AI research.

Computational lens on probability

In the late 1980s, a probabilistic renaissance swept through mainstream AI research, fueled in part by pressures for performing sound inference about likelihoods of outcomes in applications of machine reasoning to such high-stakes domains as medicine. Attempts to mechanize probability for solving challenges with inference and learning led to new insights about probability and stimulated thinking about the role of related representations and inference strategies in human cognition. Perhaps most influentially, advances in AI led to the formulation of rich network-based representations, such as Bayesian networks, broadly referred to as probabilistic graphical models (PGMs) (1, 2). Belief updating procedures were developed that use parallel and distributed computation to update constellations of random variables in the networks.

The study of PGMs has developed in numerous directions since these initial advances: efficient approximate inference methods; structure search over combinatorial spaces of network structures; hierarchical models for capturing shared structure across data sets; active learning to guide the collection of data; and probabilistic programming tools that can specify rich, context-sensitive models via compact, high-level programs. Such developments have put the notions of probabilistic inference and MEU decision-making at the heart of many contemporary AI approaches (3) and, together with ever-increasing computational power and data set availability, have been responsible for dramatic AI successes in recent years (such as IBM’s Watson, Google’s self-driving car, and Microsoft’s automated assistant). These developments also raise new computational and theoretical challenges: How can we move from the classical view of a rational agent who maximizes expected utility over an exhaustively enumerable state-action space to a theory of the decisions faced by resource-bounded AI systems deployed in the real world (Fig. 1), which place severe demands on real-time computation over complex probabilistic models?

Rational decisions under bounded computational resources

Perception and decision-making incur computational costs. Such costs may be characterized in different ways, including losses that come with delayed action in time-critical settings, interference among multiple inferential components, and measures of effort invested. Work in AI has explored the value of deliberating at the meta-level about the nature and extent of perception and inference. Metalevel analyses have been aimed at endowing computational systems with the ability to make expected utility decisions about the ideal balance between effort or delay and the quality of actions taken in the world. The use of such rational metareasoning plays a central role in decision-theoretic models of bounded rationality (10–14).

Rational metareasoning has been explored in multiple problem areas, including guiding computation in probabilistic inference and decision-making

(11, 13, 14), controlling theorem proving (15), handling proactive inference in light of incoming streams of problems (16), guiding heuristic search (13, 17), and optimizing sequences of action (18–20). Beyond real-time metareasoning, efforts have explored offline analysis to learn and optimize policies for guiding real-time meta-reasoning and for enhancing real-time inference via such methods as precomputing and caching portions of inference problems into fast-response reflexes (21).

The value of metalevel reflection in computational rationality is underscored by the complexity of probabilistic inference in Bayesian networks, which has been shown to be in the nondeterministic polynomial-time (NP)-hard complexity class (22). Such worst-case complexity highlights the importance of developing approximations that exploit the structure of real-world problems. A tapestry of approximate inferential methods have been developed, including procedures that use Monte Carlo simulation, bounding methods, and methods that decompose problems into simpler sets of subproblems (1). Some methods allow a system to trade off computation time for accuracy. For example, sampling procedures can tighten the bounds on probabilities of interest with additional computation time. Characterizations of the tradeoffs can be uncertain in themselves. Other approaches to approximation consider tradeoffs incurred with modulating the complexity of models, such as changing the size of models and the level of abstraction of evidence, actions, and outcomes considered (11, 21).

A high-level view of the interplay between the value and cost of inference at different levels of precision is captured schematically in Fig. 2A.

Here, the value of computing with additional precision on final actions and cost of delay for computation are measured in the same units of utility. A net value of action is derived as the difference between the expected value of action based on a current analysis and the cost of computation required to attain the level of analysis. In the situation portrayed, costs increase in a linear manner with a delay for additional computation, while the value of action increases with decreasing marginal returns. We see the attainment of an optimal stopping time, in which attempts to compute additional precision come at a net loss in the value of action. As portrayed in the figure, increasing the cost of computation would lead to an earlier ideal stopping time. In reality, we rarely have such a simple economics of the cost and benefits of computation. We are often uncertain about the costs and the expected value of continuing to compute and so must solve a more sophisticated analysis of the expected value of computation. A metalevel reasoner considers the current uncertainties, the time-critical losses with continuing computation, and the expected gains in precision of reasoning with additional computation.

As an example, consider a reasoning system that was implemented to study computational rationality for making inferences and providing recommendations for action in time-critical medical situations. The system needs to consider the losses incurred with increasing amounts of delay with action that stems from the time required for inference about the best decision to take in a setting. The expected value of the best decision may diminish as a system deliberates about a patient's symptoms and makes inferences about

physiology. A trace of a reasoning session guided by rational metareasoning of a time-critical respiratory situation in emergency medicine is shown in Fig. 2B (14). An inference algorithm (named Bounded Conditioning) continues to tighten the upper and lower bounds on a critical variable representing the patient's physiology, using a Bayesian network to analyze evidence. The system is uncertain about the patient's state, and each state is associated with a different time criticality and ideal action. The system continues to deliberate at the metalevel about the value of continuing to further tighten the bounds. It monitors this value via computation of the expected value of computation. When the inferred expected value of computation goes to zero, the metalevel analysis directs the base-level system to stop and take the current best inferred base-level action possible.

Computational rationality in mind and brain

In parallel with developments in AI, the study of human intelligence has charted a similar progression toward computational rationality. Beginning in the 1950s, psychologists proposed that humans are "intuitive statisticians," using Bayesian decision theory to model intuitive choices under uncertainty (23). In the 1970s and 1980s, this hypothesis met with resistance from researchers who uncovered systematic fallacies in probabilistic reasoning and decision-making (24), leading some to adopt models based on informal heuristics and biases rather than normative principles of probability and utility theory (25). The broad success of probabilistic and decision-theoretic approaches in AI over the past

two decades, however, has helped to return these ideas to the center of cognitive modeling (5, 26–28). The development of methods for approximate Bayesian updating via distributed message passing over large networks of variables suggests that similar procedures might be used for large-scale probabilistic inference in the brain (29). At the same time, researchers studying human judgment and decision-making continue to uncover ways in which people's cognitive instincts appear far from the MEU ideals that economists and policymakers might have hoped for.

Computational rationality offers a framework for reconciling these contradictory pictures of human intelligence. If the brain is adapted to compute rationally with bounded resources, then "fallacies" may arise as a natural consequence of this optimization (30). For example, a generic strategy for approximating Bayesian inference is by sampling hypotheses, with the sample-based approximation converging to the true posterior as more hypotheses are

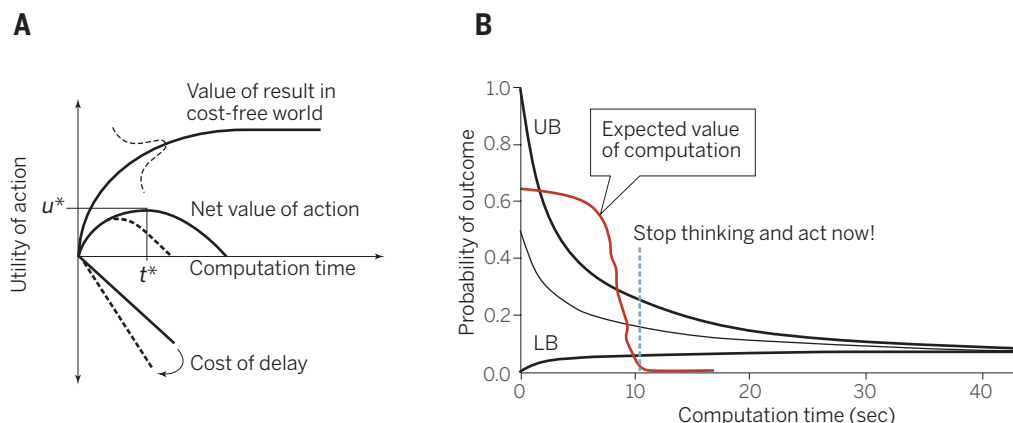


Fig. 2. Economics of thinking in computational rationality. (A) Systems must consider the expected value and cost of computation. Flexible computational procedures allow for decisions about ideal procedures and stopping times (t^*) in order to optimize the net value of action (u^*). In the general case, the cost-free value associated with obtaining a computed result at increasing degrees of precision and the cost of delay with computation are uncertain (indicated by the bell curve representing a probability distribution). Thus, the time at which further refinement of the inference should stop and action should be taken in the world are guided by computation of the expected value of computation. [Adapted from (16) with permission] (B) Trace of rational metareasoning in a time-critical medical setting. [Adapted from (14) with permission] A bounding algorithm continues to tighten the upper bound (UB) and lower bound (LB) on an important variable representing a patient's physiology. When a continually computed measure of the value of additional computation (red line) goes to zero, the base-level model is instructed to make a recommendation for immediate action.

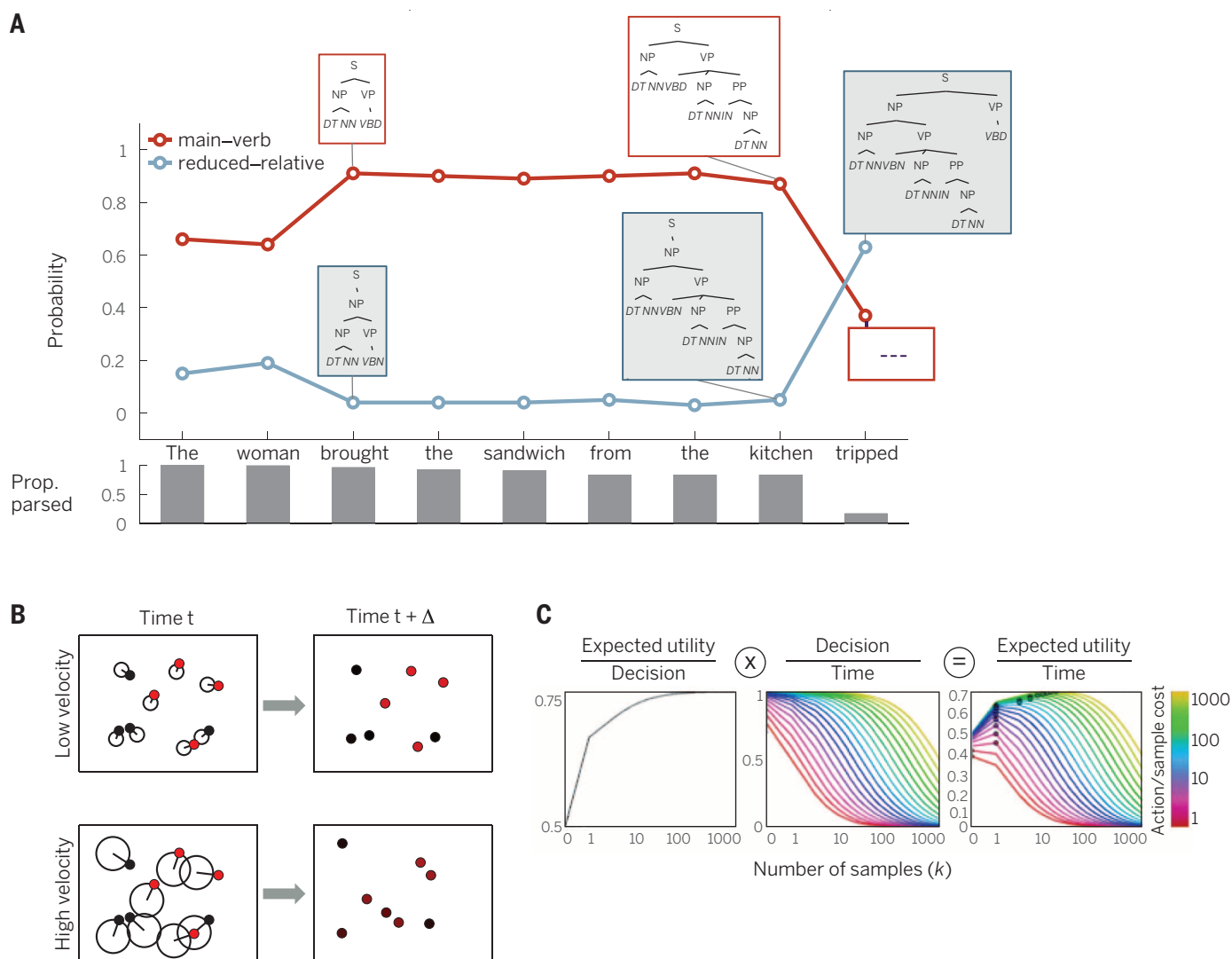


Fig. 3. Resource-constrained sampling in human cognition. (A) Incremental parsing of a garden-path sentence. [Adapted from (36)] (Top) Evolution of the posterior probability for two different syntactic parses (shown in boxes). The initially favored parse is disfavored by the end of the sentence. (Bottom) A resource-constrained probabilistic parser (based on a particle filter with few particles) may eliminate the initially unlikely parse and therefore fail to correctly parse the sentence by the end, as shown by the proportion of particle filters with 20 particles that successfully parse the sentence up to each word. **(B)** Sample-based inference for multiple-object tracking. In this task, subjects are asked to track a subset of dots over time, marked initially in red. Lines denote velocity, and circles denote uncertainty about spatial transitions. In the second frame, red shading

indicates the strength of belief that a dot is in the tracked set (More red = higher confidence) after some time interval Δ , for a particle-filter object tracker. Uncertainty scales with velocity, explaining why high-velocity objects are harder to track (32). **(C)** For a sampling-based approximate Bayesian decision-maker, facing a sequence of binary choices, expected utility per decision, and number of decisions per unit time can be combined to compute the expected utility per unit time as a function of the number of posterior samples and action/sample cost ratios. Circles in the rightmost graph indicate the optimal number of samples at a particular action/sample cost ratio. For many decisions, the optimal choice tradeoff between accuracy and computation time suggests deciding after only one sample. [Reprinted from (38) with permission]

sampled (1). Evidence suggests that humans use this strategy across several domains, including causal reasoning (31), perception (32, 33), and category learning (34). Sampling algorithms can also be implemented in biologically plausible neural circuits (35), providing a rational explanation for the intrinsic stochasticity of neurons.

We see a correspondence between the sampling algorithms humans appear to use and those used in state-of-the-art AI systems. For example, particle filters—sequential sampling algorithms

for tracking multiple objects moving in a dynamic uncertain environment (36)—are at the heart of the Google self-driving car's picture of its surroundings (Fig. 1B) and also may describe how humans track multiple objects (Fig. 3B) (32). When only a small number of hypotheses are sampled, various biases emerge that are consistent with human behavior. For instance, “garden path” effects in sentence processing, in which humans persevere on initially promising hypotheses that are disconfirmed by subse-

quent data, can be explained by particle filters for approximate online parsing in probabilistic grammars (Fig. 3A) (37). These biases may in fact be rational under the assumption that sampling is costly and most gains or losses are small, as in many everyday tasks; then, utility can be maximized by sampling as few as one or a few high-posterior probability hypotheses for each decision (Fig. 3C) (38).

This argument rests crucially on the assertion that the brain is equipped with metareasoning

mechanisms sensitive to the costs of cognition. Some such mechanisms may take the form of heuristic policies hardwired by evolutionary mechanisms; we call these “heuristic” because they would be metarational only for the range of situations that evolution has anticipated. There is also evidence that humans have more adaptive metareasoning mechanisms sensitive to the costs of cognition in online computation. In recent work with the “demand selection” task (39–41), participants are allowed to choose between two cognitive tasks that differ in cognitive demand and potential gains. Behavioral findings show that humans trade off reward and cognitive effort rationally according to a joint utility function (40). Brain imaging of the demand selection task has shown that activity in the lateral prefrontal cortex, a region implicated in the regulation of cognitive control, correlates with subjective reports of cognitive effort and individual differences in effort avoidance (41).

Several recent studies have provided support for rational metareasoning in human cognition when computational cost and reward tradeoffs are less obvious (42, 43). As an example, humans have been found to consistently choose list-sorting strategies that rationally trade time and accuracy for a particular list type (42). This study joins earlier work that has demonstrated adaptive strategy selection in humans (44, 45) but goes beyond them by explicitly modeling strategy selection using a measure of the value of computation. In another study (46), humans were found to differentially overestimate the frequency of highly stressful life events (such as lethal accidents and suicide). This “fallacy” can be viewed as rational under the assumption that only a small number of hypotheses can be sampled: Expected utility is maximized by a policy of utility-weighted sampling.

Computational tradeoffs in sequential decision-making

Computational rationality has played an important role in linking models of biological intelligence at the cognitive and neural levels in ways that can be seen most clearly in studies of sequential decision-making. Humans and other animals appear to make use of different kinds of systems for sequential decision-making: “model-based” systems that use a rich model of the environment to form plans, and a less complex “model-free” system that uses cached values to make decisions (47). Although both converge to the same behavior with enough experience, the two kinds of systems exhibit different tradeoffs in computational complexity and flexibility. Whereas model-based systems tend to be more flexible than the lighter-weight model-free systems (because they can quickly adapt to changes in environment structure), they rely on more expensive analyses (for example, tree-search or dynamic programming algorithms for computing values). In contrast, the model-free systems use inexpensive, but less flexible, look-up tables or function approximators. These efforts have conceptual links to efforts in AI that have sought to

reduce effort and to speed up responses in real time by optimizing caches of inferences via off-line precomputation (21).

Studies provide evidence that model-based and model-free systems are used in animal cognition and that they are supported by distinct regions of the prefrontal cortex (48) and striatum (49). Evidence further suggests that the brain achieves a balance between computational tradeoffs by using an adaptive arbitration between the two kinds of systems (50, 51). One way to implement such an arbitration mechanism is to view the invocation of the model-based system as a meta-action whose value is estimated by the model-free system (51).

Early during learning to solve a task, when the model-free value estimates are relatively inaccurate, the benefits of using the model-based sys-

tem outweigh its cognitive costs. Thus, moderately trained animals will be sensitive to changes in the causal structure of the environment (for example, the devaluation of a food reinforcer by pairing it with illness). After extensive training, the model-free values are sufficiently accurate to attain a superior cost-benefit tradeoff (46). This increasing reliance on the model-free system manifests behaviorally in the form of “habits”—computationally cheap but inflexible policies. For example, extensively trained animals will continue pursuing a policy that leads to previously devalued reinforcers (52).

The arbitration mechanism described above appears to adhere to the principles of computational rationality: The model-based system is invoked when deemed computationally advantageous through metareasoning (Fig. 4A). For

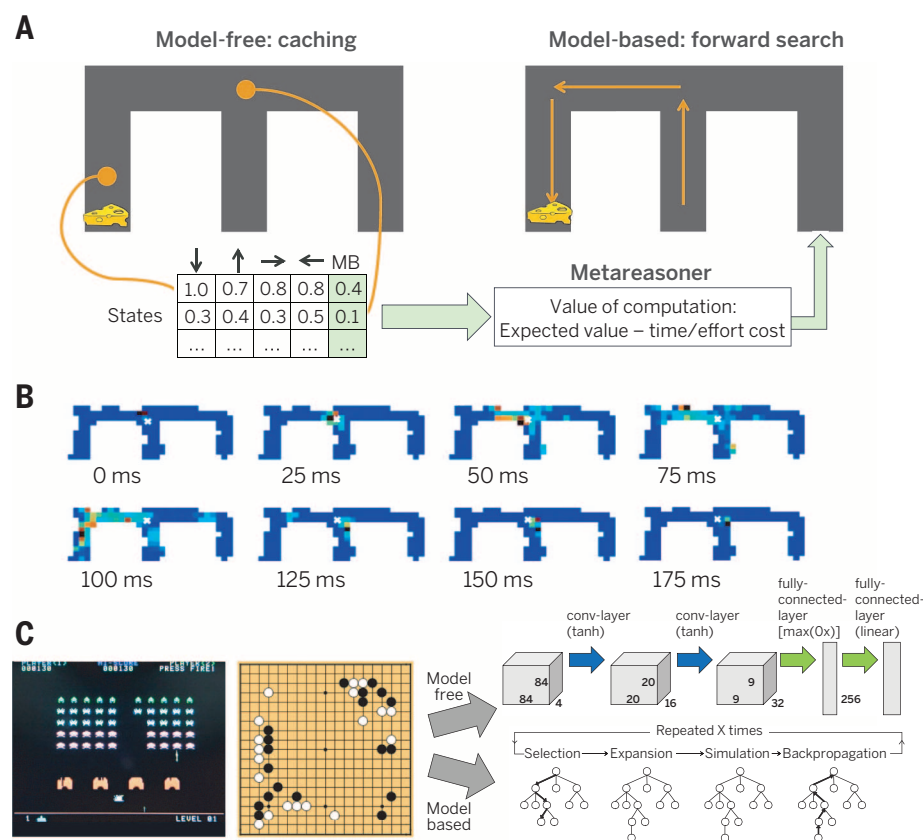


Fig. 4. Computational tradeoffs in use of different decision-making systems. (A) A fast but inflexible model-free system stores cached values in a look-up table but can also learn to invoke a slower but more flexible model-based system that uses forward search to construct an optimal plan. The cached value for invoking the model-based system (highlighted in green) is a simple form of metareasoning that weighs the expected value of forward search against time and effort costs. (B) Hippocampal place cell-firing patterns show the brain engaged in forward search at a choice point, sweeping ahead of the animal's current location. Each image shows the time-indexed representation intensity of locations in pseudocolor (red, high probability; blue, low probability). The representation intensity is computed by decoding spatial location from ensemble recordings in hippocampal area CA3. [Reprinted with permission from (56)] (C) Similar principles apply to more complex decision problems, such as Atari and Go (left). AI systems (right) use complex value function approximation architectures, such as deep convolutional nets (top right) [reprinted with permission from (60)], for model-free control, and sophisticated forward search strategies, such as Monte Carlo Tree Search (bottom right) [reprinted with permission from (61)], for model-based control.

example, reliance on the model-based system decreases when the availability of cognitive resources are transiently disrupted (53). Recent data show that the arbitration mechanism may be supported by the lateral prefrontal cortex (54), the same region involved in the registration of cognitive demand.

Finer-grained metareasoning may play a role within the richer model-based systems themselves. One way to approximate values is to adapt the sampling hypothesis to the sequential decision setting, stochastically exploring trajectories through the state space and using these sample paths to construct a Monte Carlo estimator. Recently, a class of sampling algorithms known as Monte Carlo Tree Search (MCTS) has gained considerable traction on complex problems by balancing exploration and exploitation to determine which trajectories to sample. MCTS has achieved state-of-the-art performance in computer Go as well as a number of other difficult sequential decision problems (55). A recent study analyzed MCTS within a computational rationality framework and showed how simulation decisions can be chosen to optimize the value of computation (20).

There is evidence that the brain might use an algorithm resembling MCTS to solve spatial navigation problems. In the hippocampus, “place cells” respond selectively when an animal is in a particular spatial location and are activated sequentially when an animal considers two different trajectories (Fig. 4B) (56). Pfeiffer and Foster (57) have shown that these sequences predict an animal’s immediate behavior, even for new start and goal locations. It is unknown whether forward sampling observed in place cells balances exploration and exploitation as in MCTS, exploring spatial environments the way MCTS explores game trees, or whether they are sensitive to the value of computation. These are important standing questions in the computational neuroscience of decision-making.

At the same time, AI researchers are beginning to explore powerful interactions between model-based and model-free decision-making systems parallel to the hybrid approaches that computational cognitive neuroscientists have investigated (Fig. 4C). Model-free methods for game-playing based on deep neural networks can, with extensive training, match or exceed model-based MCTS approaches in the regimes that they have been trained on (58). Yet, combinations of MCTS and deep-network approaches beat either approach on its own (59) and may be a promising route to explain how human decision-making in complex sequential tasks can be so accurate and so fast yet still flexible to replan when circumstances change—the essence of acting intelligently in an uncertain world.

Looking forward

Computational rationality offers a potential unifying framework for the study of intelligence in minds, brains, and machines, based on three core ideas: that intelligent agents fundamentally seek to form beliefs and plan actions in

support of maximizing expected utility; that ideal MEU calculations may be intractable for real-world problems, but can be effectively approximated by rational algorithms that maximize a more general expected utility incorporating the costs of computation; and that these algorithms can be rationally adapted to the organism’s specific needs, either offline through engineering or evolutionary design, or online through meta-reasoning mechanisms for selecting the best approximation strategy in a given situation. We discussed case studies in which these ideas are being fruitfully applied across the disciplines of intelligence, but we admit that a genuine unifying theory remains mostly a promise for the future. We see great value in pursuing new studies that seek additional confirmation (or disconfirmation) of the roles of machinery for cost-sensitive computation in human cognition, and for enabling advances in AI. Although we cannot foresee precisely where this road leads, our best guess is that the pursuit itself is a good bet—and as far as we can see, the best bet that we have.

REFERENCES AND NOTES

1. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA, 2009).
2. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann Publishers, Los Altos, CA, 1988).
3. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, Upper Saddle River, NJ, 2009).
4. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univ. Press, Princeton, NJ, 1947).
5. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, *Science* **331**, 1279–1285 (2011).
6. A. M. Turing, *Proc. Lond. Math. Soc.* **2**, 230–265 (1936).
7. A. M. Turing, *Mind* **59**, 433–460 (1950).
8. J. von Neumann, *The Computer and the Brain* (Yale Univ. Press, New Haven, CT, 1958).
9. H. A. Simon, *Models of Man* (Wiley, New York, 1957).
10. I. J. Good, *J. R. Stat. Soc. B* **14**, 107–114 (1952).
11. E. Horvitz, in *Proceedings of the 3rd International Conference on Uncertainty in Artificial Intelligence* (Mountain View, CA, July 1987), pp. 429–444 (1987).
12. S. Russell, E. Wefald, *Artif. Intell.* **49**, 361–395 (1991).
13. E. Horvitz, G. Cooper, D. Heckerman, in *Proceedings of IJCAI*, January 1989, pp. 1121–1127 (1989).
14. E. Horvitz, G. Rutledge, in *Proceedings of the 7th International Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, San Francisco, 1991), pp. 151–158.
15. E. Horvitz, Y. Ruan, G. Gomes, H. Kautz, B. Selman, D. M. Chickering, in *Proceedings of 17th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, San Francisco, 2001), pp. 235–244.
16. E. Horvitz, *Artif. Intell.* **126**, 159–196 (2001).
17. E. Burns, W. Ruml, M. B. Do, *J. Artif. Intell. Res.* **47**, 697–740 (2013).
18. C. H. Lin, A. Kolobov, A. Kamar, E. Horvitz, Metareasoning for planning under uncertainty. In *Proceedings of IJCAI* (2015).
19. T. Dean, L. P. Kaelbling, J. Kirman, A. Nicholson, *Artif. Intell.* **76**, 35–74 (1995).
20. N. Hay, S. Russell, D. Tolpin, S. Shimony, in *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence* (2012), pp. 346–355.
21. D. Heckerman, J. S. Breese, E. Horvitz, in *Proceedings of the 5th Conference on Uncertainty in Artificial Intelligence*, July 1989 (1989), pp. 162–173.
22. G. Cooper, *Artif. Intell.* **42**, 393–405 (1990).
23. C. R. Peterson, L. R. Beach, *Psychol. Bull.* **68**, 29–46 (1967).
24. A. Tversky, D. Kahneman, *Science* **185**, 1124–1131 (1974).
25. G. Gigerenzer, *Rationality for Mortals: How People Cope with Uncertainty* (Oxford Univ. Press, Oxford, 2008).
26. J. R. Anderson, *The Adaptive Character of Thought* (Lawrence Erlbaum, Hillsdale, NJ, 1990).
27. M. Oaksford, N. Chater, *Bayesian Rationality* (Oxford Univ. Press, Oxford, 2007).
28. T. L. Griffiths, J. B. Tenenbaum, *Cognit. Psychol.* **51**, 334–384 (2005).
29. K. Doya, S. Ishii, A. Pouget, R. P. N. Rao, Eds. *The Bayesian Brain: Probabilistic Approaches to Neural Coding* (MIT Press, Cambridge, MA, 2007).
30. T. L. Griffiths, F. Lieder, N. D. Goodman, *Top. Cogn. Sci.* **7**, 217–229 (2015).
31. S. Denison, E. Bonawitz, A. Gopnik, T. L. Griffiths, *Cognition* **126**, 285–300 (2013).
32. E. Vul, M. Frank, G. Alvarez, J. B. Tenenbaum, *Adv. Neural Inf. Process. Syst.* **29**, 1955–1963 (2009).
33. S. J. Gershman, E. Vul, J. B. Tenenbaum, *Neural Comput.* **24**, 1–24 (2012).
34. A. N. Sanborn, T. L. Griffiths, D. J. Navarro, *Psychol. Rev.* **117**, 1144–1167 (2010).
35. L. Buesing, J. Bill, B. Nessler, W. Maass, *PLOS Comput. Biol.* **7**, e1002211 (2011).
36. M. Isard, A. Blake, *Int. J. Comput. Vis.* **29**, 5–28 (1998).
37. R. Levy, F. Real, T. L. Griffiths, *Adv. Neural Inf. Process. Syst.* **21**, 937–944 (2009).
38. E. Vul, N. Goodman, T. L. Griffiths, J. B. Tenenbaum, *Cogn. Sci.* **38**, 599–637 (2014).
39. W. Kool, J. T. McGuire, Z. B. Rosen, M. M. Botvinick, *J. Exp. Psychol. Gen.* **139**, 665–682 (2010).
40. W. Kool, M. Botvinick, *J. Exp. Psychol. Gen.* **143**, 131–141 (2014).
41. J. T. McGuire, M. M. Botvinick, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7922–7926 (2010).
42. F. Lieder et al., *Adv. Neural Inf. Process. Syst.* **27**, 2870–2878 (2014).
43. R. L. Lewis, A. Howes, S. Singh, *Top. Cogn. Sci.* **6**, 279–311 (2014).
44. J. W. Payne, J. R. Bettman, E. J. Johnson, *J. Exp. Psychol. Learn. Mem. Cogn.* **14**, 534–552 (1988).
45. J. Rieskamp, P. E. Otto, *J. Exp. Psychol. Gen.* **135**, 207–236 (2006).
46. F. Lieder, M. Hsu, T. L. Griffiths, in *Proc. 36th Ann. Conf. Cognitive Science Society* (Austin, TX, 2014).
47. N. D. Daw, Y. Niv, P. Dayan, *Nat. Neurosci.* **8**, 1704–1711 (2005).
48. S. Killcross, E. Coutureau, *Cereb. Cortex* **13**, 400–408 (2003).
49. H. H. Yin, B. J. Knowlton, B. W. Balleine, *Eur. J. Neurosci.* **19**, 181–189 (2004).
50. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, *Neuron* **69**, 1204–1215 (2011).
51. M. Keramati, A. Dezfouli, P. Piray, *PLOS Comput. Biol.* **7**, e1002055 (2011).
52. A. Dickinson, *Philos. Trans. R. Soc. London B Biol. Sci.* **308**, 67–78 (1985).
53. A. R. Otto, S. J. Gershman, A. B. Markman, N. D. Daw, *Psychol. Sci.* **24**, 751–761 (2013).
54. S. W. Lee, S. Shimojo, J. P. O’Doherty, *Neuron* **81**, 687–699 (2014).
55. S. Gelly et al., *Commun. ACM* **55**, 106–113 (2012).
56. A. Johnson, A. D. Redish, *J. Neurosci.* **27**, 12176–12189 (2007).
57. B. E. Pfeiffer, D. J. Foster, *Nature* **497**, 74–79 (2013).
58. V. Mnih et al., *Nature* **518**, 529–533 (2015).
59. C. J. Maddison, A. Huang, I. Sutskever, D. Silver, <http://arxiv.org/abs/1412.6564> (2014).
60. X. Guo, S. Singh, H. Lee, R. Lewis, X. Wang, *Adv. Neural Inf. Process. Syst.* **27**, 3338–3346 (2014).
61. G. M. J.-B. Chaslot, S. Bakkes, I. Szita, P. Spronck, in *Proc. Artif. Intell. Interact. Digit. Entertain. Conf.* (Stanford, CA, 2008), pp. 216–217.

ACKNOWLEDGMENTS

We are grateful to A. Gershman and the three referees for helpful comments. This research was partly supported by the Center for Brains, Minds and Machines (CBMM), funded by National Science Foundation Science and Technology Center award CCF-1231216.

10.1126/science.aac6076

RESEARCH

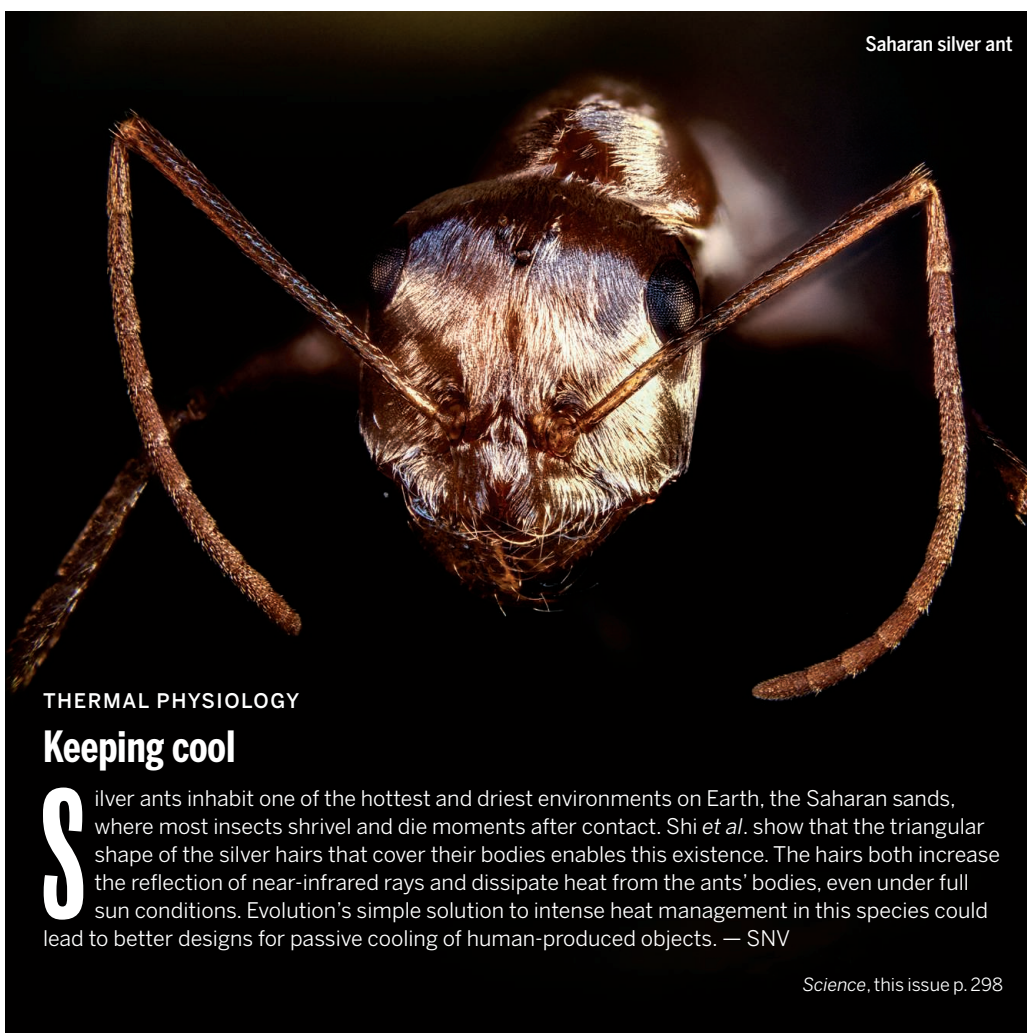
Polar bears cannot resist summer food shortages

Whiteman et al., p. 295



IN SCIENCE JOURNALS

Edited by Stella Hurtley



Saharan silver ant

THERMAL PHYSIOLOGY

Keeping cool

Silver ants inhabit one of the hottest and driest environments on Earth, the Saharan sands, where most insects shrivel and die moments after contact. Shi *et al.* show that the triangular shape of the silver hairs that cover their bodies enables this existence. The hairs both increase the reflection of near-infrared rays and dissipate heat from the ants' bodies, even under full sun conditions. Evolution's simple solution to intense heat management in this species could lead to better designs for passive cooling of human-produced objects. — SNV

Science, this issue p. 298

HIV-1 VACCINES

To defeat SIV, add a protein boost

Despite 30 years of effort, no HIV-1 vaccine exists. Barouch *et al.* evaluated one promising strategy in rhesus macaques, a preclinical model commonly used to test potential HIV-1 vaccine candidates. They immunized monkeys with adenovirus-36 vectors engineered

to express SIV (simian immunodeficiency virus) genes and then boosted them with a recombinant gp120 envelope glycoprotein (Env) from SIV. This regimen afforded greater protection than a strategy that instead used a viral vector-based boost. A parallel trial using a SHIV (simian/human immunodeficiency virus)-based vaccine and challenge model produced similar results. Whether this particular

approach will be equally successful in humans remains to be tested. — KLM

Science, this issue p. 320

MAGNETISM

Skyrmions emerge in trilayers

Skyrmions are tiny whirlpools of magnetic spin with potential to act as carriers of information in

future devices. Skyrmions have been observed in multiple materials but usually at impractically low temperatures. Jiang *et al.* used a constriction in a trilayer system to create skyrmions at room temperature (see the Perspective by von Bergmann). The authors pushed elongated magnetic domains through the constriction using an in-plane current, causing individual skyrmion bubbles to form. — JS

Science, this issue p. 283; see also p. 234

NANOPARTICLE IMAGING

Looking at teeny tiny platinum particles

Electron microscopy is a powerful technique for taking snapshots of particles or images at near-atomic resolution. Park *et al.* studied free-floating platinum nanoparticles using electron microscopy and liquid cells (see the Perspective by Colliex). Using analytical techniques developed to study biological molecules, they reconstructed the three-dimensional features of the Pt particles at near-atomic resolution. This approach has the scope to study a mixed population of particles one at a time and to study their synthesis as it occurs in solution. — MSL

Science, this issue p. 290; see also p. 232

PLANT SCIENCE

Substrate channeling in morphine biosynthesis

Poppies are still the most economically viable source of the excellent painkiller morphine. Winzer *et al.* have now identified a key enzyme in the

poppy's biosynthetic pathway for morphine. The enzyme turns out to be an unusual protein that contains both cytochrome P-450 and oxidoreductase modules. Together these modules process two subsequent steps in the biosynthetic pathway. The identification of this enzyme may enable alternate routes for morphine biosynthesis that are less dependent on poppy cultivation. — PJH

Science, this issue p. 309

SEX DETERMINATION

How germ cells become sperm or egg

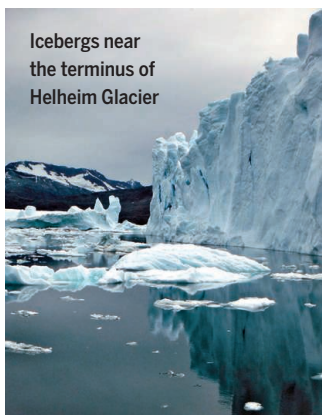
During vertebrate development, germ cells switch from a sexually indifferent to a committed state for either egg or sperm. Signals from somatic gonadal cells are generally thought to influence the sexual differentiation of germ cells. However, Nishimura *et al.* demonstrate that germ cell–intrinsic sex determination cues are at play in the teleost fish medaka. The forkhead box transcriptional factor foxl3 represses the initiation of spermatogenesis. In the absence of foxl3 function, females develop ovaries filled with functional sperm. Thus, the male gonad environment is not required for spermatogenesis. —BAP

Science, this issue p. 328

ICE SHEETS

Movers and shakers

When the edge of an ice sheet breaks off and falls into the sea (calves), the remaining section



Icebergs near the terminus of Helheim Glacier

of the ice sheet moves backward and down and can suffer a glacial earthquake. Murray *et al.* studied calving from Greenland's Helheim Glacier. The forces that cause the change in the motion of the ice sheet at its terminus also trigger the accompanying earthquakes. Because these seismic signals can be detected by instruments located all over the globe, it should be possible to use these glacial earthquakes as proxies for glacier calving. — HJS

Science, this issue p. 305

DRUG DISCOVERY

Long-acting drug to treat resistant malaria

Malaria kills 0.6 million people annually. Currently available drugs are no longer fully effective, because the malarial parasite has developed resistance. Now, Phillips *et al.* have identified a drug, DSM265, that kills both drug-sensitive and drug-resistant parasites by targeting their ability to synthesize precursors required for synthesis of DNA and RNA. DSM265 kills parasites in the blood and the liver and is sufficiently long-acting that it could potentially cure malaria after a single dose or provide effective chemoprevention if given weekly. — OMS

Sci. Transl. Med. **7**, 296ra111 (2015).

INFECTIOUS DISEASE

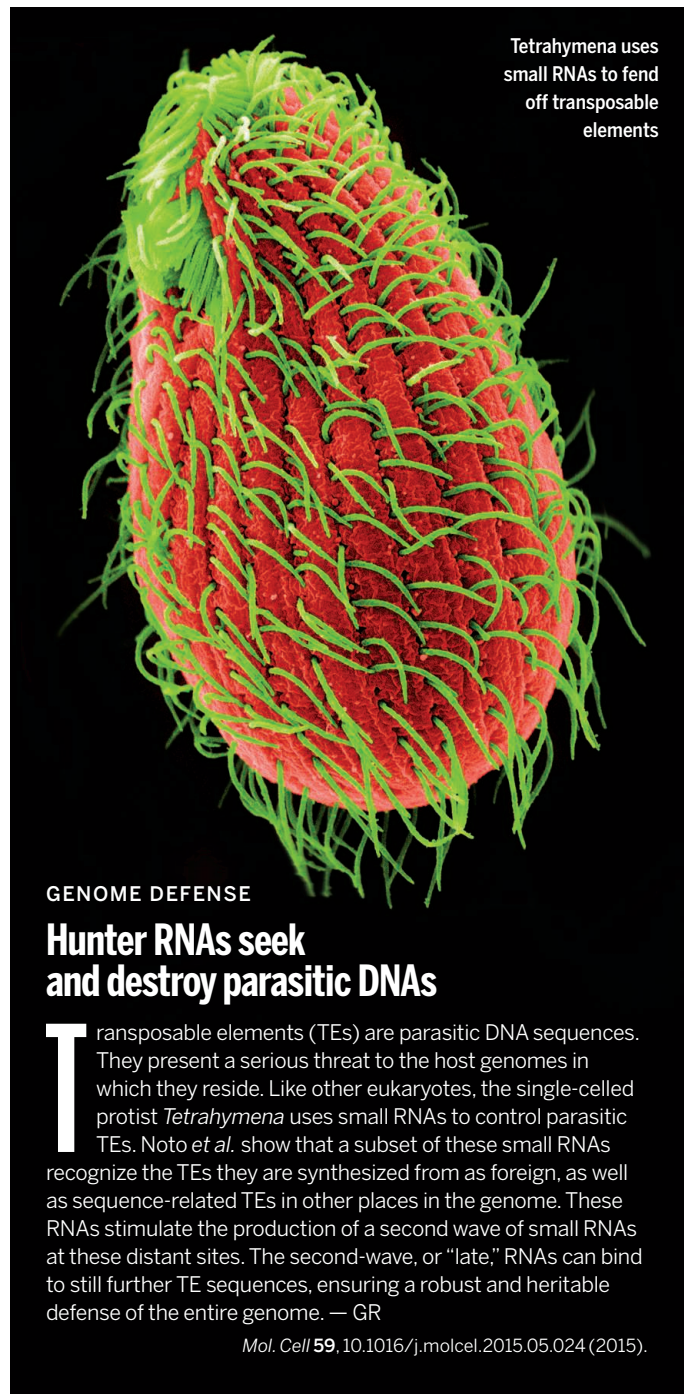
GBS toxin activates mast cells for host defense

Ascending Group B streptococcus (GBS) is a major cause of preterm birth. How the mother defends against GBS infections is not clear. Working in mice, Rajagopal *et al.* found that immune cells, called mast cells, are activated by a lipid toxin produced by GBS. This lipid toxin is an ornithine rhamno-polyene that stains the bacteria red. The toxin assists penetration of the placenta by the bacteria. The toxin also stimulates mast cell degranulation, an early step in host efforts to defeat the bacterial infection. — PLY

Sci. Adv. **10**, 1126/sciadv.1400225 (2015).

IN OTHER JOURNALS

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



Tetrahymena uses small RNAs to fend off transposable elements

GENOME DEFENSE

Hunter RNAs seek and destroy parasitic DNAs

Transposable elements (TEs) are parasitic DNA sequences. They present a serious threat to the host genomes in which they reside. Like other eukaryotes, the single-celled protist *Tetrahymena* uses small RNAs to control parasitic TEs. Noto *et al.* show that a subset of these small RNAs recognize the TEs they are synthesized from as foreign, as well as sequence-related TEs in other places in the genome. These RNAs stimulate the production of a second wave of small RNAs at these distant sites. The second-wave, or "late," RNAs can bind to still further TE sequences, ensuring a robust and heritable defense of the entire genome. — GR

Mol. Cell **59**, 10.1016/j.molcel.2015.05.024 (2015).

ASTROPHYSICS

X-ray echoes used as a cosmic yardstick

Determining accurate distances to astronomical x-ray sources is notoriously difficult, but Heinz *et al.* have developed a new technique for doing so.

They exploited "light echoes" from the object Cir X-1, in which x-rays emitted during a large flare bounced off foreground dust and appeared months later as delayed rings of emission around the source. By comparing the x-ray rings to radio data, they were able to identify the

foreground clouds of dust and thereby measure the distance to Cir X-1 with an accuracy of 10%. Previous estimates had varied by almost a factor of 3. — KTS
Astrophys. J. **806**, 265 (2015).

PHYSICS

A gap in a topological surface state

Topological insulators (TIs) have, in theory, conducting surfaces and insulating bulks. In practice, TIs often have a considerable bulk conductance that masks the conduction from the topologically protected surface states. To make the surface state nonconducting—in other words, to open a gap in its dispersion—researchers usually resort to magnetic doping. Now, Weber *et al.* demonstrate that a topological surface state of the compound Bi_4Se_3 is naturally, although partially, gapped. To show this, the authors use a combination of photoemission measurements and calculations. The results may be more general within the subclass of TIs to which Bi_4Se_3 belongs. — JS

Phys. Rev. Lett. 10.1103/PhysRevLett.114.256401 (2015).

PROTEIN FOLDING

Trapped on the wrong pathway

Protein folding can be described as diffusion over an energy landscape in conformational space to reach an energy minimum that represents a stable folded structure, and it is implicated in many diseases. A particularly dramatic example is the prion protein, PrP, that folds rapidly into a native monomeric structure but also has a stable oligomeric form that causes prion disease. Yu *et al.* used single-molecule force spectroscopy to monitor misfolding of a PrP dimer. The dimer misfolds along a single pathway involving several intermediates, one of which blocks native folding. Diffusion across the energy landscape is 1000 times slower than for the folding of native

prion monomers, probably indicating that many unproductive interactions occur during misfolding. — VV

Proc. Natl. Acad. Sci. U.S.A. 10.1073/pnas.1419197112 (2015).

GLOBAL CARBON CYCLE

Seeing the forest from the trees

Boreal forests contain nearly half of the carbon stored by the trees of the world, and thus are a critical component of the global carbon cycle. How much carbon is contained in this part of the ecosystem still is inadequately known, though. Chen and Luo studied the net annual above-ground biomass change of four

major boreal forest types in western Canada over the period between 1958 and 2011 and found that it declined for all of those groups. They attribute the loss to increased tree mortality and reduced growth, caused mostly by persistent warming and decreasing water availability, trends that are expected to escalate in the future. — HJS

Global Change Biol. 10.1111/gcb.12994 (2015).

NEUROSCIENCE

Contextual memory networks in monkey brains

Memory tasks activate a number of areas in the prefrontal cortex of the brain. Lesions in this region

cause memory deficits; however, lesions in only a fraction of the activated areas actually lead to severe memory loss. Osada *et al.* measured whole-brain activity during a memory task in monkeys. The activated areas and their task-specific functional connectivity formed a hierarchical network centered on a hub. This functional hub largely matched the documented sites where lesions had the most dramatic effect. Neighboring sites where lesions had no impact were much less affected. The functional connections of an area predicted the degree of memory loss much better than its role in the anatomical network. — PRS

PLOS Biol. 10.1371/journal.pbio.1002177 (2015).



Angiopoietin-1 expression varies with brain size in guppies

BRAIN EVOLUTION

The genetic underpinnings of brain size

What genes and selective processes lead to an increase in brain size and intelligence is one of the greatest questions in human evolution. Chen *et al.* present findings in fish that may help us understand this process. When guppies were selected for brain size, expression differences were observed in the *Angiopoietin-1* (*Ang-1*) gene. These differences seem to be due to noncoding variation; furthermore, lowering *Ang-1* expression in the zebrafish also affected brain size. On the basis of these results, the authors suggest that further study of *Ang-1* may help provide insight into the evolution of brain size and cognition in other species, including humans. — LMZ

Proc. R. Soc. London Ser. B 10.1098/rspb.2015.0872 (2015).

ALSO IN SCIENCE JOURNALS

Edited by Stella Hurtley

CHROMOSOMES

Protein partners for chromosome silencing

Female mammals have two X chromosomes, one of which is almost completely shut down during development. The long noncoding Xist RNA plays a role in this process. To understand how a whole chromosome can be stably inactivated, Minajigi *et al.* identified many of the proteins that bind to the Xist RNA, which include cohesins. Paradoxically, the interaction between Xist and cohesin subunits resulted in repulsion of cohesin complexes from the inactive X chromosome, changing the three-dimensional shape of the whole chromosome. — GR

Science, this issue p. 282

PLANT ECOLOGY

Grassland diversity and ecosystem productivity

The relationship between plant species diversity and ecosystem productivity is controversial. The debate concerns whether diversity peaks at intermediate levels of productivity—the so-called humped-back model—or whether there is no clear predictable relationship. Fraser *et al.* used a large, standardized, and geographically diverse sample of grasslands from six continents to confirm the validity and generality of the humped-back model. Their findings pave the way for a more mechanistic understanding of the factors controlling species diversity. — AMS

Science, this issue p. 302

HEAVY FERMIONS

Probing the insulating state of SmB_6

When a metal is subjected to a strong magnetic field, its

electrons start rearranging into new energy levels, causing its electronic properties to oscillate as a function of the field. Unexpectedly, Tan *et al.* observed this phenomenon, called quantum oscillations, in the Kondo insulator samarium hexaboride (SmB_6), which does not conduct electricity. They measured the magnetic torque and detected quantum oscillations originating from the bulk of this heavy fermion compound. These oscillations had an unusual temperature dependence, which presents another puzzle to theorists seeking to understand the nature of the insulating state of SmB_6 . — JS

Science, this issue p. 287

ANIMAL PHYSIOLOGY

Not that unusual after all

As polar ice recedes, polar bears are facing a changed habitat with reduced summer foraging opportunities. It has been hypothesized that they might be able to resist summer food shortages by reducing their metabolic needs in a sort of “walking hibernation.” Whiteman *et al.* monitored energy expenditure in polar bears both on and off the ice and found energy reductions, but that these were more akin to normal mammalian fasting levels. Thus, it appears that polar bears have no energetic protections against reduced summer food supplies and will face increasing starvation threats if summer foraging habitats continue to decline. — SNV

Science, this issue p. 295

CIRCADIAN RHYTHMS

Biochemical basis of a 24-hour clock

Circadian clocks keep organisms in synch with such daily cycles

as illumination, activity, and food availability. The circadian clock in cyanobacteria has the necessary 24-hour period despite its three component proteins having biochemical activities that occur on a much faster time scale. Abe *et al.* focused on the cyanobacterial clock component KaiC, an adenosine triphosphatase (ATPase) that can autophosphorylate and autodephosphorylate. The slow ATPase activity of KaiC, which is linked to a peptide isomerisation, provided the slow kinetics that set the speed of the 24-hour clock. Chang *et al.* found that another clock component, KaiB, also has slow changes in its protein conformation that help to set the oscillation period of the clock and its signaling output. — LBR

Science, this issue pp. 312 and 324

ECOLOGY

The benefits of diversity

Pathogens and parasites are an integral part of all ecosystems. But the likelihood of them causing disease depends on environmental factors. In a Perspective, Keesing and Ostfeld highlight recent studies on disease prevalence as a function of diversity. In humans, other animals, and plants, disease prevalence is lower when diversity is high. This is the case even though pathogen diversity is higher in more-diverse systems. As biodiversity is lost from ecosystems, they become more vulnerable to infections. — JFU

Science, this issue p. 235

INFLAMMATION

Neutrophil NETs drive atherosclerosis

The buildup of fats, cholesterol, and other substances in arteries causes atherosclerosis, which restricts blood flow and can

lead to heart attacks and stroke. Inflammation contributes to the pathogenesis of atherosclerosis, but exactly how is not fully understood. Warnatsch *et al.* now show that immune cells called neutrophils release NETs (neutrophil extracellular traps) (see the Perspective by Nahrendorf and Swirski). These NETs are composed of DNA and antimicrobial proteins, and in the setting of atherosclerosis they activate innate immune signaling pathways in macrophages. This causes the macrophages to secrete proinflammatory cytokines, exacerbating the disease. Indirectly, NETs also attract a specialized subset of T cells that further amplify the proinflammatory response. —KLM

Science, this issue p. 316; see also p. 237

IMMUNOLOGY

Fine-tuning the inflammatory response

The binding of the bacterial product lipopolysaccharide (LPS) to the receptor TLR4 on macrophages triggers inflammatory responses that require the transcription factor NF- κ B. TLR4 recruits the adaptor protein MyD88 when the receptor is at the plasma membrane and a different adaptor protein, TRIF, after internalization into endosomes. Cheng *et al.* found that MyD88 was required for the initial peak of transient NF- κ B activation in all LPS-stimulated cells (see also the Focus by Williams *et al.*). In contrast, TRIF was required for more sustained NF- κ B activation in a subset of cells. Thus, macrophages use both adaptor proteins to fine-tune NF- κ B activation to induce an appropriate inflammatory response. — JFF

Sci. Signal. **8**, ra69 and fs13 (2014).

RESEARCH ARTICLE SUMMARY

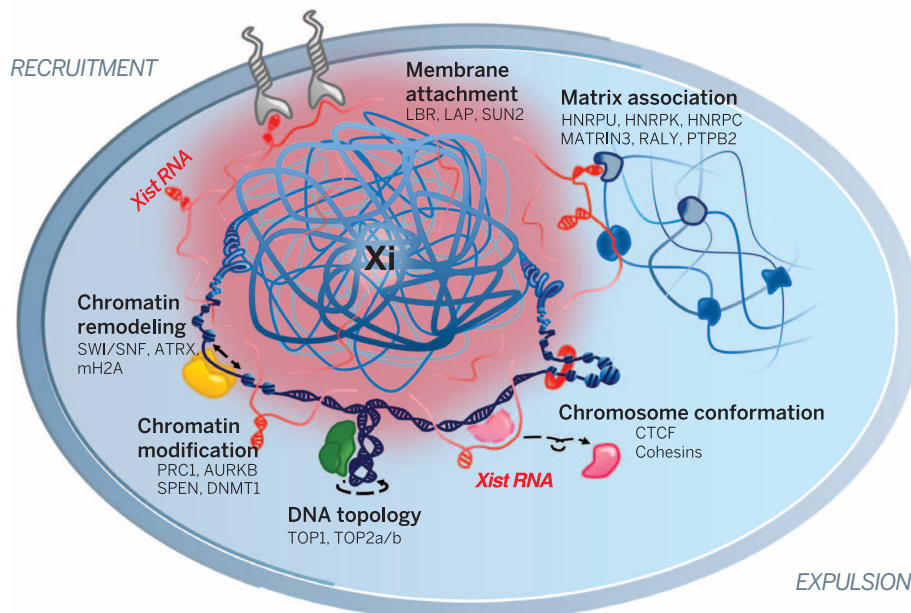
CHROMOSOMES

A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation

Anand Minajigi,* John E. Froberg,* Chunyao Wei, Hongjae Sunwoo, Barry Kesner, David Colognori, Derek Lessing, Bernhard Payer, Myriam Boukhali, Wilhelm Haas, Jeannie T. Lee†

INTRODUCTION: The mammal has evolved an epigenetic mechanism to silence one of two X chromosomes in the XX female to equalize gene dosages with the XY male. Once established, the inactivated X chromosome (Xi) is extremely stable and is maintained through the lifetime of the female mammal. The principal regulator, Xist, is a long noncoding RNA that orchestrates the silencing process along the Xi. Xist is believed to operate as a scaffold to recruit and spread repressive complexes, such as Polycomb Repressive Complex 2, along the X chromosome. The identities of crucial interacting factors, however, have remained largely unknown.

RATIONALE: Although the Xi's epigenetic stability is a necessary homeostatic property, an ability to unlock this epigenetic state is of great current interest. The X chromosome is home to nearly 1000 genes, at least 50 of which have been implicated in X-linked diseases, such as Rett syndrome and fragile X syndrome. The Xi is therefore a reservoir of functional genes that could be tapped to replace expression of a disease allele on the active X (Xa). A major gap in current understanding is the lack of a comprehensive Xist interactome. Progress toward a full interactome would advance knowledge of epigenetic regulation by long noncoding RNA and potentially inform treatment of X-linked diseases.



An operational model for how Xist RNA orchestrates the Xi state. Xist is a multitasking RNA that brings many layers of repression to the Xi. Although Xist RNA recruits repressive complexes (such as PRC1, PRC2, DNMT1, macroH2A, and SmcHD1) to establish and maintain the inactive state, it also actively repels activating factors and architectural proteins (such as the cohesins and CTCF) to avoid acquisition of a transcription-favorable chromatin conformation.

RESULTS: We have developed an RNA-centric proteomic method called iDRiP (identification of direct RNA-interacting proteins).

Using iDRiP, we identified 80 to 200 proteins in the Xist interactome. The interactors fall into several functional categories, including cohesins, condensins, topoisomerases, RNA helicases, chromatin remodelers, histone modifiers, DNA methyltransferases, nucleoskeletal factors, and nuclear matrix proteins. Targeted inhibition demonstrates that Xi silencing can be destabilized by disrupting multiple components of the interactome, consistent with the idea that these factors synergistically repress Xi transcription. Triple-drug treatments lead to a net increase of Xi expression and up-regulation of ~100 to 200 Xi genes. We then carry out a focused study of X-linked cohesin sites. Chromatin immunoprecipitation sequencing analysis demonstrates three types of cohesin sites on the X chromosome: Xi-specific sites, Xa-specific sites, and biallelic sites. We find that the Xa-specific binding sites represent a default state. Ablating Xist results in restoration of Xa-specific sites on the Xi. These findings demonstrate that, while Xist attracts repressive complexes to the Xi, it actively repels chromosomal architectural factors such as the cohesins from the Xi. Finally, we examine how Xist and the repulsion of cohesins affect Xi chromosome structure. In wild-type cells, the Xa is characterized by ~112 topologically associated domains (TADs) and the Xi by two megadomains. Intriguingly, loss of Xist and restoration of cohesin binding result in a reversion of the Xi to an Xa-like chromosome conformation. Hi-C analysis shows that TADs return to the Xi in a manner correlated with the reappearance of cohesins and with a transcriptionally permissive state.

ON OUR WEB SITE

Read the full article at <http://dx.doi.org/10.1126/science.aab2276>

CONCLUSION: Our study unveils many layers of Xi repression and demonstrates a central role for RNA in the topological organization of mammalian chromosomes. Our study also supports a model in which Xist RNA simultaneously acts as (i) a scaffold for the recruitment of repressive complexes to establish and maintain the inactive state and (ii) a repulsion mechanism to extrude architectural factors such as cohesins to avoid acquisition of a transcription-favorable chromatin conformation. Finally, our findings indicate that the stability of the Xi can be perturbed by targeted inhibition of multiple components of the Xist interactome. ■

CONCLUSION: Our study unveils many layers of Xi repression and demonstrates a central role for RNA in the topological organization of mammalian chromosomes. Our study also supports a model in which Xist RNA simultaneously acts as (i) a scaffold for the recruitment of repressive complexes to establish and maintain the inactive state and (ii) a repulsion mechanism to extrude architectural factors such as cohesins to avoid acquisition of a transcription-favorable chromatin conformation. Finally, our findings indicate that the stability of the Xi can be perturbed by targeted inhibition of multiple components of the Xist interactome. ■

The list of author affiliations is available in the full article online.

*These authors contributed equally to this work.

†Corresponding author: lee@molbio.mgh.harvard.edu
Cite this article as A. M. Minajigi et al., *Science* 349, aab2276 (2015). DOI: 10.1126/science. aab2276

RESEARCH ARTICLE

CHROMOSOMES

A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation

Anand Minajigi,^{1*} John E. Froberg,^{1*} Chunyao Wei,¹ Hongjae Sunwoo,¹ Barry Kesner,¹ David Colognori,¹ Derek Lessing,¹ Bernhard Payer,^{1†} Myriam Boukhali,² Wilhelm Haas,² Jeannie T. Lee^{1‡}

The inactive X chromosome (Xi) serves as a model to understand gene silencing on a global scale. Here, we perform “identification of direct RNA interacting proteins” (iDRiP) to isolate a comprehensive protein interactome for Xist, an RNA required for Xi silencing. We discover multiple classes of interactors—including cohesins, condensins, topoisomerases, RNA helicases, chromatin remodelers, and modifiers—that synergistically repress Xi transcription. Inhibiting two or three interactors destabilizes silencing. Although Xist attracts some interactors, it repels architectural factors. Xist evicts cohesins from the Xi and directs an Xi-specific chromosome conformation. Upon deleting *Xist*, the Xi acquires the cohesin-binding and chromosomal architecture of the active X. Our study unveils many layers of Xi repression and demonstrates a central role for RNA in the topological organization of mammalian chromosomes.

The mammalian X chromosome is unique in its ability to undergo whole-chromosome silencing. In the early female embryo, X-chromosome inactivation (XCI) enables mammals to achieve gene dosage equivalence between the XX female and the XY male (1–3). XCI depends on Xist RNA, a 17-kb long noncoding RNA (lncRNA) expressed only from the inactive X chromosome (Xi) (4) and that implements silencing by recruiting repressive complexes (5–8). Whereas XCI initiates only once during development, the female mammal stably maintains the Xi through her lifetime. In mice, a germline deletion of *Xist* results in peri-implantation lethality due to a failure of Xi establishment (9), whereas a lineage-specific deletion of *Xist* causes a lethal blood cancer due to a failure of Xi maintenance (10). Thus, both the de novo establishment and proper maintenance of the Xi are crucial for viability and homeostasis. There are therefore two critical phases of XCI: (i) A one-time initiation phase in peri-implantation embryonic development that is recapitulated by differentiating embryonic stem (ES) cells in culture, and (ii) a lifelong maintenance phase that persists in all somatic lineages.

Once established, the Xi is extremely stable and difficult to disrupt genetically and pharma-

cologically (11–13). In mice, X reactivation is programmed to occur only twice: once in the blastocyst to erase the imprinted XCI pattern and a second time in the germ line before meiosis (14, 15). Although the Xi's epigenetic stability is a homeostatic asset, an ability to unlock this epigenetic state is of great current interest. The X chromosome is home to nearly 1000 genes, at least 50 of which have been implicated in X-linked diseases, such as Rett syndrome and fragile X syndrome. The Xi is therefore a reservoir of functional genes that could be tapped to replace expression of a disease allele on the active X (Xa). A better understanding of Xi repression would inform both basic biological mechanisms and treatment of X-linked diseases.

It is believed that Xist RNA silences the Xi through conjugate protein partners. A major gap in current understanding is the lack of a comprehensive Xist interactome. Despite multiple attempts to define the complete interactome, only four directly interacting partners have been identified over the past two decades, including PRC2, ATRX, YY1, and HNRPU: Polycomb repressive complex 2 (PRC2) is targeted by Xist RNA to the Xi; the ATRX RNA helicase is required for the specific association between Xist and PRC2 (16, 17); YY1 tethers the Xist-PRC2 complex to the Xi nucleation center (18); and the nuclear matrix factor, HNRPU/SAF-A, enables stable association of Xist with the chromosomal territory (19). Many additional interacting partners are expected, given the large size of Xist RNA and its numerous conserved modular domains. We have developed an RNA-based proteomic method and implement an unbiased screen for Xist's comprehensive interactome. We identify a large number of high-

confidence candidates, demonstrate that it is possible to destabilize Xi repression by inhibiting multiple interacting components, and then delve into a focused set of interactors with the cohesins.

Results

iDRiP identifies multiple classes of Xist-interacting proteins

A systematic identification of interacting factors has been challenging because of Xist's large size, the expected complexity of the interactome, and the persistent problem of high background with existing biochemical approaches (20). A high background could be particularly problematic for chemical cross-linkers that create extensive covalent networks of proteins, which could in turn mask specific and direct interactions. We therefore developed iDRiP (identification of direct RNA interacting proteins) using the zero-length cross-linker, ultraviolet (UV) light, to implement an unbiased screen of directly interacting proteins in female mouse fibroblasts expressing physiological levels of Xist RNA (Fig. 1A). We performed in vivo UV cross-linking, prepared nuclei, and solubilized chromatin by DNase I digestion. Xist-specific complexes were captured using nine complementary oligonucleotide probes spaced across the 17-kb RNA, with a 25-nucleotide probe length designed to maximize RNA capture while reducing nonspecific hybridization. The complexes were washed under denaturing conditions to eliminate factors not covalently linked by UV to Xist RNA. To minimize background due to DNA-bound proteins, a key step was inclusion of DNase I treatment before elution of complexes (see the supplementary text). We observed substantial enrichment of Xist RNA over highly abundant cytoplasmic and nuclear RNAs (U6, Jpx, and 18S ribosomal RNA) in eluates of female fibroblasts (Fig. 1B). Enrichment was not observed in male eluates or with luciferase capture probes. Eluted proteins were subjected to quantitative mass spectrometry (MS), with spectral counting (21) and multiplexed quantitative proteomics (22) yielding similar enrichment sets (table S1).

From three independent replicates, iDRiP-MS revealed a large Xist protein interactome (Fig. 1C and table S1). Recovery of known Xist interactors, PRC2, ATRX, and HNRPU, provided a first validation of the iDRiP technique. Also recovered were macrohistone H2A (mH2A), RING1 (PRC1), and the condensin component, SmcHD1—proteins known to be enriched on the Xi (19, 23, 24) but not previously shown to interact directly with Xist. More than 80 proteins were found to be ≥threefold enriched over background; >200 proteins were ≥twofold enriched (table S1). In many cases, multiple subunits of the epigenetic complex were identified, boosting our confidence in them as interactors. We verified select interactions by performing a test of reciprocity: By baiting with candidate proteins in an antibody capture, RIP-qPCR (RNA immunoprecipitation–quantitative polymerase chain reaction) of UV-cross-linked cells reciprocally identified Xist RNA in the pulldowns (Fig. 1D). Called on the basis of high enrichment

¹Howard Hughes Medical Institute; Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA; Department of Genetics, Harvard Medical School, Boston, MA, USA. ²Massachusetts General Hospital Cancer Center, Charlestown, Boston, MA; Department of Medicine, Harvard Medical School, Boston, MA, USA.

*These authors contributed equally to this work. †Present address: Centre for Genomic Regulation, Barcelona, Spain. ‡Corresponding author: lee@molbio.mgh.harvard.edu

values, presence of multiple subunits within a candidate epigenetic complex, and tests of reciprocity, novel high-confidence interactors fell into several functional categories: (i) Cohesin complex proteins SMC1a, SMC3, RAD21, WAPL, and PDS5a/b, as well as CCCTC-binding factor (CTCF) (25), which are collectively implicated in chromosome looping (26–28); (ii) histone modifiers such as aurora kinase B (AURKB), a serine/threonine kinase that phosphorylates histone H3 (29); RING1, the catalytic subunit of polycomb repressive complex 1 (PRC1) for H2A-K119 ubiquitylation (23); SPEN and RBM15, which associate with histone deacetylases; (iii) switch/sucrose non-fermentable (SWI/SNF) chromatin remodeling factors; (iv) topoisomerases TOP2a, TOP2b, and TOP1, which relieve torsional stress during transcription and DNA replication; (v) miscellaneous transcriptional regulators like MYEF2 and ELAV1; (vi) nucleoskeletal proteins that anchor chromosomes to the nuclear envelope, SUN2, lamin-B receptor (LBR), and LAP2; (vii) nuclear matrix proteins hnRPU/SAF-A, hnRKP, and MATRIN3; and (viii) the DNA methyltransferase DNMT1, known as a maintenance methylase for CpG dinucleotides (30).

To study their function, we first performed RNA immunofluorescence in situ hybridization (immunofISH) of female cells and observed several patterns of Xi coverage relative to the surrounding nucleoplasm (Fig. 1E). Like PRC2, RING1 (PRC1) has been shown to be enriched on the Xi (23) and is therefore not pursued further. TOP1 and TOP2a/b appeared neither enriched nor depleted on the Xi (100%, $n > 50$ nuclei). AURKB showed two patterns of localization: pericentric enrichment (20%, $n > 50$) and a more diffuse localization pattern (80%, data not shown), consistent with its cell-cycle-dependent chromosomal localization (29). On the other hand, whereas SUN2 was depleted on the Xi (100%, $n = 52$), it often appeared as pinpoints around the Xi in both day 7 differentiating female ES cells (establishment phase; 44%, $n = 307$) and in fibroblasts (maintenance phase; 38.5%, $n = 52$), consistent with SUN2's function in tethering telomeres to the nuclear envelope. Finally, the cohesins and SWI/SNF remodelers unexpectedly showed a depletion relative to the surrounding nucleoplasm (100%, $n = 50$ to 100). These patterns suggest that the Xist interactors operate in different XCI pathways.

To ask if the factors intersect the PRC2 pathway, we stably knocked down (KD) top candidates using short-hairpin RNAs (shRNAs) (table S2) and performed RNA immunofISH to examine trimethylation of histone H3-lysine 27 (H3K27me3) (Fig. 2, A and B). No major changes to Xist localization or H3K27me3 were evident in d7 ES cells (fig. S1). There were, however, long-term effects in fibroblasts: The decrease in H3K27me3 enrichment in shSMARCC1 and shSMARCA5 cells (Fig. 2, A and B) indicated that SWI/SNF interaction with Xist is required for proper maintenance of PRC2 function on the Xi. Steady-state Xist levels did not change by more than twofold (Fig. 2C) and were therefore unlikely to be the cause of the polycomb defect. Knockdowns of other factors (cohesins,

topoisomerases, SUN2, and AURKB) had no obvious effects on Xist localization and H3K27me3. Thus, whereas the SWI/SNF factors intersect the PRC2 pathway, other interactors do not overtly affect PRC2.

Xi reactivation via targeted inhibition of synergistic interactors

Given the large number of interactors, we created a screen to analyze effects on Xi gene expression. We derived clonal fibroblast lines harboring a transgenic green fluorescent protein (GFP) reporter on the Xi (fig. S2) and shRNAs against Xist interactors. Knockdown of any one interactor did not reactivate GFP by more than fourfold (Fig. 3A, shControl + none, and fig. S3A). Suspecting synergistic repression, we targeted multiple pathways using a combination of drugs. To target DNMT1, we employed the small molecule 5'-azacytidine (aza) (30) at a nontoxic concentration of 0.3 μ M [\leq median inhibitory concentration (IC_{50})], which minimally reactivated GFP (Fig. 3A, shControl + aza). To target TOP2a/b (31), we employed etoposide (eto) at 0.3 μ M ($\leq IC_{50}$), which also minimally reactivated GFP (Fig. 3A, shControl + eto). Combining 0.3 μ M aza + eto led to an 80- to 90-fold reactivation—a level that was almost half of GFP levels on the Xa (Xa-GFP) (Fig. 3A), suggesting strong synergy between DNMT1 and TOP2 inhibitors. Using aza + eto as priming agents, we designed triple-drug combinations inclusive of shRNAs for proteins that have no specific small-molecule inhibitors. In various shRNA + aza + eto combinations, we achieved up to 230-fold GFP reactivation—levels that equaled or exceeded Xa-GFP levels (Fig. 3A). The greatest effects were observed for combinations using shSMARCC2 (227x), shSMARCA4 (180x), and shRAD21 (211x). shTOP1 and shCTCF were also effective (175x and 154x, respectively). Combinations involving remaining interactors yielded 63 to 94x reactivation.

We then performed allele-specific RNA sequencing (RNA-seq) to investigate native Xi genes. In an F1 hybrid fibroblast line in which the Xi is of *Mus musculus* (mus) origin and the Xa of *Mus castaneus* (cas) origin, >600,000 X-linked sequence polymorphisms enabled allele-specific calls (32). Two biological replicates of each of the most promising triple-drug treatments showed good correlation (figs. S4 to S6). RNA-seq analysis showed reactivation of 75 to 100 Xi-specific genes in one replicate (Fig. 3B) and up to 200 in a second replicate (fig. S3B), representing a large fraction of expressed X-linked genes, considering that only ~210 X-linked genes have an fragments per kilobase of transcript per million mapped reads (FPKM) ≥ 1.0 in this hybrid fibroblast line. Heat map analysis demonstrated that, for individual Xi genes, reactivation levels ranged from 2 to 80x for various combinatorial treatments (Fig. 3C). There was a net increase in expression level (Δ FPKM) from the Xi in the triple-drug-treated samples relative to the shControl + aza + eto, whereas the Xa and autosomes showed no obvious net increase, thereby suggesting preferential effects on the Xi due to targeting synergistic components of the Xist interactome. Reactivation

was not specific to any one Xi region (Fig. 3D). Most effective were shRAD21, shSMC3, shSMC1a, shSMARCA4, shTOP2a, and shAURKB drug combinations. Genic examination confirmed increased representation of mus-specific tags (red) relative to the shControl (Fig. 3E). Such allelic effects were not observed at imprinted loci and other autosomal genes (fig. S7), further suggesting Xi-specific allelic effects. The set of reactivated genes varied among drug treatments, although some genes (e.g., *Rbbp7*, *G6pdx*, and *Fmr1*) appeared more prone to reactivation. Thus, the Xi is maintained by multiple synergistic pathways, and Xi genes can be reactivated preferentially by targeting two or more synergistic Xist interactors.

Xist interaction leads to cohesin repulsion

To investigate the mechanism, we focused on one group of interactors—the cohesins—because they were among the highest-confidence hits and their knockdowns consistently destabilized Xi repression. To obtain Xa and Xi binding patterns, we performed allele-specific chromatin immunoprecipitation sequencing (ChIP-seq) for two cohesin subunits, SMC1a and RAD21, and for CTCF, which works together with cohesins (28, 33–35). In wild-type cells, CTCF binding was enriched on Xa (cas) but also showed a number of Xi (mus)-specific sites (Fig. 4A) (25, 36). Allelic ratios ranged from equal to nearly complete Xa or Xi skewing (Fig. 4A). For the cohesins, 1490 SMC1a and 871 RAD21 binding sites were mapped onto the X chromosome in total, of which allelic calls could be made on ~50% of sites (Fig. 4, B and C). Although the Xa and Xi each showed significant cohesin binding, Xa-specific sites greatly outnumbered Xi-specific sites. For SMC1a, 717 sites were called on Xa, of which 589 were Xa-specific; 203 sites were called on Xi, of which 20 were Xi-specific. For RAD21, 476 sites were called on Xa, of which 336 were Xa-specific; 162 sites were called on Xi, of which 18 were Xi-specific. Biological replicates showed similar trends (fig. S8, A and B).

Cohesin's Xa preference was unexpected in light of Xist's physical interaction with cohesins—an interaction suggesting that Xist might recruit cohesins to the Xi. We therefore conditionally ablated Xist from the Xi (Xi^{AXist}) and repeated ChIP-seq analysis in the Xi^{AXist}/Xa^{WT} fibroblasts (37). Surprisingly, Xi^{AXist} acquired 106 SMC1a and 48 RAD21 sites in cis at positions that were previously Xa-specific (Fig. 4, C and D). Biological replicates trended similarly (figs. S8 and S9). In nearly all cases, acquired sites represented a restoration of Xa sites, rather than binding to random positions. By contrast, sites that were previously Xi-specific remained intact (Fig. 4, C and E, and fig. S8B), suggesting that they do not require Xist for their maintenance. The changes in cohesin peak densities were Xi-specific and significant (Fig. 4F). Cohesin restoration occurred throughout Xi^{AXist} , resulting in domains of bi-allelic binding (Fig. 4G and figs. S10 to S12), and often favored regions that harbor genes that escape XCI (e.g., *Bgn*) (38, 39). There were also shifts in CTCF binding, more noticeable at a locus-specific level than at a chromosomal level (Fig. 4,

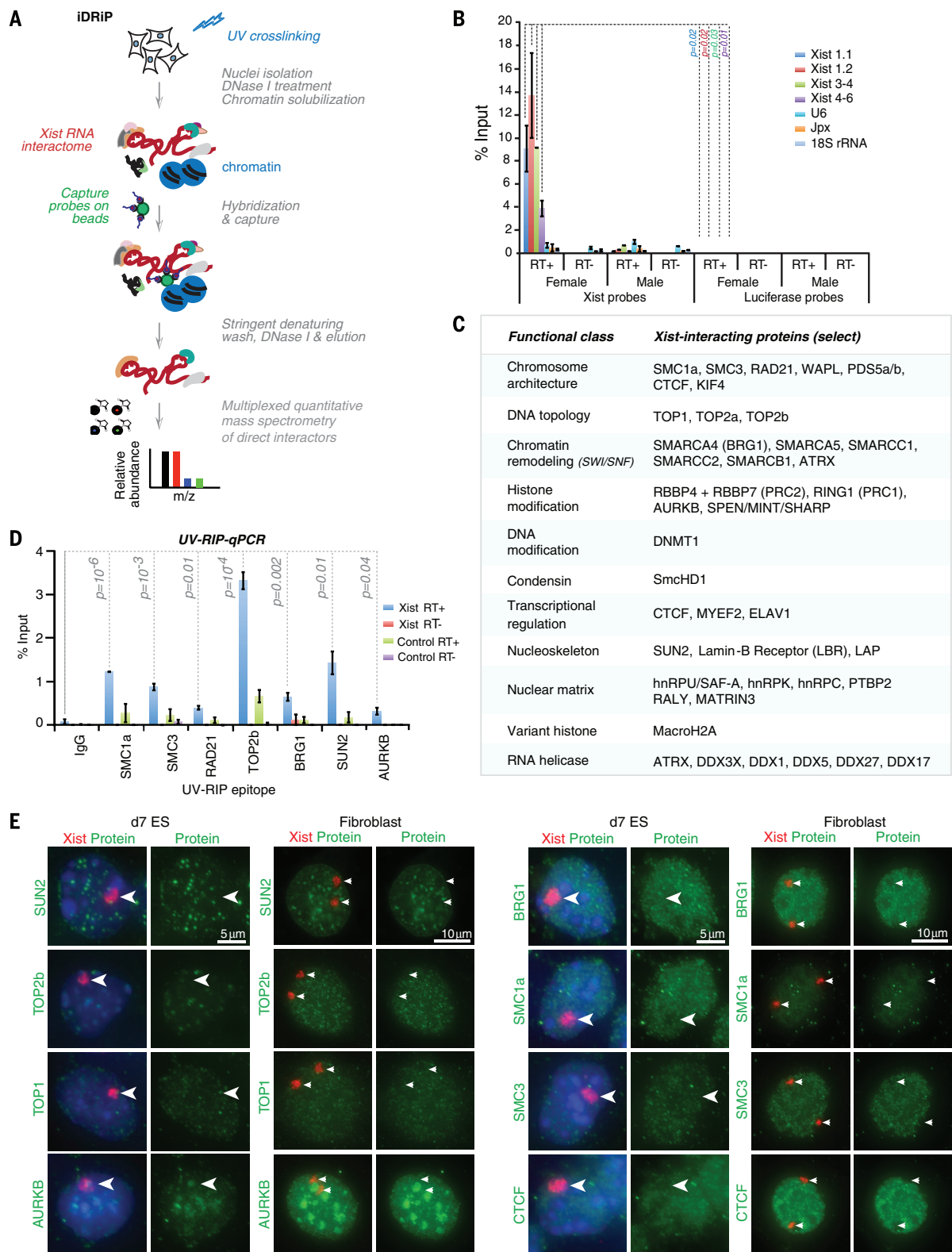


Fig. 1. iDRiP-MS reveals a large Xist interactome. (A) iDRiP schematic. (B) RT-qPCR demonstrated the specificity of Xist pulldown by iDRiP. Xist and control luciferase probes were used for pulldown from UV-cross-linked female and control male fibroblasts. Efficiency of Xist pulldown was calculated by comparing to a standard curve generated using 10-fold dilutions of input. Mean \pm standard error (SE) of three independent experiments shown. *P* determined by Student's *t* test. (C) Selected high-confidence candidates from three

biological replicates are grouped into functional classes. Additional candidates are shown in table S1. (D) UV-RIP-qPCR validation of candidate interactors. Enrichment calculated as percentage input, as in (B). Mean \pm SE of three independent experiments shown. *P* determined by Student's *t* test. (E) RNA immunoFISH to examine localization of candidate interactors (green) in relation to Xist RNA (red). Immortalized MEF cells are tetraploid and harbor two Xi.

A and G), suggesting that CTCF and cohesins do not necessarily track together on the Xi. The observed dynamics were X chromosome-specific and were not observed on autosomes (fig. S13). To determine whether there were restoration hotspots, we plotted restored SMC1a and RAD21 sites (Fig. 4H, purple) on Xi^{AXist} and observed clustering within gene-rich regions. We conclude that Xist does not recruit cohesins to the Xi-specific sites. Instead, Xist actively repels cohesins in cis to prevent establishment of the Xa pattern.

Xist RNA directs an Xi-specific chromosome conformation

Cohesins and CTCF have been shown to facilitate formation of large chromosomal domains called TADs (topologically associated domains) (27, 28, 34, 35, 40–42). The function of TADs is currently not understood, because TADs are largely

invariant across development. However, X-linked domains are exceptions to this rule and are therefore compelling models to study function of topological structures (43–46). By carrying out allele-specific Hi-C, we asked whether cohesin restoration altered the chromosomal architecture of Xi^{AXist}. First, we observed that, in wild-type cells, our TADs called on autosomal contact maps at 40-kb resolution resembled published composite (nonallelic) maps (27) (Fig. 5A, bottom). Our X chromosome contact maps were also consistent, with TADs being less distinct due to a summation of Xa and Xi reads in the composite profiles (Fig. 5A, top). Using the 44% of reads with allelic information, our allelic analysis yielded high-quality contact maps at 100-kb resolution by combining replicates (Fig. 5B and fig. S14A) or at 200-kb resolution with a single replicate. In wild-type cells, we deduced 112 TADs at 40-kb resolution on the X chromosome using the

method of Dixon *et al.* (27). We attempted TAD calling for the Xi on the 100-kb contact map but were unable to obtain obvious TADs, suggesting that the 112 TADs are present only on the Xa. The Xi instead appeared to be partitioned into two megadomains at the *DXZ4* region (fig. S14A) (46). Thus, although the Xa is topologically organized into structured domains, the Xi is devoid of such megabase-scale structures across its full length.

When Xist was ablated, however, TADs were restored in cis, and the Xi reverted to an Xa-like conformation (Fig. 5B and fig. S14B). In mutant cells, ~30 TADs were gained on Xi^{AXist} in each biological replicate. Where TADs were restored, Xi^{AXist} patterns (red) became nearly identical to those of the Xa (blue), with similar interaction frequencies. These Xi^{AXist} regions now bore little resemblance to the Xi of wild-type cells (Xi^{WT}, orange). Overall, the difference in the average interaction scores between Xi^{WT} and Xi^{AXist} was

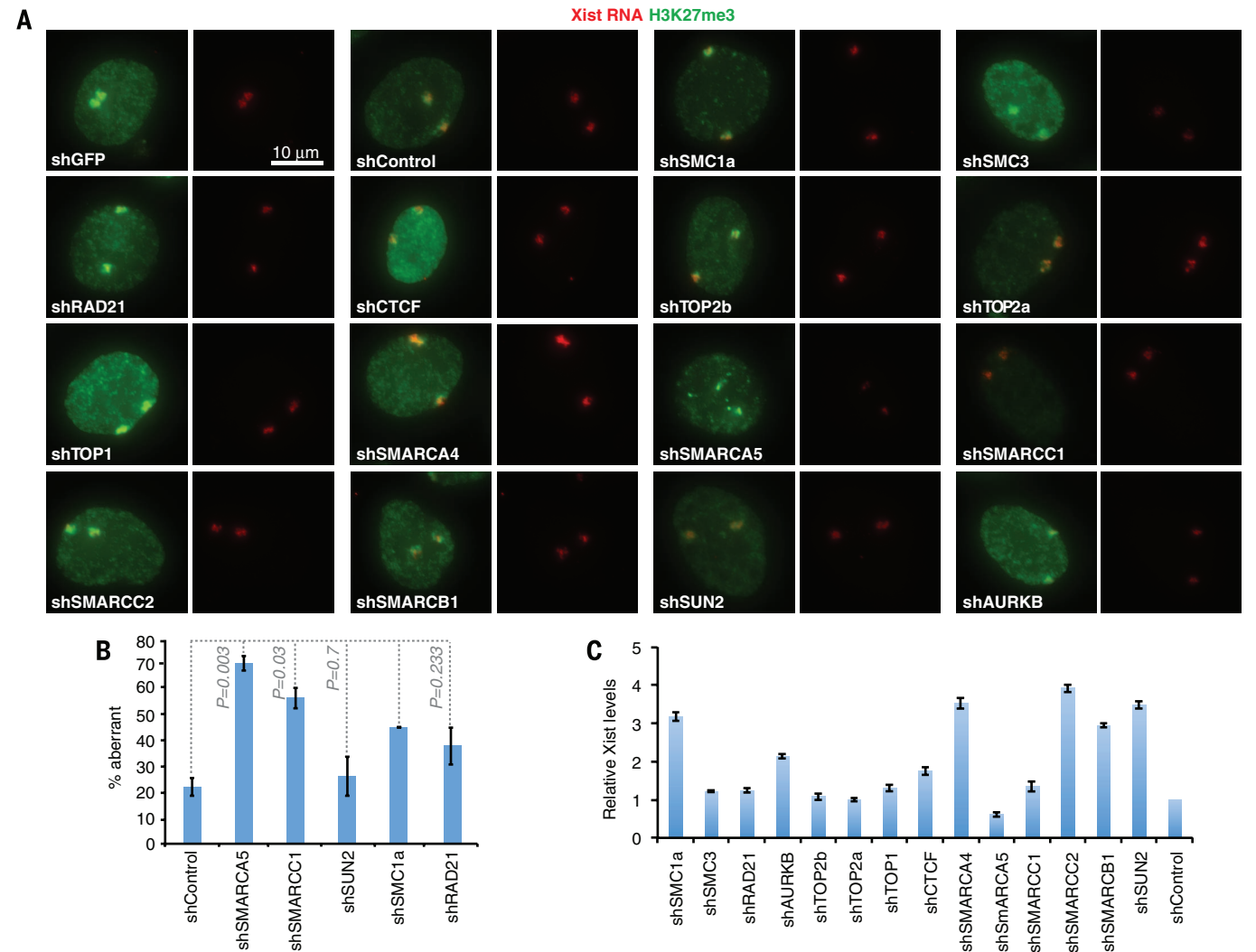


Fig. 2. Effect of depleting Xist interactors on H3K27me3. (A) RNA immunofluorescence of Xist (red) and H3K27me3 (green) after shRNA KD of interactors in fibroblasts (tetraploid; two Xist clouds). KD efficiencies (fraction remaining): SMC1a-0.48, SMC3-0.39, RAD21-0.15, AURKB-0.27, TOP2b-0.20, TOP2a-0.42, TOP1-0.34, CTCF-0.62, SMARCA4-0.52, SMARCA5-0.18, SMARCC1-0.25, SMARCC2-0.32, SMARCB1-0.52, and SUN2-0.72. Some fac-

tors are essential; therefore, high-percentage KD may be inviable. All images are presented at the same photographic exposure and contrast. **(B)** Quantitation of RNA immunofluorescence results. *n*, sample size. Percentages of aberrant Xist/H3K27me3 associations are shown. **(C)** RT-qPCR of Xist levels in KD fibroblasts, normalized to shControls. Means \pm SD of two independent experiments are shown.

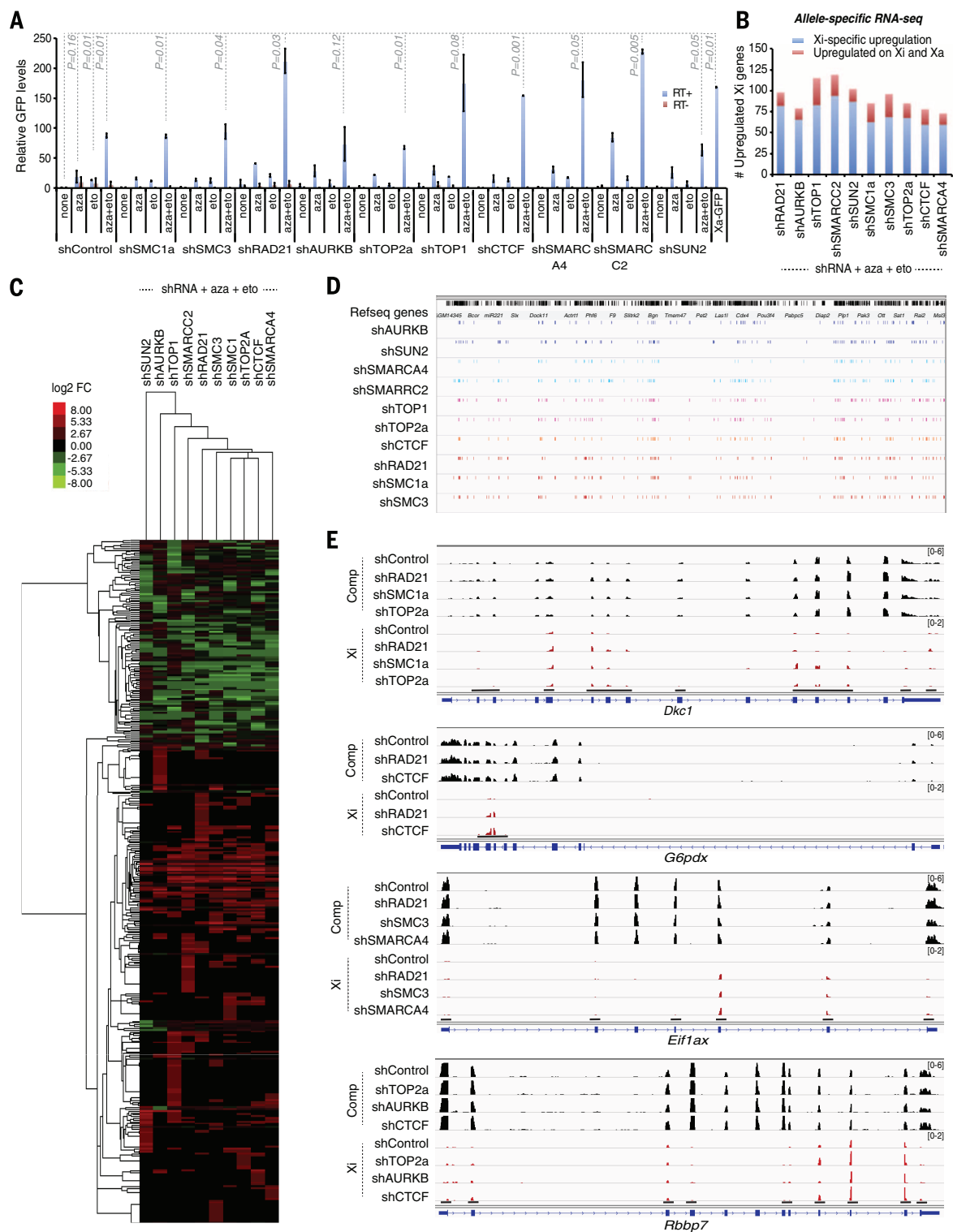


Fig. 3. Derepression of Xi genes by targeting Xist interactors. (A) Relative GFP levels by RT-qPCR analysis in female fibroblasts stably knocked down for indicated interactors \pm 0.3 μ M 5'-azacytidine (aza) \pm 0.3 μ M etoposide (eto). Xa-GFP, control male fibroblasts with X-linked GFP. Means \pm SE of two independent experiments are shown. *P* determined by Student's *t* test. (B) Allele-specific RNA-seq analysis: Number of up-regulated Xi genes for each indicated triple-drug treatment (aza + eto + shRNA). Blue, genes specifically reactivated on Xi [fold change (FC) > 2]; red, genes also up-regulated on Xa (FC > 1.3). (C) RNA-seq heat

map indicating that a large number of genes on the Xi were reactivated. X-linked genes reactivated in at least one of the triple-drug treatment (aza + eto + shRNA) were shown in the heat map. Color key, Log2 FC. Cluster analysis performed based on similarity of KD profiles (across) and on the sensitivity and selectivity of various genes to reactivation (down). (D) Chromosomal locations of Xi reactivated genes (colored ticks) for various aza + eto + shRNA combinations. (E) Read coverage of four reactivated Xi genes after triple-drug treatment. Xi, mus reads (scale, 0 to 2). Comp, total reads (scale, 0 to 6). Red tags appear only in exons with SNPs.

highly significant (Fig. 5C and fig. S15A). Intersecting TADs with SMC1a sites on Xi^{ΔXist} revealed that 61 restored cohesin sites overlapped restored TADs (61 did not overlap). In general, restored cohesin sites occurred both within TADs and at

TAD borders. TADs overlapping restored peaks had larger increases in interaction scores relative to all other TADs (Fig. 5D and fig. S15B), and we observed an excellent correlation between the restored cohesin sites and the restored TADs

(Fig. 5E and fig. S15C), consistent with a role of cohesins in reestablishing TADs after Xist deletion. Taken together, these data uncover a role for RNA in establishing topological domains of mammalian chromosomes and demonstrate that Xist must

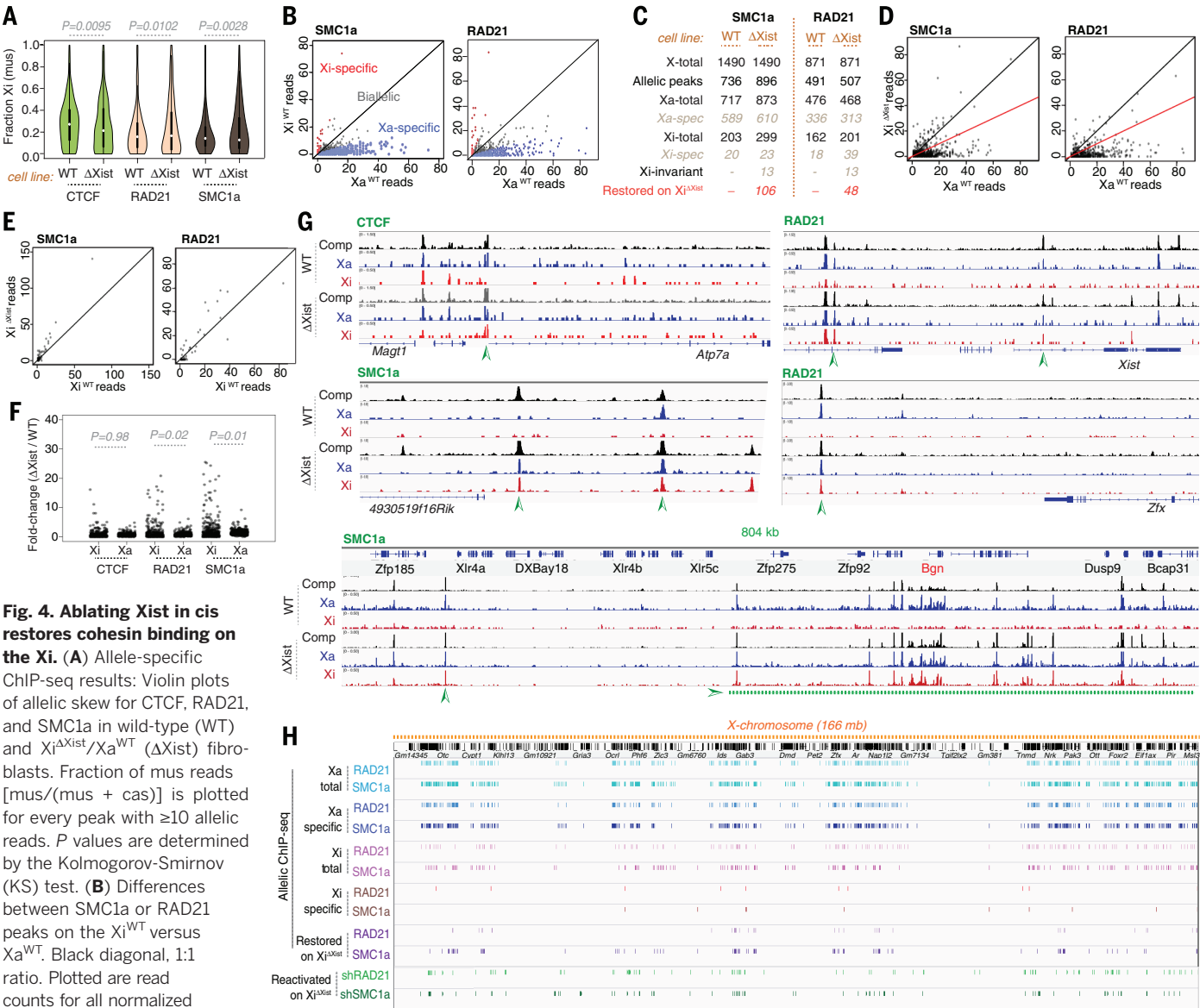
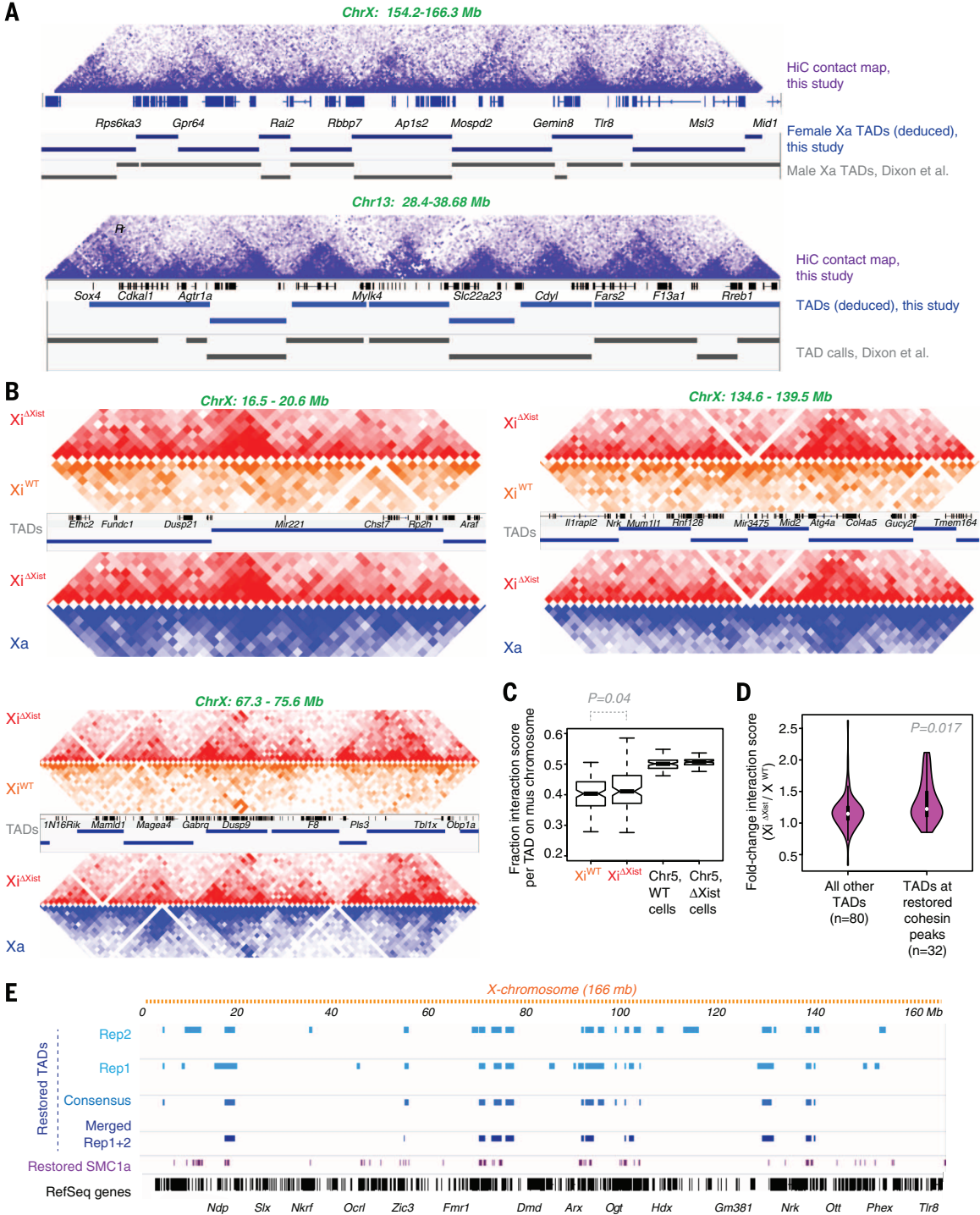


Fig. 4. Ablating Xist in cis restores cohesin binding on the Xi. (A) Allele-specific ChIP-seq results: Violin plots of allelic skew for CTCF, RAD21, and SMC1a in wild-type (WT) and Xi^{ΔXist}/Xa^{WT} (ΔXist) fibroblasts. Fraction of mus reads [mus/(mus + cas)] is plotted for every peak with ≥10 allelic reads. *P* values are determined by the Kolmogorov-Smirnov (KS) test. (B) Differences between SMC1a or RAD21 peaks on the Xi^{WT} versus Xa^{WT}. Black diagonal, 1:1 ratio. Plotted are read counts for all normalized SMC1a or RAD21 peaks.

Allele-specific skewing is defined as ≥threefold skew toward either Xa (cas, blue dots) or Xi (mus, red dots). Biallelic peaks, gray dots. (C) Table of total, Xa-specific, and Xi-specific cohesin binding sites in WT versus ΔXist (Xi^{ΔXist}/Xa^{WT}) cells. Significant SMC1a and RAD21 allelic peaks with ≥5 reads were analyzed. Allele-specific skewing is defined as ≥threefold skew toward Xa or Xi. Sites were considered “restored” if Xi^{ΔXist}’s read counts were ≥50% of Xa’s. X-total, all X-linked binding sites; allelic peaks, sites with allelic information; Xa-total, all Xa sites; Xi-total, all sites; Xa-spec, Xa-specific; Xi-spec, Xi-specific; Xi-invariant, Xi-specific in both WT and Xi^{ΔXist}/Xa^{WT} cells. There is a net gain of 96 sites on the Xi in the mutant, a number different from the number of restored sites (106). This difference arises because restored peaks are defined as sites that are heavily Xa-skewed in WT but acquire substantial Xi-binding after Xist deletion; thus, the number of restored sites is not simply the net change in Xi sites. (D) Partial restoration of SMC1a or RAD21 peaks on the Xi^{ΔXist} to an Xa pattern. Plotted are peaks with read counts with ≥threefold skew toward Xa^{WT} (Xa-specific). x axis, normalized Xa^{WT} read counts; y axis, normalized Xi^{ΔXist}

read counts; black diagonal, 1:1 Xi^{ΔXist}/Xa^{WT} ratio; red diagonal, 1:2 ratio. (E) Xi-specific SMC1a or RAD21 peaks remained on Xi^{ΔXist}. Plotted are read counts for SMC1a or RAD21 peaks with ≥threefold skew toward Xi^{WT} (“Xi-specific”). (F) Comparison of fold changes for CTCF, RAD21, and SMC1 binding in Xi^{ΔXist} cells relative to WT cells. Shown are fold changes for Xi versus Xa. The Xi showed significant gains in RAD21 and SMC1a binding, but not in CTCF binding. Method: X^{WT} and Xi^{ΔXist} ChIP samples were normalized by scaling to equal read counts. Fold changes for Xi were computed by dividing the normalized mus read count in Xi^{ΔXist} by the mus read count X^{WT}; fold changes for Xa were computed by dividing the normalized cas read count in Xi^{ΔXist} by the cas read count X^{WT}. To eliminate noise, peaks with <10 allelic reads were eliminated from analysis. *P* values were determined by a paired Wilcoxon signed rank test. (G) The representative examples of cohesin restoration on Xi^{ΔXist}. Arrowheads, restored peaks. (H) Allelic-specific cohesin binding profiles of Xa, Xi^{WT}, and Xi^{ΔXist}. Shown below restored sites are regions of Xi reactivation after shSMC1a and shRAD21 triple-drug treatments, as defined in Fig. 3.

Fig. 5. Ablating Xist results in Xi reversion to an Xa-like chromosome conformation. (A) Chr13 and X chromosome contact maps showing triangular domains representative of TADs. Purple shades correspond to varying interaction frequencies (dark, greater interactions). TADs called from our composite (nonallelic) HiC data at 40-kb resolution (blue bars) are highly similar to those (gray bars) called previously (27). **(B)** Allele-specific HiC-seq analysis: Contact maps for three different X chromosome regions at 100-kb resolution comparing $Xi^{ΔXist}$ (red) to Xi^{WT} (orange), and $Xi^{ΔXist}$ (red) versus Xa (blue) of the mutant cell line. Our TAD calls are shown with reference sequence (RefSeq) genes. **(C)** Fraction of interaction frequency per TAD on the Xi (mus) chromosome. The positions of our TAD borders were rounded to the nearest 100 kb, and submatrices were generated from all pixels between the two end points of the TAD border for each TAD. We calculated the average interaction score for each TAD by summing the interaction scores for all pixels in the submatrix defined by a TAD and dividing by the total number of pixels in the TAD. We then averaged the normalized interaction scores across all bins in a TAD in the Xi (mus) and Xa (cas) contact maps and computed the fraction of averaged interaction scores from mus chromosomes. The X chromosome and a representative autosome, Chr5, are shown for the WT cell line and the $Xist^{ΔXist/+}$ cell line. *P* values were determined by paired Wilcoxon signed rank test. **(D)** Violin plots showing that TADs overlapping restored peaks have larger increases in interaction scores relative to all other TADs. We calculated the fold change in average interaction scores on the Xi for all X-linked TADs and intersected the TADs with SMC1a sites ($Xi^{ΔXist}/Xi^{WT}$). Thirty-two TADs occurred at restored cohesin sites; 80 TADs did not overlap restored cohesin sites. Violin plot shows distributions of fold change average interaction scores between Xi^{WT} and $Xi^{ΔXist}$. *P* values were determined by Wilcoxon ranked sum test. **(E)** Restored TADs



actively and continually repulse cohesins from the Xi, even during the maintenance phase, to prevent formation of an Xa chromosomal architecture.

Discussion

Using iDRiP, we have identified a comprehensive Xist interactome and revealed multiple synergistic pathways to Xi repression (Fig. 6). With Xist physically contacting 80 to 250 proteins at any given time, the Xist ribonucleoprotein particle may be as large as the ribosome. Our study supports a model in which Xist RNA simultaneously acts as (i) scaffold for the recruitment of repressive complexes (such as PRC1, PRC2, ATRX, mH2A, and SmcHD1) to establish and maintain the inactive state; and as (ii) a repulsion mechanism to extrude architectural factors such as cohesins to avoid acquisition of a transcription-favorable chromatin conformation. Without Xist, cohesins return to their default Xa binding state. Repulsion could be based on eviction, with Xist releasing cohesins as it extrudes them, or on sequestration, with Xist sheltering cohesins to prevent Xi binding. Our study shows that the Xi harbors three types of cohesin sites: (i) Xi-specific sites that do not depend on Xist; (ii) biallelic sites that are also Xist-independent; and (iii) Xa-specific sites, many of which cannot be established on the Xi because of active repulsion by Xist. The type (i) and type (iii) sites likely explain the paradoxical observations that, on the one hand, depleting cohesins leads to Xi reactivation but, on the other, loss of Xist-mediated cohesin recruitment leads to an Xa-like chromosome conformation that is permissive for transcription. In essence, modulating the type (i) and type (iii) sites both have the effect of destabilizing the Xi, rendering the Xi more accessible to transcription. Disrupting type (i) sites by cohesin knockdown would change the repressive Xi structure, while ablating Xist would restore the type (iii) sites that promote an Xa-like conformation. Our study has focused on cohesins, but RNA-mediated repulsion may be an outcome for other Xist interactors and may be as prevalent an epigenetic mechanism as RNA-mediated recruitment (47).

The robustness of Xi silencing is demonstrated by the observation that we destabilized the Xi only after pharmacologically targeting two or three distinct pathways. The fact that the triple-drug treatments varied with respect to reactivated loci and depth of derepression creates the possibility of treating X-linked disease in a locus-specific manner by administering unique drug combinations. Given the existence of many other disease-associated lncRNAs, the iDRiP technique could be applied systematically toward identifying new drug targets for other diseases and generally for elucidating mechanisms of epigenetic regulation by lncRNA.

Materials and methods

Identification of Direct RNA interacting Proteins (iDRiP)

Mouse embryonic fibroblasts (MEFs) were irradiated with UV light at 200 mJ energy (Stratagene 2400) after rinsing with phosphate-buffered

saline (PBS). The pellets were resuspended in cytoskeleton buffer with 0.5% Triton X-100 (CSKT)-0.5% [10 mM piperazinediethanesulfonic acid, pH 6.8, 100 mM NaCl, 3 mM MgCl₂, 0.3 M sucrose, 0.5% Triton X-100, 1 mM phenylmethylsulfonyl fluoride (PMSF)] for 10 min at 4°C followed by a spin. The pellets were again resuspended in nuclear isolation buffer (10 mM Tris pH 7.5, 10 mM KCl, 0.5% Nonidet-P 40, 1x protease inhibitors, 1 mM PMSF), and rotated at 4°C for 10 min (optional step). The pellets were collected after a spin, weighed, flash frozen in liquid nitrogen, and stored at -80°C until use.

Approximately equal amounts of female and male UV cross-linked pellets were thawed and resuspended for treatment with Turbo DNase I in the DNase I digestion buffer (50 mM Tris pH 7.5, 0.5% Nonidet-P 40, 0.1% sodium lauroyl sarcosine, 1x protease inhibitors, SuperaseIn). The tubes were rotated at 37°C for 45 min with intermittent mixing or vortexing. The nuclear lysates were further solubilized by adding 1% sodium lauroyl sarcosine, 0.3 M lithium chloride, 25 mM EDTA, and 25 mM EGTA to final concentrations. After brief vortexing, continue incubation at 37°C for 15 min. The lysates were mixed with biotinylated DNA probes (table S3) prebound to the streptavidin magnetic beads (MyOne streptavidin C1 Dyna beads, Invitrogen) and incubated at 55°C for 1 hour before overnight incubation at 37°C in the hybridization chamber. The beads were washed three times in wash buffer (10 mM Tris, pH 7.5, 0.3 M LiCl, 1% LDS, 0.5% Nonidet-P 40, 1x protease inhibitor) at room temperature followed by treatment with Turbo DNase I in DNase I digestion buffer with the addition of 0.3 M LiCl, protease inhibitors, and superaseIn at 37°C for 20 min. Then, beads were resuspended and washed two more times in the wash buffer.

For MS analysis, elution was done in elution buffer (10 mM Tris, pH 7.5, 1 mM EDTA) at 70°C for 4 min followed by brief sonication in Covaris. For the quantification of pulldown efficiency, MEFs, without cross-linking, were used and elution was done at 95°C. The elute was used for RNA isolation and reverse transcription qPCR (RT-qPCR). When cross-linked MEFs were used, elute was subjected for proteinase-K treatment (50 mM Tris pH 7.5, 100 mM NaCl, 0.5% SDS, 10 µg proteinase K) for 1 hour at 55°C. RNA was isolated by Trizol and quantified with SYBR green qPCR. Input samples were used to make standard curve by 10-fold dilutions, to which the RNA pulldown efficiencies were compared and calculated. The efficiency of Xist pulldown was relatively lower after UV cross-linking, similar to (48, 49).

Quantitative proteomics

Proteins co-enriched with Xist from female or male cells were quantitatively analyzed either using a label-free approach based on spectral counting (21) or by multiplexed quantitative proteomics using tandem-mass tag (TMT) reagents (50, 51) on an Orbitrap Fusion mass spectrometer (Thermo Scientific). Disulfide bonds were reduced with dithiothreitol (DTT) and free thiols alkylated with iodoacetamide as described previously (22). Proteins were then precipitated with trichloroacetic acid, resuspended in 50 mM HEPES (pH 8.5) and 1 M urea, and digested first with endoproteinase Lys-C (Wako) for 17 hours at room temperature and then with sequencing-grade trypsin (Promega) for 6 hours at 37°C. Peptides were desalted over Sep-Pak C₁₈ solid-phase extraction (SPE) cartridges (Waters), and the peptide concentration was determined using a bicinchoninic acid (BCA) assay (Thermo Scientific). For the label-free analysis,

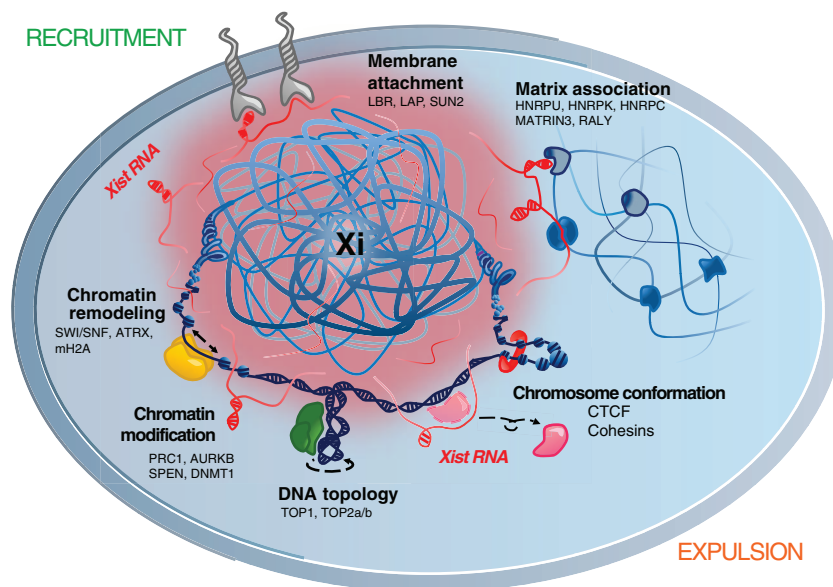


Fig. 6. The Xi is suppressed by multiple synergistic mechanisms. Xist RNA (red) suppresses the Xi by either recruiting repressive factors (e.g., PRC1 and PRC2) or expelling architectural factors (e.g., cohesins).

peptides were then dried and resuspended in 5% formic acid (FA) and 5% acetonitrile (ACN), and 5 µg of peptides were analyzed by mass spectrometry as described below. For the multiplexed quantitative analysis, a maximum of 50 µg of peptides were labeled with one out of the available TMT-10plex reagents (Thermo Scientific) (50). To achieve this, peptides were dried and resuspended in 50 µl of 200 mM HEPES (pH 8.5) and 30% (ACN), and 10 µg of the TMT in reagent in 5 µl of anhydrous ACN was added to the solution, which was incubated at room temperature (RT) for 1 hour. The reaction was then quenched by adding 6 µl of 5% (w/v) hydroxylamine in 200 mM HEPES (pH 8.5) and incubated for 15 min at RT. The labeled peptide mixture was then subjected to a fractionation using basic pH reversed-phase liquid chromatography (bRPLC) on an Agilent 1260 Infinity high-performance liquid chromatography (HPLC) system equipped with an Agilent Extend-C18 column (4.6 x 250 mm; particle size, 5 µm), basically as described previously (52). Peptides were fractionated using a gradient from 22 to 35% ACN in 10 mM ammonium bicarbonate over 58 min at a flow rate of 0.5 ml/min. Fractions of 0.3 ml were collected into a 96-well plate to then be pooled into a total 12 fractions (A1 to A12, B1 to B12, etc.) that were dried and resuspended in 8 µl of 5% FA and 5% ACN, 3 of which were analyzed by microcapillary liquid chromatography tandem mass spectrometry on an Orbitrap Fusion mass spectrometer, and a recently introduced multistage (MS3) method was used to provide highly accurate quantification (53).

The mass spectrometer was equipped with an EASY-nLC 1000 integrated autosampler and HPLC pump system. Peptides were separated over a 100-µm inner diameter microcapillary column in-house packed with first 0.5 cm of Magic C4 resin (5 µm, 100 Å, Michrom Bioresources), then with 0.5 cm of Maccell C₁₈ resin (3 µm, 200 Å, Nest Group) and 29 cm of GP-C18 resin (1.8 µm, 120 Å, Sepax Technologies). Peptides were eluted applying a gradient of 8 to 27% ACN in 0.125% formic acid over 60 min (label-free) and 165 min (TMT) at a flow rate of 300 nl/min. For label-free analyses, we applied a tandem-MS method where a full-MS spectrum [MS1; mass-to-charge ratio (m/z) 375 to 1500; resolution 6×10^4 ; automated gain control (AGC) target, 5×10^5 ; maximum injection time, 100 ms] was acquired using the Orbitrap, after which the most abundant peptide ions were selected for linear ion trap CID-MS2 in an automated fashion. MS2 scans were done in the linear ion trap using the following settings: quadrupole isolation at an isolation width of 0.5 Th; fragmentation method, CID; AGC target, 1×10^4 ; maximum injection time, 35 ms; normalized collision energy, 30%. The number of acquired MS2 spectra was defined by setting the maximum time of one experimental cycle of MS1 and MS2 spectra to 3 s (top speed). To identify and quantify the TMT-labeled peptides, we applied a synchronous precursor selection MS3 method (22, 53, 54) in a data-dependent mode. The scan sequence was started with the acquisition of a full MS or MS1 spectrum acquired in the Orbitrap

(m/z range, 500 to 1200; other parameters were set as described above), and the most intense peptide ions detected in the full MS spectrum were then subjected to MS2 and MS3 analysis, while the acquisition time was optimized in an automated fashion (top speed, 5 s). MS2 scans were performed as described above. Using synchronous precursor selection, the 10 most abundant fragment ions were selected for the MS3 experiment after each MS2 scan. The fragment ions were further fragmented using the higher-energy collisional dissociation (HCD) fragmentation (normalized collision energy, 50%), and the MS3 spectrum was acquired in the Orbitrap (resolution, 60,000; AGC target, 5×10^4 ; maximum injection time, 250 ms).

Data analysis was performed on an in-house generated SEQUEST-based (55) software platform. RAW files were converted into the mzXML format using a modified version of ReAdW.exe. MS2 spectra were searched against a protein sequence database containing all protein sequences in the mouse UniProt database (downloaded 4 February 2014), as well as that of known contaminants such as porcine trypsin. This target component of the database was followed by a decoy component containing the same protein sequences but in flipped (or reversed) order (56). MS2 spectra were matched against peptide sequences, with both termini consistent with trypsin specificity and allowing two missed trypsin cleavages. The precursor ion m/z tolerance was set to 50 parts per million; TMT tags on the N terminus and on lysine residues (229.162932 Da, only for TMT analyses), as well as carbamidomethylation (57.021464 Da) on cysteine residues were set as static modification; and oxidation (15.994915 Da) of methionines was set as variable modification. Using the target-decoy database search strategy (56), a spectra assignment false discovery rate (FDR) of less than 1% was achieved through using linear discriminant analysis, with a single discriminant score calculated from the following SEQUEST search score and peptide sequence properties: mass deviation, XCorr, dCn, number of missed trypsin cleavages, and peptide length (57). The probability of a peptide assignment to be correct was calculated using a posterior error histogram, and the probabilities for all peptides assigned to a protein were combined to filter the data set for a protein FDR of less than 1%. Peptides with sequences that were contained in more than one protein sequence from the UniProt database were assigned to the protein with the most matching peptides (57).

For a quantitative estimation of protein concentration using spectral counts, we counted the number of MS2 spectra assigned to a given protein (table S1). TMT reporter ion intensities were extracted as that of the most intense ion within a 0.03 Th window around the predicted reporter ion intensities in the collected MS3 spectra. Only MS3 with an average signal-to-noise value of larger than 28 per reporter ion, as well as with an isolation specificity (22) of larger than 0.75, were considered for quantification. Reporter ions from all peptides assigned to a protein were summed to

define the protein intensity. A two-step normalization of the protein TMT intensities was performed by first normalizing the protein intensities over all acquired TMT channels for each protein based to the median average protein intensity calculated for all proteins. To correct for slight mixing errors of the peptide mixture from each sample, a median of the normalized intensities was calculated from all protein intensities in each TMT channel, and the protein intensities were normalized to the median value of these median intensities.

UV RIP

The protocol followed is similar to the one described in (18). Briefly, MEFs were cross-linked with UV light at 200 mJ and collected by scraping in PBS. Cell pellets were resuspended in CSKT-0.5% for 10 min at 4°C followed by a spin. The nuclei were resuspended in the UV RIP buffer [PBS buffer containing 300 mM NaCl (total), 0.5% Nonidet-P 40, 0.5% sodium deoxycholate, 200 U Protector RNase Inhibitor and 1x protease inhibitors] with Turbo DNase I 30 U/IP for 30 min at 37°C. Supernatants were collected after a spin and incubated with 5 µg specific antibodies prebound to 40 µl protein-G magnetic beads (Invitrogen) at 4°C overnight. Beads were washed three times with cold UV RIP buffer. The beads were resuspended in 200 µl Turbo DNase I buffer with 20 U Turbo DNase, SuperscriptIN, 1x protease inhibitors for 30 min at 37°C. The beads were resuspended and washed three more times in the UV RIP washing buffer containing 10 mM EDTA. The final three washes were given after threefold dilution of UV RIP washing buffer. The beads were resuspended in 200 µl proteinase-K buffer with 10 µg proteinase-K and incubated at 55°C for 1 hour. RNA was isolated by Trizol, and pull-down efficiencies were calculated by SYBR qPCR using input for the standard curve.

Generation of Xi-TgGFP clonal fibroblasts

Xi-TgGFP (68-5-11) tail-tip fibroblasts (TTF) were initially derived from a single female pup, a daughter of a cross between a *M. castaneus* male and a *M. musculus* female, homozygous for an X-linked GFP transgene driven by a strong, ubiquitous promoter (58). The fibroblasts were immortalized by SV40 transformation, and clonal lines were derived from individual GFP-negative cells selected by fluorescence-activated cell sorting. In our experience, occasional clones with undetectable GFP expression nevertheless have the transgene located on the active X chromosome. Thus, we confirmed the GFP transgene location on the inactive X for the particular clone used here, 68-5-11 (see fig. S2).

Generation of stable KD of Xi-TgGFP TTF and 16.7 ES cells

The protocol is as described in <http://www.broadinstitute.org/rnai/public/resources/protocols>

A cocktail of three shRNA viruses was used for infections (table S2) followed with puromycin selection. In all the experiments, nonclonal knock-down cells were used.

Assay for the reactivation of Xi-TgGFP

About 125,000 to 150,000 Xi-TgGFP (68-5-11) cells were plated along with control (shControl) cells treated with dimethyl sulfoxide or stable KD cells treated with 0.3 μ M azacytidine and 0.3 μ M Etoposide for 3 days in six well plates. RNA was isolated by Trizol twice, with an intermittent TurboDNase treatment after the first isolation for 30 min at 37°C. One μ g RNA was used for each of the RT+ and RT- reactions (Superscript III, Invitrogen) followed by the SYBR green qPCR using the primers listed in table S3, with annealing temperature of 60°C for 45 cycles. The relative efficiency of Xi-TgGFP reactivations was calculated by comparing to U1 snRNA as the internal control.

ImmunofISH

Cells were grown on coverslips, rinsed in PBS, pre-extracted in 0.5% CSKT on ice, and washed once in CSK, followed by fixation with 4% paraformaldehyde in PBS at room temperature. After blocking in 1% bovine serum albumin in PBS for 20 min supplemented with 10 mM ribonucleoside-vanadyl complex (VRC) (New England Biolabs) and RNase inhibitor (Roche), incubation was carried out with primary antibodies (table S3) at room temperature for 1 hour. Cells were washed three times in PBST-0.02% Tween-20. After incubating with secondary antibody at room temperature for 30 min, cells were washed three times by PBS/0.02% Tween-20. Cells were fixed again in 4% paraformaldehyde and dehydrated in ethanol series. RNA FISH was performed using a pool of Cy3B- or Alexa 568-labeled Xist oligonucleotides for 4 to 6 hours at 42°C in a humid chamber. Cells were washed three times in 2X SSC, and nuclei were counter-stained by Hoechst 33342. Cells were observed under Nikon 90i microscope equipped with 60X/1.4 N.A. objective lens, Orca ER charge-coupled device camera (Hamamatsu), and Volocity software (Perkin Elmer). Xist RNA FISH probes, a set of total 37 oligonucleotides with 5' amine modification (IDT), were labeled with NHS-Cy3B (GE Healthcare) overnight at room temperature followed by ethanol precipitation. In the case of confirmation of Xi-TgGFP cells, probes were made by nick translation of a GFP PCR product with Cy3-dUTP and of a plasmid containing the first exon of the mouse Xist gene, with FITC-dUTP.

Allelic ChIP-seq

Allele-specific ChIP-seq was performed according to the method of Kung *et al.* (25), in two biological replicates. To increase available read depth, we pooled together two technical replicates for Xi^{Axist}/Xa^{WT} Rad21 replicate 1 sequenced on a 2 x 50 bp HiSeq. 2500 rapid run, and we also pooled two technical replicates of wild-type Rad21 replicate 1, one sequenced on a HiSeq. 2 x 50 bp run and one on a MiSeq. 2 x 50 bp run. All other libraries were sequenced on using 2 x 50 bp HiSeq. 2500 rapid runs. To visualize ChIP binding signal, we generated fragments per million (fpm)-normalized bigWig files from the raw ChIP read counts for all reads (comp), mus-specific

(mus), and cas-specific reads separately. For Smc1a, CTCF, and Rad21, peaks were called using macs2 with default settings. To generate consensus peak sets for all three epitopes, peaks for the two wild-type and Xi^{Axist}/Xa^{WT} replicates were pooled, and peaks present in at least two experiments were used as the common peak set. To make comparisons between allelic read counts between different experiments, we defined a scaling factor as the ratio of the total read numbers for the two experiments and multiplied the allelic reads for each peak in the larger sample by the scaling factor. We plotted the number of reads on Xi vs Xa in wild-type for all peaks on the X-chromosome to determine whether there is a general bias toward binding to the Xa or the Xi. To evaluate allelic skew on an autosome, we generated plots of mus read counts versus cas read counts for all peaks on chromosome 5 from 1 to 140,000,000. We used this particular region of chromosome 5 because Xi^{Axist}/Xa^{WT} is not fully hybrid, and this is a large region of an autosome that is fully hybrid based on even numbers of read counts from input and from our Hi-Cs over this region in Xi^{Axist}/Xa^{WT} (data not shown). To identify peaks that are highly Xa-skewed in wild-type but bind substantially to the Xi in Xi^{Axist}/Xa^{WT} (restored peaks), for Xa-skewed peaks in wild-type, we plotted normalized read counts on Xi in Xi^{Axist}/Xa^{WT} versus read counts on Xa in wild-type. We defined restored peaks as peaks that are (i) more than 3X Xa-skewed in wild-type, (ii) have at least five allelic reads in wild-type, and (iii) exhibit normalized read counts on Xi in Xi^{Axist}/Xa^{WT} that are at least half the level of Xa in wild-type. This threshold ensures that all restored peaks have at least a 2X increase in binding to the Xi in Xi^{Axist}/Xa^{WT} relative to wild-type. We identified restored peaks using these criteria in both replicates of Smc1a and Rad21 ChIP separately, and to merge these calls into a consensus set for each epitope, we took all peaks that met criteria for restoration in at least one replicate and had at least 50% wild-type Xa read counts on Xi in Xi^{Axist}/Xa^{WT} in both replicates.

Allele-specific RNA-seq

Xi-TgGFP TTFs (68-5-11) with the stable knock-down of candidates were treated with 5'-azacytidine and etoposide at 0.3 μ M each for 3 days. Strand-specific RNA-seq, the library preparation, deep sequencing, and data analysis was followed as described in (59). Two biological replicates of each drug treatment were produced. All libraries were sequenced with Illumina HiSeq. 2000 or 2500 using 50 cycles to obtain paired-end reads. To determine the allelic origin of each sequencing read from the hybrid cells, reads were first depleted of adaptors dimers and PCR duplicates, followed by the alignment to custom mus/129 and cas genomes to separate mus and cas reads. After removal of PCR duplicates, ~90% of reads were mappable. Discordant pairs and multimapped reads were discarded. Reads were then mapped back to reference mm9 genome using Tophat v2.0.10 (-g 1-no-coverage-search-read-edit-dist 3-read-mismatches 3-read-gap-length 3-b2-very-

sensitive-mate-inner-dist 50-mate-std-dev 50-library-type fr-firststrand), as previously described (25, 32, 59). After alignment, gene expression levels within each library were quantified using Homer v4.7 (rna mm9 -count genes -strand + -noadj -condenseGenes) (59), and the normalized differential expression analyses across samples were performed by using EdgeR (60).

Hi-C library preparation and analysis

Hi-C libraries were generated according to the protocol in Lieberman-Aiden *et al.* (61). Two biological replicate libraries were prepared for wild-type and Xi^{Axist}/Xa^{WT} fibroblasts each. We obtained 150 to 220 million 2 x 50 bp paired-end reads per library. The individual ends of the read-pairs were aligned to the mus and cas reference genomes separately using novoalign with default parameters for single-end alignments, and the quality score of the alignment was used to determine whether each end could be assigned to either the mus or the cas haplotype (62). The single-end alignments were merged into a Hi-C summary file using custom scripts. Reads were filtered for self-ligation events and short fragments (less than 1.5X the estimated insert length) likely to be random shears using Homer (59, 63). Hi-C contact maps were generated using Homer. "Comp" maps were made from all reads. "Xi" and "Xa" reads were from reads where at least one read-end could be assigned to either the mus or cas haplotype, respectively. A small fraction of reads (~5% of all allelic reads) aligned such that one end aligned to mus, the other to cas. These "discordant" reads were excluded from further analysis, because they are likely to be noise arising due to random ligation events and/or improper single-nucleotide polymorphism (SNP) annotation (46, 64). All contact maps were normalized using the matrix balancing algorithm of Knight and Ruiz (65), similar to iterative correction (46, 66), using the MATLAB script provided at the end of their paper. We were able to generate robust contact maps using the comp reads in one replicate at 40-kb resolution, but because only ~44% of reads align allele-specifically, we were only able to generate contact maps for the cas and mus haplotypes at 200 kb. To increase our resolution, we pooled together both biological replicates and analyzed the comp contact map at 40-kb resolution and the mus and cas contact maps at 100 kb. We called TADs at 40 kb on the X chromosome Chr5, and Chr13 using the method of Dixon *et al.* (27). Specifically, we processed the normalized comp 40-kb contact maps separately into a vector of directionality indices using DI_from_matrix.pl with a bin size of 40,000 and a window size of 200,000. We used this vector of directionality indices as input for the HMM_calls.m script, and after HMM_generation, we processed the HMM output to file_ends_cleaner.pl, converter_7col.pl, hmm_probability_corrector.pl, hmm-state_caller.pl, and, finally, hmm-state_domains.pl. We used parameters of min = 2, prob = 0.99, binsize = 40,000 as input to the HMM probability correction script.

To create a general metric describing interaction frequencies within TADs at resolution available in the allele-specific interaction maps, for each TAD, on the X chromosome and Chr5 we averaged the normalized interaction scores for all bins within each TAD, excluding the main diagonal. To make comparisons between interaction frequency over TADs between the cas (Xa) and mus (Xi) haplotypes at the resolution available with our current sequencing depth, we defend the “fraction mus” as the average interaction score for a TAD in the mus contact map divided by the sum of the average interaction scores in the mus and cas contact maps.

To discover TADs that show significantly increased interaction frequency in Xi^{Xist}/Xa^{WT} , we generated a null distribution of changes in average normalized interaction scores for all TADs on chromosome 5, 1 to 140 Mb using the cas and mus contact maps. We reasoned that there would be few changes in interaction frequency on an autosome between the mus or cas contact maps for wild-type and Xi^{Xist}/Xa^{WT} ; thus, the distribution of fold changes in interaction score on an autosome constitutes a null distribution. Using this distribution of fold changes allowed us to calculate a threshold fold change for an empirical FDR of 0.05, and all TADs that had a greater increase in average normalized interaction score on Xi between wild-type and Xi^{Xist}/Xa^{WT} were considered restored TADs. We performed this analysis of restored TADs separately in each biological replicate using the 200-kb contact maps to generate interaction scores over TADs and using the combined data at 100-kb resolution.

REFERENCES AND NOTES

1. J. Starmer, T. Magnuson, A new model for random X chromosome inactivation. *Development* **136**, 1–10 (2009). doi: [10.1242/dev.025908](https://doi.org/10.1242/dev.025908); pmid: [19036804](https://pubmed.ncbi.nlm.nih.gov/19036804/)
2. C. M. Disteche, Dosage compensation of the sex chromosomes. *Annu. Rev. Genet.* **46**, 537–560 (2012). doi: [10.1146/annurev-genet-110711-155454](https://doi.org/10.1146/annurev-genet-110711-155454); pmid: [22974302](https://pubmed.ncbi.nlm.nih.gov/22974302/)
3. A. Wutz, R. Agrelo, Response: The diversity of proteins linking Xist to gene silencing. *Dev. Cell* **23**, 680 (2012). doi: [10.1016/j.devcel.2012.09.017](https://doi.org/10.1016/j.devcel.2012.09.017); pmid: [23079595](https://pubmed.ncbi.nlm.nih.gov/23079595/)
4. C. J. Brown *et al.*, The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527–542 (1992). doi: [10.1016/0092-8674\(92\)90520-M](https://doi.org/10.1016/0092-8674(92)90520-M); pmid: [1423611](https://pubmed.ncbi.nlm.nih.gov/1423611/)
5. J. Wang *et al.*, Imprinted X inactivation maintained by a mouse Polycomb group gene. *Nat. Genet.* **28**, 371–375 (2001). doi: [10.1038/ng574](https://doi.org/10.1038/ng574); pmid: [11479595](https://pubmed.ncbi.nlm.nih.gov/11479595/)
6. A. Kohlmaier *et al.*, A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biol.* **2**, e171 (2004). doi: [10.1371/journal.pbio.0020171](https://doi.org/10.1371/journal.pbio.0020171); pmid: [15252442](https://pubmed.ncbi.nlm.nih.gov/15252442/)
7. K. Plath *et al.*, Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *J. Cell Biol.* **167**, 1025–1035 (2004). doi: [10.1083/jcb.200409026](https://doi.org/10.1083/jcb.200409026); pmid: [15596546](https://pubmed.ncbi.nlm.nih.gov/15596546/)
8. J. Zhao *et al.*, Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* **40**, 939–953 (2010). doi: [10.1016/j.molcel.2010.12.011](https://doi.org/10.1016/j.molcel.2010.12.011); pmid: [21172659](https://pubmed.ncbi.nlm.nih.gov/21172659/)
9. Y. Marahrens, B. Panning, J. Dausman, W. Strauss, R. Jaenisch, Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev.* **11**, 156–166 (1997). doi: [10.1101/gad.11.2.156](https://doi.org/10.1101/gad.11.2.156); pmid: [9009199](https://pubmed.ncbi.nlm.nih.gov/9009199/)
10. E. Yildirim *et al.*, Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell* **152**, 727–742 (2013). doi: [10.1016/j.cell.2013.01.034](https://doi.org/10.1016/j.cell.2013.01.034); pmid: [23415223](https://pubmed.ncbi.nlm.nih.gov/23415223/)
11. C. J. Brown, H. F. Willard, The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. *Nature* **368**, 154–156 (1994). doi: [10.1038/368154a0](https://doi.org/10.1038/368154a0); pmid: [8139659](https://pubmed.ncbi.nlm.nih.gov/8139659/)
12. G. Csankovszki, A. Nagy, R. Jaenisch, Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J. Cell Biol.* **153**, 773–784 (2001). pmid: [11352938](https://pubmed.ncbi.nlm.nih.gov/11352938/)
13. S. Bhatnagar *et al.*, Genetic and pharmacological reactivation of the mammalian inactive X chromosome. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12591–12598 (2014). doi: [10.1073/pnas.1413620111](https://doi.org/10.1073/pnas.1413620111); pmid: [25136103](https://pubmed.ncbi.nlm.nih.gov/25136103/)
14. W. Mak *et al.*, Reactivation of the paternal X chromosome in early mouse embryos. *Science* **303**, 666–669 (2004). doi: [10.1126/science.1092674](https://doi.org/10.1126/science.1092674); pmid: [14752160](https://pubmed.ncbi.nlm.nih.gov/14752160/)
15. M. Sugimoto, K. Abe, X chromosome reactivation initiates in nascent primordial germ cells in mice. *PLOS Genet.* **3**, e116 (2007). pmid: [17676999](https://pubmed.ncbi.nlm.nih.gov/17676999/)
16. J. Zhao, B. K. Sun, J. A. Erwin, J. J. Song, J. T. Lee, Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750–756 (2008). pmid: [18974356](https://pubmed.ncbi.nlm.nih.gov/18974356/)
17. K. Sarma *et al.*, ATRX directs binding of PRC2 to Xist RNA and Polycomb targets. *Cell* **159**, 869–883 (2014). doi: [10.1016/j.cell.2014.10.019](https://doi.org/10.1016/j.cell.2014.10.019); pmid: [25417162](https://pubmed.ncbi.nlm.nih.gov/25417162/)
18. Y. Jeon, J. T. Lee, YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146**, 119–133 (2011). doi: [10.1016/j.cell.2011.06.026](https://doi.org/10.1016/j.cell.2011.06.026); pmid: [21297874](https://pubmed.ncbi.nlm.nih.gov/21297874/)
19. Y. Hasegawa *et al.*, The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev. Cell* **19**, 469–476 (2010). doi: [10.1016/j.devcel.2010.08.006](https://doi.org/10.1016/j.devcel.2010.08.006); pmid: [20833368](https://pubmed.ncbi.nlm.nih.gov/20833368/)
20. A. Wutz, Gene silencing in X-chromosome inactivation: Advances in understanding facultative heterochromatin formation. *Nat. Rev. Genet.* **12**, 542–553 (2011). doi: [10.1038/nrg3035](https://doi.org/10.1038/nrg3035); pmid: [21765457](https://pubmed.ncbi.nlm.nih.gov/21765457/)
21. D. H. Lundgren, S. I. Hwang, L. Wu, D. K. Han, Role of spectral counting in quantitative proteomics. *Expert Rev. Proteomics* **7**, 39–53 (2010). doi: [10.1586/epr.09.69](https://doi.org/10.1586/epr.09.69); pmid: [20121475](https://pubmed.ncbi.nlm.nih.gov/20121475/)
22. L. Ting, R. Rad, S. P. Gygi, W. Haas, MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011). doi: [10.1038/nmeth.1714](https://doi.org/10.1038/nmeth.1714); pmid: [21963607](https://pubmed.ncbi.nlm.nih.gov/21963607/)
23. S. Schoettner *et al.*, Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *EMBO J.* **25**, 3110–3122 (2006). doi: [10.1038/sj.emboj.7601187](https://doi.org/10.1038/sj.emboj.7601187); pmid: [16763550](https://pubmed.ncbi.nlm.nih.gov/16763550/)
24. M. E. Blewitt *et al.*, SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. *Nat. Genet.* **40**, 663–669 (2008). doi: [10.1038/ng.142](https://doi.org/10.1038/ng.142); pmid: [18425126](https://pubmed.ncbi.nlm.nih.gov/18425126/)
25. J. T. Kung *et al.*, Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. *Mol. Cell* **57**, 361–375 (2015). doi: [10.1016/j.molcel.2014.12.006](https://doi.org/10.1016/j.molcel.2014.12.006); pmid: [25578877](https://pubmed.ncbi.nlm.nih.gov/25578877/)
26. M. H. Kagey *et al.*, Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010). doi: [10.1038/nature09380](https://doi.org/10.1038/nature09380); pmid: [20720539](https://pubmed.ncbi.nlm.nih.gov/20720539/)
27. J. R. Dixon *et al.*, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012). doi: [10.1038/nature11082](https://doi.org/10.1038/nature11082); pmid: [22495300](https://pubmed.ncbi.nlm.nih.gov/22495300/)
28. M. Merkenschlager, D. T. Odom, CTCF and cohesin: Linking gene regulatory elements with their targets. *Cell* **152**, 1285–1297 (2013). doi: [10.1016/j.cell.2013.02.029](https://doi.org/10.1016/j.cell.2013.02.029); pmid: [23498937](https://pubmed.ncbi.nlm.nih.gov/23498937/)
29. L. L. Hall, M. Byron, G. Pageau, J. B. Lawrence, AURKB-mediated effects on chromatin regulate binding versus release of Xist RNA to the inactive chromosome. *J. Cell Biol.* **186**, 491–507 (2009). doi: [10.1083/jcb.200811143](https://doi.org/10.1083/jcb.200811143); pmid: [19704020](https://pubmed.ncbi.nlm.nih.gov/19704020/)
30. V. Singh, P. Sharma, N. Capalash, DNA methyltransferase-1 inhibitors as epigenetic therapy for cancer. *Curr. Cancer Drug Targets* **13**, 379–399 (2013). doi: [10.2174/15680096113139990077](https://doi.org/10.2174/15680096113139990077); pmid: [23517596](https://pubmed.ncbi.nlm.nih.gov/23517596/)
31. M. E. Ashour, R. Attaya, S. F. El-Khamisy, Topoisomerase-mediated chromosomal break repair: An emerging player in many games. *Nat. Rev. Cancer* **15**, 137–151 (2015). doi: [10.1038/nrc3892](https://doi.org/10.1038/nrc3892); pmid: [25693836](https://pubmed.ncbi.nlm.nih.gov/25693836/)
32. S. F. Pinter *et al.*, Spreading of X chromosome inactivation via a hierarchy of defined Polycomb stations. *Genome Res.* **22**, 1864–1876 (2012). doi: [10.1101/gr.133751.111](https://doi.org/10.1101/gr.133751.111); pmid: [22948768](https://pubmed.ncbi.nlm.nih.gov/22948768/)
33. S. Lin, A. C. Ferguson-Smith, R. M. Schultz, M. S. Bartolomei, Nonallelic transcriptional roles of CTCF and cohesins at imprinted loci. *Mol. Cell Biol.* **31**, 3094–3104 (2011). doi: [10.1128/MCB.01449-10](https://doi.org/10.1128/MCB.01449-10); pmid: [21628529](https://pubmed.ncbi.nlm.nih.gov/21628529/)
34. W. Li *et al.*, Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516–520 (2013). doi: [10.1038/nature12210](https://doi.org/10.1038/nature12210); pmid: [23728302](https://pubmed.ncbi.nlm.nih.gov/23728302/)
35. J. M. Down *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](https://doi.org/10.1016/j.cell.2014.09.030); pmid: [25303531](https://pubmed.ncbi.nlm.nih.gov/25303531/)
36. J. M. Calabrese *et al.*, Site-specific silencing of regulatory elements as a mechanism of X inactivation. *Cell* **151**, 951–963 (2012). doi: [10.1016/j.cell.2012.10.037](https://doi.org/10.1016/j.cell.2012.10.037); pmid: [23178118](https://pubmed.ncbi.nlm.nih.gov/23178118/)
37. L. F. Zhang, K. D. Huynh, J. T. Lee, Perinuclear targeting of the inactive X during S phase: Evidence for a role in the maintenance of silencing. *Cell* **129**, 693–706 (2007). doi: [10.1016/j.cell.2007.03.036](https://doi.org/10.1016/j.cell.2007.03.036); pmid: [17512404](https://pubmed.ncbi.nlm.nih.gov/17512404/)
38. L. Carrel, H. F. Willard, X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404 (2005). doi: [10.1038/nature03479](https://doi.org/10.1038/nature03479); pmid: [15772666](https://pubmed.ncbi.nlm.nih.gov/15772666/)
39. J. B. Berletch, F. Yang, J. Xu, L. Carrel, C. M. Disteche, Genes that escape from X inactivation. *Hum. Genet.* **130**, 237–245 (2011). doi: [10.1007/s00439-011-1011-z](https://doi.org/10.1007/s00439-011-1011-z); pmid: [21614513](https://pubmed.ncbi.nlm.nih.gov/21614513/)
40. C. Feig, D. T. Odom, Cohesin's role as an active chromatin domain anchorage revealed. *EMBO J.* **32**, 3114–3115 (2013). doi: [10.1038/emboj.2013.248](https://doi.org/10.1038/emboj.2013.248); pmid: [24270571](https://pubmed.ncbi.nlm.nih.gov/24270571/)
41. C. T. Ong, V. G. Corces, CTCF: An architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014). doi: [10.1038/nrg3663](https://doi.org/10.1038/nrg3663); pmid: [24614316](https://pubmed.ncbi.nlm.nih.gov/24614316/)
42. M. Vietri Rudan *et al.*, Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports* **10**, 1297–1309 (2015). doi: [10.1016/j.celrep.2015.02.004](https://doi.org/10.1016/j.celrep.2015.02.004); pmid: [25732821](https://pubmed.ncbi.nlm.nih.gov/25732821/)
43. E. Splinter *et al.*, The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev.* **25**, 1371–1383 (2011). doi: [10.1101/gad.633311](https://doi.org/10.1101/gad.633311); pmid: [21690198](https://pubmed.ncbi.nlm.nih.gov/21690198/)
44. E. P. Nora *et al.*, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012). doi: [10.1038/nature11049](https://doi.org/10.1038/nature11049); pmid: [22495304](https://pubmed.ncbi.nlm.nih.gov/22495304/)
45. T. Nagano *et al.*, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013). doi: [10.1038/nature12593](https://doi.org/10.1038/nature12593); pmid: [24067610](https://pubmed.ncbi.nlm.nih.gov/24067610/)
46. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014). doi: [10.1016/j.cell.2014.11.021](https://doi.org/10.1016/j.cell.2014.11.021); pmid: [25497547](https://pubmed.ncbi.nlm.nih.gov/25497547/)
47. S. Sun *et al.*, Jpx RNA activates Xist by evicting CTCF. *Cell* **153**, 1537–1551 (2013). doi: [10.1016/j.cell.2013.05.028](https://doi.org/10.1016/j.cell.2013.05.028); pmid: [23791181](https://pubmed.ncbi.nlm.nih.gov/23791181/)
48. A. Castello *et al.*, Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012). doi: [10.1016/j.cell.2012.04.031](https://doi.org/10.1016/j.cell.2012.04.031); pmid: [22658674](https://pubmed.ncbi.nlm.nih.gov/22658674/)
49. S. C. Kwon *et al.*, The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1122–1130 (2013). doi: [10.1038/nsmb.2638](https://doi.org/10.1038/nsmb.2638); pmid: [23912277](https://pubmed.ncbi.nlm.nih.gov/23912277/)
50. G. C. McAlister *et al.*, Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* **84**, 7469–7478 (2012). doi: [10.1021/ac301572t](https://doi.org/10.1021/ac301572t); pmid: [22880955](https://pubmed.ncbi.nlm.nih.gov/22880955/)
51. A. Thompson *et al.*, Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003). doi: [10.1021/ac0262560](https://doi.org/10.1021/ac0262560); pmid: [12713048](https://pubmed.ncbi.nlm.nih.gov/12713048/)
52. A. C. Tolonen, W. Haas, Quantitative proteomics using reductive dimethylation for stable isotope labeling. *J. Vis. Exp.* (89) (2014). pmid: [25045933](https://pubmed.ncbi.nlm.nih.gov/25045933/)
53. G. C. McAlister *et al.*, MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150–7158 (2014). doi: [10.1021/ac502040v](https://doi.org/10.1021/ac502040v); pmid: [24927332](https://pubmed.ncbi.nlm.nih.gov/24927332/)
54. M. P. Weekes *et al.*, Quantitative temporal viromics: An approach to investigate host-pathogen interaction. *Cell* **157**, 1460–1472 (2014). doi: [10.1016/j.cell.2014.04.028](https://doi.org/10.1016/j.cell.2014.04.028); pmid: [24906157](https://pubmed.ncbi.nlm.nih.gov/24906157/)
55. J. K. Eng, A. L. McCormack, J. R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994). doi: [10.1016/1044-0305\(94\)80016-2](https://doi.org/10.1016/1044-0305(94)80016-2); pmid: [24226387](https://pubmed.ncbi.nlm.nih.gov/24226387/)
56. J. E. Elias, S. P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007). doi: [10.1038/nmeth1019](https://doi.org/10.1038/nmeth1019); pmid: [17327847](https://pubmed.ncbi.nlm.nih.gov/17327847/)
57. E. L. Huttlin *et al.*, A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010). pmid: [21183079](https://pubmed.ncbi.nlm.nih.gov/21183079/)
58. A. K. Hadjantonakis, L. L. Cox, P. P. Tam, A. Nagy, An X-linked GFP transgene reveals unexpected paternal X-chromosome activity in trophoblastic giant cells of the mouse placenta. *Genesis* **29**, 133–140 (2001). doi: [10.1002/gene.1016](https://doi.org/10.1002/gene.1016); pmid: [11252054](https://pubmed.ncbi.nlm.nih.gov/11252054/)
59. S. Heinz *et al.*, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required

- for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010). doi: [10.1016/j.molcel.2010.05.004](https://doi.org/10.1016/j.molcel.2010.05.004); pmid: [20513432](https://pubmed.ncbi.nlm.nih.gov/20513432/)
60. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010). doi: [10.1093/bioinformatics/btp616](https://doi.org/10.1093/bioinformatics/btp616); pmid: [19910308](https://pubmed.ncbi.nlm.nih.gov/19910308/)
61. E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009). doi: [10.1126/science.1181369](https://doi.org/10.1126/science.1181369); pmid: [19815776](https://pubmed.ncbi.nlm.nih.gov/19815776/)
62. E. Yildirim, R. I. Sadreyev, S. F. Pinter, J. T. Lee, X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. *Nat. Struct. Mol. Biol.* **19**, 56–61 (2011). doi: [10.1038/nsmb.2195](https://doi.org/10.1038/nsmb.2195); pmid: [22139016](https://pubmed.ncbi.nlm.nih.gov/22139016/)
63. Y. C. Lin *et al.*, Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat. Immunol.* **13**, 1196–1204 (2012). doi: [10.1038/ni.2432](https://doi.org/10.1038/ni.2432); pmid: [23064439](https://pubmed.ncbi.nlm.nih.gov/23064439/)
64. S. Selvaraj, J. R. Dixon, V. Bansal, B. Ren, Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013). doi: [10.1038/nbt.2728](https://doi.org/10.1038/nbt.2728); pmid: [24185094](https://pubmed.ncbi.nlm.nih.gov/24185094/)
65. P. A. Knight, D. Ruiz, A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013). doi: [10.1093/imanum/drs019](https://doi.org/10.1093/imanum/drs019)
66. M. Imakaev *et al.*, Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012). doi: [10.1038/nmeth.2148](https://doi.org/10.1038/nmeth.2148); pmid: [22941365](https://pubmed.ncbi.nlm.nih.gov/22941365/)

ACKNOWLEDGMENTS

We thank S. Gygi for access to computational resources for proteomic analysis and all members of the Lee and Haas laboratories for valuable discussions. This work was supported by grants from NIH (R01-DA-38695 and R03-MH97478), the Rett Syndrome Research Trust, and the International Rett Syndrome Foundation to J.T.L.; a National Science Foundation predoctoral

award to J.E.F.; and the MGH Fund for Medical Discovery to H.S. J.T.L. is an Investigator of the HHMI. J.T.L., A.M., J.E.F., C.W., and the Massachusetts General Hospital have filed patent applications (USSN 62/144,219 and 62/168,528) that relate to leveraging the Xist interactome to reactivate the Xi. The GEO accession code for data in the paper is GSE67516. RNA-seq, ChIP-seq, and HiC-seq data are deposited in GEO.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/aab2276/suppl/DC1

Supplementary Text

Figs. S1 to S15

Tables S1 to S3

References (67, 68)

30 March 2015; accepted 4 June 2015

Published online 18 June 2015;

10.1126/science.aab2276

RESEARCH ARTICLE

MAGNETISM

Blowing magnetic skyrmion bubbles

Wanjuan Jiang,¹ Pramey Upadhyaya,² Wei Zhang,¹ Guoqiang Yu,²
 M. Benjamin Jungfleisch,¹ Frank Y. Fradin,¹ John E. Pearson,¹ Yaroslav Tserkovnyak,³
 Kang L. Wang,² Olle Heinonen,^{1,4,5,6} Suzanne G. E. te Velthuis,¹ Axel Hoffmann^{1*}

The formation of soap bubbles from thin films is accompanied by topological transitions. Here we show how a magnetic topological structure, a skyrmion bubble, can be generated in a solid-state system in a similar manner. Using an inhomogeneous in-plane current in a system with broken inversion symmetry, we experimentally “blow” magnetic skyrmion bubbles from a geometrical constriction. The presence of a spatially divergent spin-orbit torque gives rise to instabilities of the magnetic domain structures that are reminiscent of Rayleigh-Plateau instabilities in fluid flows. We determine a phase diagram for skyrmion formation and reveal the efficient manipulation of these dynamically created skyrmions, including depinning and motion. The demonstrated current-driven transformation from stripe domains to magnetic skyrmion bubbles could lead to progress in skyrmion-based spintronics.

Magnetic skyrmions are topological spin textures that can be stabilized by Dzyaloshinskii-Moriya interactions (DMIs) (1–9) in chiral bulk magnets such as MnSi, FeGe, etc. Owing to their distinct vortex-like spin texture, skyrmions exhibit many fascinating features, including emergent electromagnetic fields, which enable their efficient manipulation (4, 5, 8–10). A particularly technologically interesting property is that skyrmions can be driven by a spin-transfer torque mechanism at a very low current density, which has been demonstrated at cryogenic temperatures (4, 5, 8, 10). Besides bulk chiral magnetic interactions, the interfacial symmetry breaking in heavy metal/ultrathin ferromagnet/insulator (HM/F/I) trilayers introduces an interfacial DMI (11–14) between neighboring atomic spins, which stabilizes Néel walls (cycloidal rotation of the magnetization direction) with a fixed chirality over the Bloch walls (spiral rotation of the magnetization direction) (15–20). This is expected to result in the formation of skyrmions with a “hedgehog” configuration (14, 18, 21–25). This commonly accessible material system exhibits spin Hall effects from heavy metals with strong spin-orbit interactions (26), which in turn give rise to well-defined spin-orbit torques (SOTs) (17, 19, 27–29) that can control magnetization dynamics efficiently. However, it has been experimentally challenging to use the electric current and/or its induced SOTs (8, 21, 23, 24, 27, 30–32) for dynamically creating and/or manipulating hedgehog skyrmions. Here we address that issue.

Central to this work is how electric currents can manipulate a chiral magnetic domain wall (DW); that is, the chirality of the magnetization rotation (as shown in Fig. 1A) is identical for every DW. This fixed chirality is stabilized by the interfacial DMI (17–19, 21, 28). In HM/F/I heterostructures, the current flowing through the heavy metal generates a transverse vertical spin current attributable to the spin Hall effect (27), which results in spin accumulation at the interface with the ferromagnetic layer. This spin accumulation gives rise to a SOT acting on the chiral DW (Fig. 1A). The resultant effective spin Hall field can be expressed as (17–19, 27)

$$\vec{B}_{\text{sh}} = B_{\text{sh}}^0 [\hat{m} \times (\hat{z} \times \hat{j}_e)] \quad (1)$$

where \hat{m} is the magnetization unit vector, \hat{z} is the unit vector normal to the film plane, and \hat{j}_e is the direction of electron particle flux. Here B_{sh}^0 can be written as $(\hbar/2|e|) \cdot (\theta_{\text{sh}} J_c / t_f M_s)$, where $\hbar/2$ is the spin of an electron (and \hbar is Planck's constant h divided by 2π), e is the charge of an electron, t_f is the thickness of the ferromagnetic layer, and M_s is the saturation (volume) magnetization. The spin Hall angle $\theta_{\text{sh}} = J_s / J_c$ is defined by the ratio between spin current density (J_s) and charge current density (J_c). For homogeneous current flow along the x axis (Fig. 1B), a chiral SOT enables efficient DW motion (17–19). In the case of a stripe domain with a chiral DW (Fig. 1B), the symmetry of Eq. 1 leads to a vanishing torque on the side walls parallel to the current, and therefore only the end of the stripe domain is moved. If the opposite end is pinned, this results in an elongation of the stripe.

The situation becomes more complex when the stripe domain is subjected to an inhomogeneous current flow. This can be achieved by introducing a geometrical constriction into a current-carrying trilayer wire (Fig. 1C). Such a constriction results in an additional current component along the y axis: j_y around the narrow neck (Fig. 1D). The total current j is spatially convergent (or divergent) to

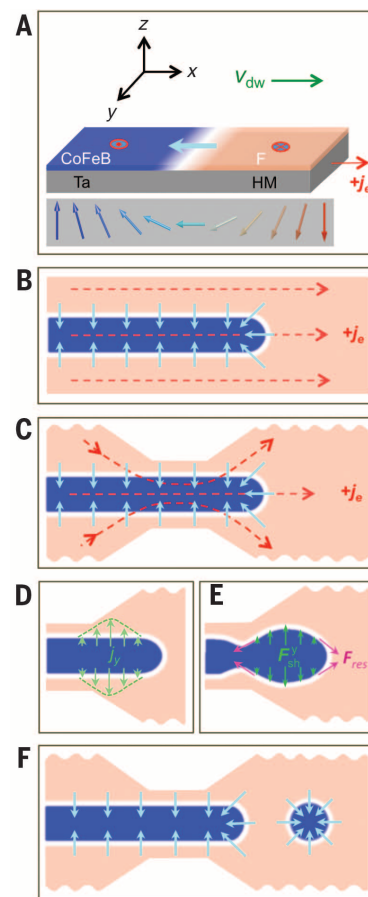


Fig. 1. Schematic of the transformation of stripe domains into magnetic skyrmion bubbles. (A)

Infinitesimal section of a chiral DW in a ferromagnet (F)/heavy metal (HM) bilayer illustrating the relation between local magnetization vectors and the SOT-induced chiral DW motion of velocity V_{dw} in a device with a homogeneous electron current flow j_e along the $+x$ axis. Blue corresponds to upward orientation of magnetization, whereas orange represents the downward orientation of magnetization. The bottom panel illustrates the magnetization directions inside of the Néel wall. (B) Top view of a trilayer device. The blue region is a stripe-shaped domain. Light blue arrows show the in-plane magnetization direction of the DW [as shown in the bottom panel of (A)] and indicate that the domain has left-handed chirality. The red arrows correspond to the current distribution. (C) Introducing a geometrical constriction into the device gives rise to an inhomogeneous current distribution, which generates a flow along the y axis (j_y) around the narrow neck. This current distribution is spatially divergent to the right and convergent to the left of the constriction. The y component of the current distribution is highlighted in (D). This introduces an effective spin Hall force (F_{sh}^y) along the y axis that (E) locally expands the stripe domain on the right side. (F) Once the expansion approaches a critical point, the resultant restoring forces (F_{res}) associated with the surface tension of the DWs are no longer able to maintain the shape, and the stripe domains break into circular bubble domains, resulting in the formation of synthetic Néel skyrmions.

¹Materials Science Division, Argonne National Laboratory, Lemont, IL 60439, USA. ²Device Research Laboratory, Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA. ³Department of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA. ⁴Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA. ⁵Northwestern-Argonne Institute of Science and Engineering, Northwestern University, Evanston, IL 60208, USA. ⁶Computation Institute, University of Chicago, Chicago, IL 60637, USA.

*Corresponding author. E-mail: hoffmann@anl.gov

the left (or right) of the constriction (33). Consequently, inhomogeneous effective forces (F_{sh}^y) on the DWs (caused by the spin Hall field) are created along the y axis; these forces act to expand the end of the domain (Fig. 1E). As the domain end continually expands its radius, the surface tension in the DW (resulting from the increasing DW energy determined by the combination of exchange and anisotropy fields) increases (34), which results in breaking the stripes into circular domains (Fig. 1F).

This process resembles how soap bubbles develop out of soap films upon blowing air through a straw, or how liquid droplets form in fluid flow jets (35). Because of the interfacial DMI in the present system, the spin structures of the newly formed circular domains maintain a well-defined (left-handed) chirality (13, 14, 23, 24). Once formed, these created synthetic hedgehog (Néel) skyrmions (14, 23) are stable due to topological protection and move very efficiently following the current direction, a process that can be described on the basis of a modified Thiele equation (36). The dynamic skyrmion conversion could, in principle, happen at the other side of the device, where the spatially convergent current compresses stripe domains. However, sizeable currents and SOTs are required to compensate the enhanced (repulsive) dipolar interaction. The proposed mechanism differs from a recent theoretical proposal with similar geometry, in which skyrmions are formed from the coalescence of two independent DWs extending over the full width of a narrow constriction at a current density $\sim 10^8$ A/cm² (32). For repeated skyrmion generation, this latter mechanism requires a continuous generation of paired DWs in the constriction, which is inconsistent with the experimental observations described below.

Transforming chiral stripe domains into skyrmions

We demonstrated this idea experimentally with a Ta(5 nm)/Co₂₀Fe₆₀B₂₀(CoFeB)(1.1 nm)/TaO_x(3 nm) trilayer grown by magnetron sputtering (37, 38) and patterned into constricted wires via photolithography and ion milling (33). The wires have a width of 60 μ m with a 3- μ m-wide and 20- μ m-long geometrical constriction in the center. Our devices are symmetrically designed across the narrow neck to maintain balanced demagnetization energy. A polar magneto-optical Kerr effect (MOKE) microscope in a differential mode (39) was used for dynamic imaging experiments at room temperature. Before applying a current, the sample was first saturated at positive magnetic fields and subsequently a perpendicular magnetic field of $B_{\perp} = +0.5$ mT was applied; sparse magnetic stripe and bubble domains prevail at both sides of the wire (Fig. 2A). The lighter area corresponds to positive perpendicular magnetization orientation, and the darker area corresponds to negative orientation, respectively.

In contrast to the initial magnetic domain configuration, after passing a 1-s single pulse of current density $j_e = +5 \times 10^5$ A/cm² (normalized by the width of the device: 60 μ m), it is observed that the stripe domains started to migrate, subsequently forming extended stripe domains on the left side. These domains were mostly aligned with the charge current

flow and converged at the left side of constriction. The stripes were transformed into skyrmion bubbles immediately after passing through the constriction (Fig. 2B). These dynamically created skyrmions, varying in size between 700 nm and 2 μ m (depending on the strength of the external magnetic field), are stable and do not decay on the scale of a typical laboratory testing period (at least 8 hours). The size of the skyrmions is determined by the interplay between Zeeman, magnetostatic interaction and interfacial DMI. In the presence of a constant electron current density of $j_e = +5 \times 10^5$ A/cm², these skyrmions are created with a high speed close to the central constriction and destroyed at the end of the wire. Capturing the transformation dynamics of skyrmions from stripe domains is beyond the temporal resolution of the present setup. Reproducible generation of skyrmions is demonstrated by repeating pulsed experiments several times (33). The left side of the device remains mainly in the labyrinthine stripe domain state after removing the pulse current, which indicates that both skyrmion bubbles and stripe domains are metastable.

When the polarity of the charge current is reversed to $j_e = -5 \times 10^5$ A/cm², the skyrmions are formed at the left side of the device (Fig. 2, C and D). This directional dependence indicates that the spatially divergent current and SOT, determined by the geometry of the device, are most likely responsible for slicing stripe DWs into magnetic

skyrmion bubbles, qualitatively consistent with the schematic presented in Fig. 1.

At a negative magnetic field $B_{\perp} = -0.5$ mT and current $j_e = +5 \times 10^5$ A/cm² (Fig. 2, E and F), a reversed contrast, resulting from opposite inner and outer magnetization orientations, is observed compared with positive fields. We varied the external magnetic field and charge current density systematically and determined the phase diagram for skyrmion formation shown in Fig. 2G. A large population of synthetic skyrmions is found only in the shadowed region, whereas in the rest of phase diagram the initial domain configurations remain either stationary or flowing smoothly, depending on the strength of current density, as discussed below. This phase diagram is independent of pulse duration for pulses longer than 1 μ s. No creation of skyrmions in a regular-shaped device with a homogeneous current flow (as illustrated in Fig. 1B) is observed up to a current density of $j_e = +5 \times 10^6$ A/cm².

Capturing the transformational dynamics

The conversion from chiral stripe domains into magnetic skyrmions can be captured by decreasing the driving current, which slows down the transformational dynamics. Figure 3, A to D, shows the dynamics for a constant current density of $j_e = +6.4 \times 10^4$ A/cm² at $B_{\perp} = +0.46$ mT. The original (disordered) labyrinthine domains on the left side squeeze to pass through the constriction (Fig.

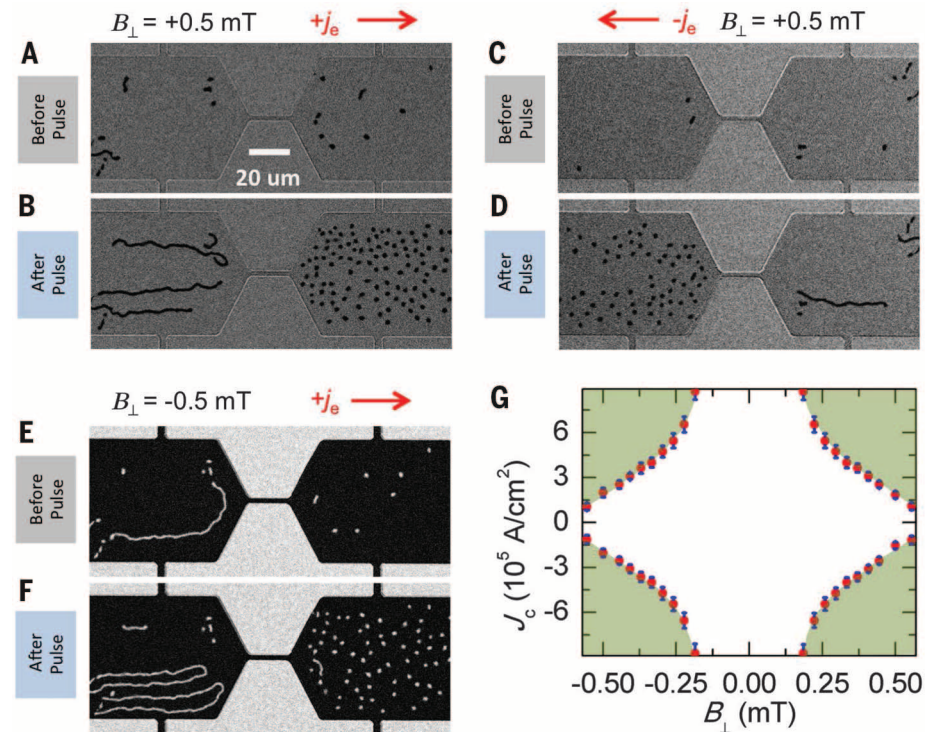


Fig. 2. Experimental generation of magnetic skyrmions. (A) Sparse irregular domain structures are observed at both sides of the device at a perpendicular magnetic field of $B_{\perp} = +0.5$ mT. (B) Upon passing a current of $j_e = +5 \times 10^5$ A/cm² through the device, the left side of the device develops predominantly elongated stripe domains, whereas the right side converts into dense skyrmion bubbles. (C and D) By reversing the current direction to $j_e = -5 \times 10^5$ A/cm², the dynamically created skyrmions are forming at the left side of the device. (E and F) Changing the polarity of the external magnetic field reverses the internal and external magnetization of these skyrmions. (G) Phase diagram for skyrmion formation. The shaded area indicates field-current combinations that result in the persistent generation of skyrmions after each current pulse.

3B). The stripe domains become unstable after passing through the constriction and are eventually converted into skyrmions on the right side of the device, as shown in Fig. 3, C and D. This can be seen in more detail in the supplementary MOKE movies (movies S1 and S2). Because the x component of the current results in an efficient motion of DWs, the skyrmion formation can occur away from the constriction. The synthetic skyrmions do not merge into stripe domains and, in fact, repel each other, indicating their topological protection as well as magnetostatic interactions.

Some important features should be noticed. There exists a threshold current $j_{e-sk} = \pm 6 \times 10^4$ A/cm² for persistently generating skyrmion bubbles from stripe domains for pulses longer than 1 μ s. Above this current, the enhanced spin-orbit torques produce the instability of the DWs, which results in the continuous formation of skyrmions. The present geometry for skyrmion generation is very efficient, resulting in the observed threshold current that is three orders of magnitude smaller than suggested by previous simulation studies

(10^7 to 10^8 A/cm²) in MnSi thin films with a bulk DMI where the driving mechanism is the conventional spin-transfer torque (30). Below this threshold for continuous skyrmion generation, there is a threshold depinning current $j_{e-st} = \pm 4.1 \times 10^4$ A/cm² that produces a steady motion of stripe domains. The force (pressure) on the stripe from SOT at this current exceeds the one required to maintain its shape. When $j_{e-st} < j_e < j_{e-sk}$, the stripe domains are moving smoothly through the constriction and prevail at both sides of devices, with just the occasional formation of skyrmions.

Depinning and motion of synthesized $S = 1$ skyrmions

The magnetic skyrmion bubbles discussed so far have a topological charge given by the skyrmion number $S = 1$, as is determined by wrapping the unit magnetization vector (\mathbf{m}) over the sphere $S = \frac{1}{4\pi} \int \mathbf{m} \cdot (\partial_x \mathbf{m} \times \partial_y \mathbf{m}) dx dy$ (1, 8). These $S = 1$ synthetic skyrmions move because of the opposite direction of effective SOTs on the opposite sides of the skyrmion (Fig. 3E). Following the

initialization by a current pulse $j_e = +5 \times 10^5$ A/cm² (which is larger than the threshold current j_{e-sk} for generating skyrmions), we studied the efficient depinning and motion of synthetic skyrmions (Fig. 3, F to I) at $B_{\perp} = -0.5$ mT. At the current density $j_e = +3 \times 10^4$ A/cm², there is no migration of stripe domains through the constriction (hence an absence of newly formed skyrmions). However, it is clear that the previously generated skyrmions at the right side of the device are gradually moving away following the electron flow direction. During the motion, no measurable distortion of these synthetic skyrmions is observed within the experimental resolution, consistent with the well-defined chirality of the skyrmion bubble. The average velocity ($\bar{v} = \ell / \Delta t$) is determined by dividing the displacement (ℓ) with the total time period (Δt). For the present current density, the motion of synthetic skyrmion is stochastic and influenced by random pinning with an average velocity of ≈ 10 μ m/s, the current dependence of which is summarized in Fig. 3J. The ratio of the velocity to the applied current is comparable to what is observed for the chiral DW motion in the related systems (17, 19).

Current characteristics of $S = 0$ magnetic bubbles

Because of the competition between long-range dipolar and short-range exchange interaction, a system with a weak perpendicular magnetic anisotropy undergoes a spin reorientation transition with in-plane magnetic fields that is typified by a stripe-to-bubble domain phase transition (39, 40). Such an in-plane field-induced bubble state is established by sweeping the magnetic field from $B_{\parallel} = +100$ to $+10$ mT. Current-driven characteristics of the in-plane field-induced magnetic bubbles are in stark contrast to the mobile magnetic skyrmions generated from SOTs. These bubbles shrink and vanish in the presence of a positive electron current density (Fig. 4, A to E) or elongate and transform into stripe domains in the presence of negative electron current density (Fig. 4, F to J). Such a distinct difference directly indicates the different spin structures surrounding these field-induced bubbles and, therefore, the different skyrmion numbers.

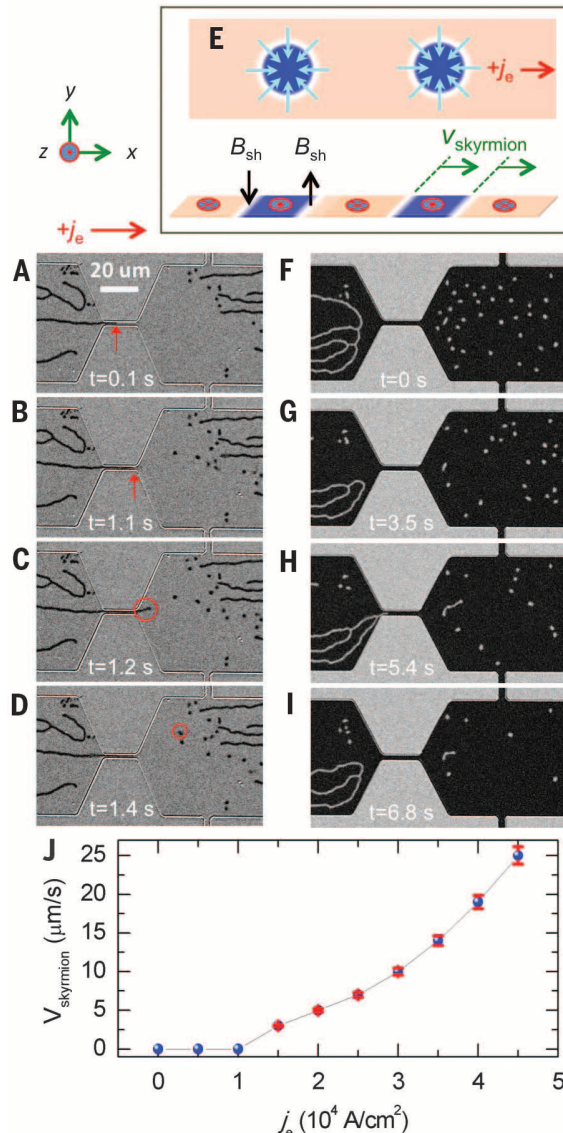
For the in-plane field-induced magnetic bubbles, because the spin structures of DWs follow the external magnetic fields (18, 41, 42) (Fig. 4K), the corresponding skyrmion number is $S = 0$. Because of the same direction of the effective spin Hall fields given by the reversed DW orientations, topologically trivial $S = 0$ magnetic bubbles experience opposite forces on the DWs at opposite ends. This leads to either a shrinking or elongation of the bubbles, depending on the direction of currents, which is consistent with our experimental observation. This also explains the in-plane current-induced perpendicular magnetization switching in the presence of in-plane fields (27, 42).

Perspectives

Recent experimental efforts toward creating individual magnetic skyrmions use either tunneling current from a low-temperature spin-polarized scanning tunneling microscope (43) or geometrical confinement via sophisticated nanopatterning

Fig. 3. Capturing the transformational dynamics from stripe domains to skyrmions and motion of skyrmions. (A to D)

At a constant current density $j_e = +6.4 \times 10^4$ A/cm² and $B_{\perp} = +0.46$ mT, the disordered stripe domains are forced to pass through the constriction and are eventually converted into skyrmions at the right side of the device. Red circles highlight the resultant newly formed skyrmions. (E) Illustration of the effective spin Hall field acting on these dynamically created skyrmions; the direction of motion follows the electron current. (F to I) Efficient motion of these skyrmions for a current density $j_e = +3 \times 10^4$ A/cm². (F) First, a 1-s-long single pulse $j_e = +5 \times 10^5$ A/cm² initializes the skyrmion state. (G to I) Subsequently, smaller currents (below the threshold current to avoid generating additional skyrmions through the constriction) are used to probe the current-velocity relation. These skyrmions are migrating stochastically and moving out of the field of view. See supplementary MOKE movies S1, S2, and S4 for the corresponding temporal dynamics. (J) The current-velocity dependence of skyrmions is acquired by studying ≈ 20 skyrmions via averaging their velocities by dividing the total displacement with the total time period.



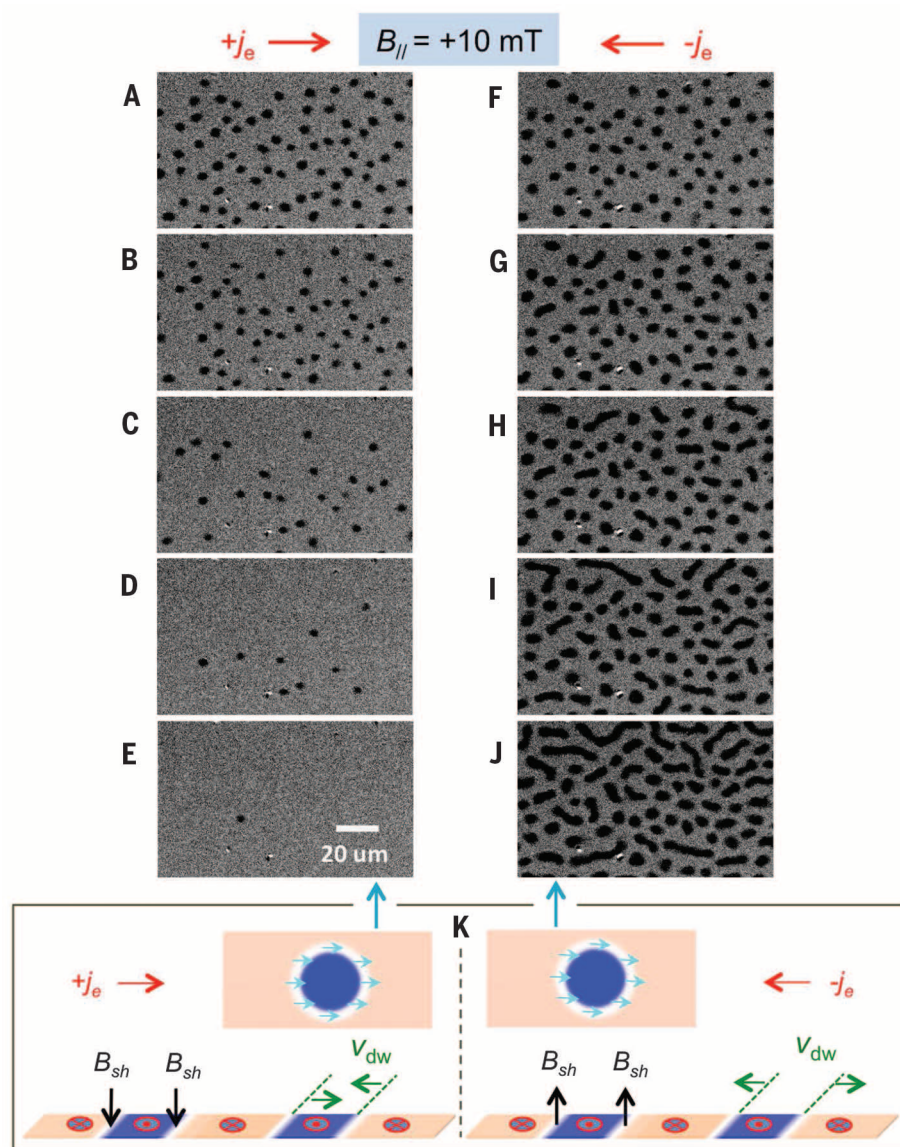


Fig. 4. Absence of motion for the in-plane magnetic fields stabilized $S = 0$ magnetic bubbles. (A to E) (A) In-plane magnetic field-induced bubbles are created by first saturating at in-plane field $B_{||} = +100$ mT and subsequently decreasing to $B_{||} = +10$ mT. Depending on the direction of the current, these magnetic bubbles either shrink or expand. (A to E) Shrinking bubbles are observed upon increasing the current density from $j_e = +5 \times 10^4$ to $+2.5 \times 10^5$ A/cm² in steps of 5×10^4 A/cm². (F to J) Expansion of bubbles is revealed for currents from $j_e = -0.5 \times 10^5$ to -2.5×10^5 A/cm² in steps of 5×10^4 A/cm². (K) These results are linked to the different spin textures (namely, $S = 0$ skyrmion bubbles) that were stabilized along the DW by the in-plane magnetic fields, which lead to different orientations of the effective spin Hall fields and different directions of DW motion, as illustrated.

(44–46). Our results demonstrate that spatially divergent current-induced SOTs can be an effective way for dynamically generating mobile magnetic skyrmions at room temperature in commonly accessible material systems. The size of these synthetic skyrmions could be scaled down by properly engineering the material-specific parameters that control the various competing interactions in magnetic nanostructures (23, 24, 47). We expect that similar instabilities will be generated from divergent charge current flows. Whereas the mechanism for synthetic skyrmion generation can be qualitatively linked to the spatially divergent spin Hall spin torque, a comprehensive understanding of this dy-

namical conversion, particularly at the picosecond or nanosecond time scale where the intriguing magnetization dynamics occurs, requires further experimental and theoretical investigations. Spatially divergent SOT-driven structures also offer a readily accessible model system for studying topological transitions and complex flow instabilities (35), where the parameters governing the flow, such as surface tension, can be systematically tuned by the magnetic interactions. At the same time, this dynamic approach for skyrmion generation in the near future could enable the demonstration of advanced skyrmionic device concepts; for example, functional skyrmion racetrack memory (14, 23, 36, 48).

REFERENCES AND NOTES

- U. K. Röbler, A. N. Bogdanov, C. Pfleiderer, *Nature* **442**, 797–801 (2006).
- S. Mühlbauer et al., *Science* **323**, 915–919 (2009).
- X. Z. Yu et al., *Nature* **465**, 901–904 (2010).
- F. Jonietz et al., *Science* **330**, 1648–1651 (2010).
- X. Z. Yu et al., *Nat. Mater.* **10**, 106–109 (2011).
- S. Seki, X. Z. Yu, S. Ishiwata, Y. Tokura, *Science* **336**, 198–201 (2012).
- H.-B. Braun, *Adv. Phys.* **61**, 1–116 (2012).
- N. Nagaosa, Y. Tokura, *Nat. Nanotechnol.* **8**, 899–911 (2013).
- S. Z. Lin, C. Reichhardt, C. D. Batista, A. Saxena, *Phys. Rev. Lett.* **110**, 207202 (2013).
- J. Zang, M. Mostovoy, J. H. Han, N. Nagaosa, *Phys. Rev. Lett.* **107**, 136804 (2011).
- M. Bode et al., *Nature* **447**, 190–193 (2007).
- S. Heinze et al., *Nat. Phys.* **7**, 713–718 (2011).
- A. Thiaville, S. Rohart, E. Jue, V. Cros, A. Fert, *Europhys. Lett.* **100**, 57002 (2012).
- A. Fert, V. Cros, J. Sampaio, *Nat. Nanotechnol.* **8**, 152–156 (2013).
- G. Chen et al., *Nat. Commun.* **4**, 2671 (2013).
- G. Chen et al., *Phys. Rev. Lett.* **110**, 177204 (2013).
- S. Emori, U. Bauer, S. M. Ahn, E. Martinez, G. S. Beach, *Nat. Mater.* **12**, 611–616 (2013).
- S. Emori et al., *Phys. Rev. B* **90**, 184427 (2014).
- K. S. Ryu, L. Thomas, S. H. Yang, S. Parkin, *Nat. Nanotechnol.* **8**, 527–533 (2013).
- O. Boulle et al., *Phys. Rev. Lett.* **111**, 217203 (2013).
- N. Perez et al., *Appl. Phys. Lett.* **104**, 092403 (2014).
- K. W. Kim, H. W. Lee, K. J. Lee, M. D. Stiles, *Phys. Rev. Lett.* **111**, 216601 (2013).
- J. Sampaio, V. Cros, S. Rohart, A. Thiaville, A. Fert, *Nat. Nanotechnol.* **8**, 839–844 (2013).
- S. Rohart, A. Thiaville, *Phys. Rev. B* **88**, 184422 (2013).
- B. Dupé, M. Hoffmann, C. Paillard, S. Heinze, *Nat. Commun.* **5**, 4030 (2014).
- A. Hoffmann, *IEEE Trans. Magn.* **49**, 5172–5193 (2013).
- L. Liu et al., *Science* **336**, 555–558 (2012).
- A. V. Khvalkovskiy et al., *Phys. Rev. B* **87**, 020402 (2013).
- I. M. Miron et al., *Nature* **476**, 189–193 (2011).
- J. Iwasaki, M. Mochizuki, N. Nagaosa, *Nat. Nanotechnol.* **8**, 742–747 (2013).
- Y. Tchoe, J. H. Han, *Phys. Rev. B* **85**, 174416 (2012).
- Y. Zhou, M. Ezawa, *Nat. Commun.* **5**, 4652 (2014).
- Supplementary materials are available on Science Online.
- A. P. Malozemoff, J. C. Slonczewski, *Magnetic Domain Walls in Bubble Materials* (Academic Press, New York, 1979).
- J. Eggers, *Rev. Mod. Phys.* **69**, 865–930 (1997).
- R. Tomasello et al., *Sci. Rep.* **4**, 6784 (2014).
- G. Yu et al., *Nat. Nanotechnol.* **9**, 548–554 (2014).
- G. Q. Yu et al., *Phys. Rev. B* **89**, 104421 (2014).
- A. Hubert, R. Schäfer, *Magnetic Domains: The Analysis of Magnetic Microstructures* (Springer, Berlin, Heidelberg, New York, 2008).
- J. Choi et al., *Phys. Rev. Lett.* **98**, 207205 (2007).
- J. H. Franken, M. Herps, H. J. M. Swagten, B. Koopmans, *Sci. Rep.* **4**, 5248 (2014).
- O. J. Lee et al., *Phys. Rev. B* **89**, 024418 (2014).
- N. Romming et al., *Science* **341**, 636–639 (2013).
- L. Sun et al., *Phys. Rev. Lett.* **110**, 167201 (2013).
- J. Li et al., *Nat. Commun.* **5**, 4704 (2014).
- B. F. Miao et al., *Phys. Rev. B* **90**, 174411 (2014).
- A. Hrabec et al., *Phys. Rev. B* **90**, 020402(R) (2014).
- S. S. P. Parkin, M. Hayashi, L. Thomas, *Science* **320**, 190–194 (2008).

ACKNOWLEDGMENTS

Work carried out at Argonne National Laboratory was supported by the U.S. Department of Energy (DOE), Office of Science, Materials Science and Engineering Division. Lithography was carried out at the Center for Nanoscale Materials, an Office of Science user facility, which is supported by the DOE, Office of Science, Basic Energy Sciences, under contract no. DE-AC02-06CH11357. Work performed at the University of California, Los Angeles, was partially supported by the NSF Nanosystems Engineering Research Center for Translational Applications of Nanoscale Multiferric Systems. We thank I. Martin and I. Aronson for insightful discussion.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/283/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
References
Movies S1 to S5

26 October 2014; accepted 28 May 2015
Published online 11 June 2015
10.1126/science.aaa1442

REPORTS

HEAVY FERMIONS

Unconventional Fermi surface in an insulating state

B. S. Tan,¹ Y.-T. Hsu,¹ B. Zeng,² M. Ciomaga Hatnean,³ N. Harrison,⁴ Z. Zhu,⁴ M. Hartstein,¹ M. Kiourlappou,¹ A. Srivastava,¹ M. D. Johannes,⁵ T. P. Murphy,² J.-H. Park,² L. Balicas,² G. G. Lonzarich,¹ G. Balakrishnan,³ Suchitra E. Sebastian^{1*}

Insulators occur in more than one guise; a recent finding was a class of topological insulators, which host a conducting surface juxtaposed with an insulating bulk. Here, we report the observation of an unusual insulating state with an electrically insulating bulk that simultaneously yields bulk quantum oscillations with characteristics of an unconventional Fermi liquid. We present quantum oscillation measurements of magnetic torque in high-purity single crystals of the Kondo insulator SmB_6 , which reveal quantum oscillation frequencies characteristic of a large three-dimensional conduction electron Fermi surface similar to the metallic rare earth hexaborides such as PrB_6 and LaB_6 . The quantum oscillation amplitude strongly increases at low temperatures, appearing strikingly at variance with conventional metallic behavior.

Kondo insulators, a class of materials positioned close to the border between insulating and metallic behavior, provide fertile ground for unusual physics (1–14). This class of strongly correlated materials is thought to be characterized by a ground state

with a small energy gap at the Fermi energy owing to the collective hybridization of conduction and f -electrons. The observation of quantum oscillations has traditionally been associated with a Fermi liquid state; here, we present the surprising measurement of quantum oscillations in the Kondo insulator SmB_6 (15) that originate from a large three-dimensional Fermi surface occupying half the Brillouin zone and strongly resembling the conduction electron Fermi surface in the metallic rare earth hexaborides (16, 17). Our measurements in SmB_6 reveal a dramatic departure from conventional metallic Lifshitz-Kosevich behavior (18); instead of the expected saturation at low temperatures, a striking increase is

observed in the quantum oscillation amplitude at low temperatures.

Single crystals of SmB_6 used in the present study were grown by means of the image furnace technique (19) in order to achieve high purities as characterized by the high inverse residual resistivity ratio. Single crystals with inverse resistance ratios $[\text{IRR} = R(T = 1.8 \text{ K})/R(T = 300 \text{ K})]$, where R is resistance and T is temperature] of the order of 10^5 were selected for this study; the IRR has been shown to characterize crystal quality, with the introduction of point defects through radiation damage (20) or through off-stoichiometry (21), resulting in a decrease in low-temperature resistance and an increase in high-temperature resistance. The resistance of a SmB_6 single crystal is shown in Fig. 1B measured as a function of temperature at zero magnetic field and in an applied DC magnetic field of 45 T, demonstrating that activated electrical conductivity characteristic of an energy gap $\approx 40 \text{ K}$ at the Fermi energy persists up to high magnetic fields. The non-magnetic ground state of SmB_6 is evidenced by the linear magnetization up to 60 T (Fig. 1B, bottom inset).

We observed quantum oscillations in SmB_6 by measuring the magnetic torque. The measurements were done in magnetic fields up to 40 T and down to $T = 0.4 \text{ K}$ and in magnetic fields up to 35 T and down to $T = 0.03 \text{ K}$. Quantum oscillations periodic in inverse magnetic field are observed against a quadratic background, with frequencies ranging from 50 to 15,000 T (Fig. 1, A, C, and D). A Fourier transform of the quantum oscillations is shown in Fig. 2A as a function of inverse magnetic field, revealing well-defined peaks corresponding to multiple frequencies. The periodicity of the quantum oscillations in inverse magnetic field is revealed by the linear Landau index plot in Fig. 2B.

The observation, especially of rapid quantum oscillations with frequencies higher than 10 kT

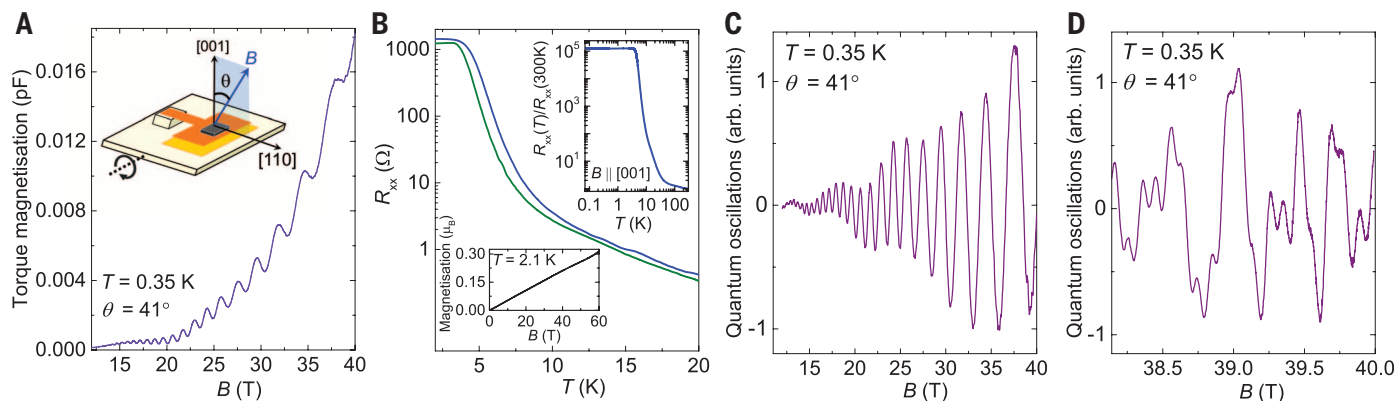


Fig. 1. Quantum oscillations in the magnetic torque in SmB_6 . (A) Quantum oscillations in magnetic torque are visible against a quadratic background. (Inset) Schematic of the magnetic torque measurement setup using a capacitive cantilever and the notation for angular rotation by angle θ . (B) Resistance as a function of temperature in zero magnetic field (blue line) and at 45 T (green line) using an unchanged measurement configuration on a SmB_6 sample of dimensions 1.1 by 0.3 by 0.1 mm. (Top inset) Measured resistance from 80 mK up to high temperatures, from which the high IRR

can be ascertained [a fit to activated electrical conductivity is provided in (23)]. (Bottom inset) Magnetization of SmB_6 at 2.1 K remains linear up to 60 T. (C) Dominant low-frequency quantum oscillations can be discerned after background subtraction of a sixth-order polynomial. (D) Magnetic torque at the highest measured fields after the subtraction of the low-frequency background torque. Quantum oscillations are visible in an intermediate-frequency range (between 2000 and 4000 T) as well as a high-frequency range up to 15,000 T.

(corresponding to approximately half the volume of the cubic Brillouin zone) in SmB_6 is striking. This observation is in contrast to previous reports of very-low-frequency quantum oscillations corresponding to a few percent of the Brillouin zone in SmB_6 , attributed to a two-dimensional surface contribution (22). Our observation of very high quantum oscillation frequencies requiring mean free paths on the order of a few micrometers would be challenging to explain from a surface layer of a few atomic lengths' thickness, which would typically yield such rapid frequencies only at a special angle of inclination at which the cyclotron orbit lies completely within the surface layer. Key to identifying the Fermi surface from which the observed quantum oscillation frequencies originate is a comparison with previous quantum oscillation measurements on metallic hexaborides, such as nonmagnetic LaB_6 , antiferromagnetic CeB_6 , and antiferromagnetic PrB_6 (16, 17). These materials exhibit a metallic ground state involving predominantly conduction electrons, with a low residual resistivity of the order of one microhm cm ($\approx 10^6$ times lower than in Kondo insulating SmB_6) and are characterized by a multiply connected Fermi surface of prolate spheroids (Fig. 3, D and E). Strikingly, the angular dependence of the various quantum oscillation frequencies in SmB_6 reveals characteristic signatures of the three-dimensional Fermi surface identified in the metallic rare earth hexaborides (Fig. 3, A to C). In particular, the high observed α frequencies (Fig. 3A) reveal the characteristic symmetry of large prolate spheroids centered at X-points of the Brillouin zone (Fig. 3, D and E), whereas the lower observed frequencies (Fig. 3A) reveal the characteristic symmetry of small ellipsoids located at the neck positions. Both of these types of ellipsoids are universal Fermi surface features identified from experiment and band structure calculations in the metallic rare earth hexaborides (16, 17). Similar features are also revealed in density functional calculations of SmB_6 when the Fermi energy is shifted from its calculated position in the insulating gap either up into the conduction or down into the valence bands (Fig. 3, D and E) (23).

The observed angular dependence of quantum oscillations in SmB_6 remains the same irrespective of whether the sample is prepared as a thin plate with a large plane face perpendicular to the [110] direction, or to the [100] direction (fig. S1), and exhibits the same characteristic signatures with respect to the orientation of the magnetic field to the crystallographic symmetry axes of the bulk crystal (Fig. 3A). The bulk quantum oscillations we measure in SmB_6 corresponding to the three-dimensional Fermi surface mapped out in the metallic rare earth hexaborides may not be directly related to the potential topological character of SmB_6 , which would have as its signature a conducting surface (24). In addition to the magnetic torque signal from the atomically thin surface region being several orders of magnitude smaller than the signal from the bulk, the observation of surface quantum oscillations would be rendered more challenging by the reported Sm depletion and resulting recon-

struction of Sm ions at the surface layer of SmB_6 (25).

The unconventional character of the state we measure in SmB_6 becomes apparent upon investigating the temperature dependence of the quantum oscillation amplitude in SmB_6 . We found that between $T = 25$ K and 2 K, the quantum oscillation amplitude exhibits a Lifshitz-Kosevich-like temperature dependence (Fig. 4) characteristic of a low effective mass similar to that of metallic LaB_6 , which has only conduction electrons (16). The comparable size of low-temperature electronic heat capacity measured for our SmB_6 single crystals to that of metallic LaB_6 (23) also seems to suggest a large Fermi surface with low effective mass in SmB_6 . However, instead of saturating at lower temperatures as would be expected for the Lifshitz-Kosevich distribution characteristic of quasiparticles with Fermi-Dirac statistics (18), the quantum oscillation amplitude increases dramatically as low temperatures down to 30 mK are approached (Fig. 4). Such non-Lifshitz-Kosevich temperature dependence is remarkable, given the robust adherence to Lifshitz-Kosevich temperature dependence in most examples of strongly correlated electron systems, from the underdoped cuprate superconductors (26) to heavy fermion systems (27, 28) to systems displaying signatures

of quantum criticality (29), a notable exception being fractional quantum Hall systems (30, 31). The possibility of a subtle departure from Lifshitz-Kosevich temperature dependence has been reported in a few materials (32, 33).

The ground state of SmB_6 is fairly insensitive to applied magnetic fields, with activated electrical conductivity behavior across a gap remaining largely unchanged up to at least 45 T (Fig. 1B). Such a weak coupling to the magnetic field is in contrast to unconventional states in other materials that are tuned by an applied magnetic field (6–11, 13). Furthermore, this rules out the possibility of quantum oscillations in SmB_6 arising from a high magnetic field state in which the energy gap is closed. The possibility that quantum oscillations arise from static, spatially disconnected metallic patches of at least 1- μm length scale that do not contribute to the electrical transport also appears unlikely. Similar quantum oscillations are observed in all (more than 10) measured high-quality samples in multiple high-magnetic-field experiments, with the best samples yielding magnetic quantum oscillations of amplitude corresponding to a substantial fraction of the expected size from bulk SmB_6 . The presence of rare earths other than Sm has been ruled out to within 0.01% by means of chemical

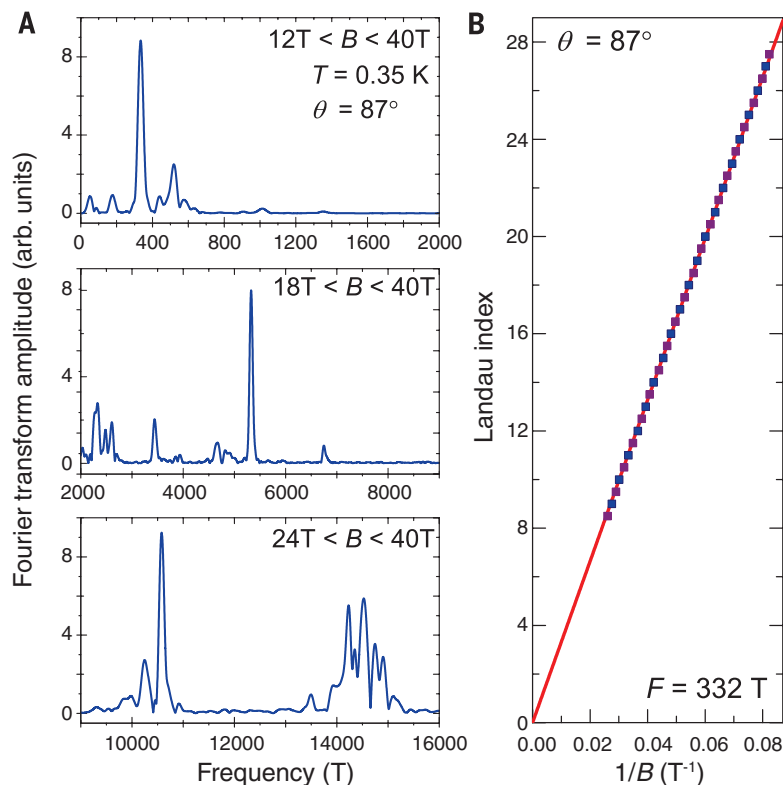


Fig. 2. Landau quantization in SmB_6 . (A) Fourier transforms of magnetic torque as a function of inverse magnetic field, from which a polynomial background has been subtracted, revealing multiple quantum oscillatory frequencies ranging from 50 T to 15,000 T. Field ranges for analysis have been chosen that best capture the observed oscillations, with the highest frequencies only appearing in the higher field ranges. (B) The maxima and minima in the derivative of magnetic torque with respect to the magnetic field, corresponding to the dominant low-frequency oscillation, are plotted as a function of inverse magnetic field; the linear dependence signals Landau quantization.

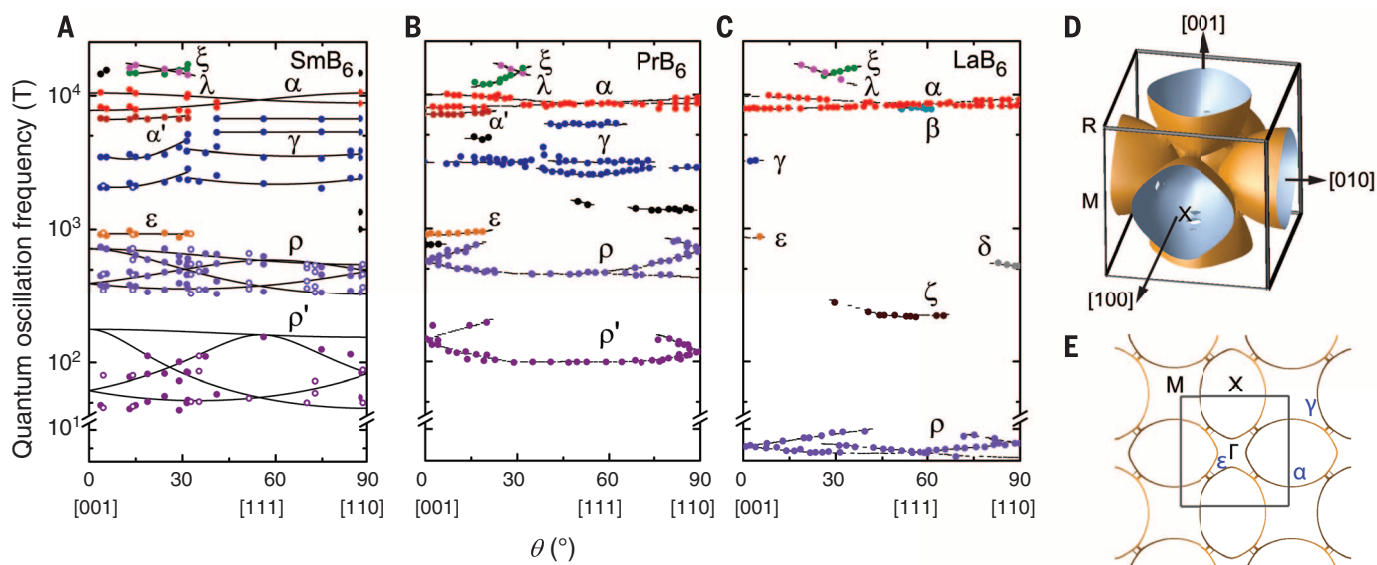
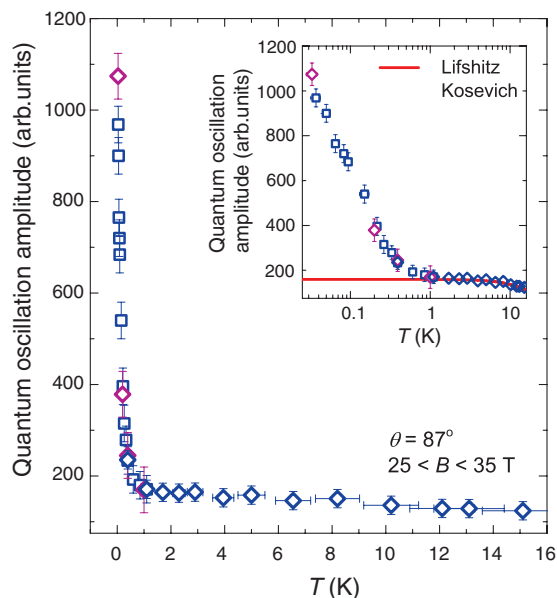


Fig. 3. Angular dependence of the quantum oscillation frequencies in SmB_6 . (A) Data from two of the SmB_6 samples in which oscillations were observed are shown, indicated by solid and open circles. One of the samples (solid circles) was prepared as a thin plate with the dominant face perpendicular to the [100] axis [sample 1 (23)], and the second sample (open circles) was prepared as a thin plate with the dominant face perpendicular to the [110] axis [sample 2 (23)]. (B and C) The angular dependence strongly resembles that of the three-dimensional Fermi surface in antiferromagnetic PrB_6 shown in (B), and nonmagnetic LaB_6 shown in (C) (17). (D and E) The α orbit in red in all the rare earth hexaborides is fit to large multiply connected

prolate spheroids centered at the X points of the Brillouin zone, shown in (D); a cross section in the XM plane is shown in (E). The ρ and ρ' orbits in each of the rare earth hexaborides are fit to small ellipsoids located at the neck positions [not shown in (D) and (E)]. More details of the fits are provided in (23). The remaining intermediate orbits are shown with lines as a guide to the eye. All orbit identifications have been made after measured frequencies and band structure calculations in PrB_6 and LaB_6 (17). (D) and (E) show Fermi surfaces calculated for SmB_6 using density functional theory (23), with a downward shift of the Fermi energy from its calculated position within the gap to expose the unhybridized bands and yield pocket sizes similar to experiment.

Fig. 4. Temperature dependence of quantum oscillation amplitude.

The dominant 330 T frequency over the magnetic field range 25 to 35 T is shown, revealing a steep increase in amplitude at low temperatures. The measurements in the temperature range from 25 K down to 0.35 K were performed in a ^3He fridge in the hybrid magnet [sample 1 (23), blue diamonds], whereas the measurements at temperatures in the range from 1 K down to 30 mK were done in a dilution fridge in the resistive magnet on two different samples [sample 1, purple diamonds; sample 3 (23), blue squares]. At low temperatures, a strong deviation from the conventional Lifshitz-Kosevich form can be seen in the inset by comparison with a simulated Lifshitz-Kosevich form for effective mass $m^* = 0.18m_0$. A logarithmic temperature scale is used in the inset for clarity.



analysis and scanning electron microscopy (23). Off-stoichiometric metallic regions of SmB_6 appear an unlikely explanation for our results, given reports that up to 30% Sm depletion does not close the energy gap (20), whereas scanning electron microscopy of our samples reveals a homogeneity of within 1% of Sm concentration

over the sample area (23). The possibility of spatially disconnected strained regions of SmB_6 , which is known to become metallic under applied pressures on the order of 10 GPa, or static spatially disconnected islands of hybridized and unhybridized Sm f -electrons also seems unlikely. An improvement in the IRR by means of the

removal of strain with electropolishing strengthens the quantum oscillation signal, whereas straining the sample by means of thermal cycling weakens the quantum oscillation signal (23). Further, the interplay between hybridized and unhybridized Sm f -electrons, which may be an important ingredient in the physics of SmB_6 , has been revealed by Mössbauer and muon-spin relaxation experiments to be homogeneous and dynamically fluctuating, rather than being manifested as static spatially inhomogeneous regions (34, 35).

The insulating state in SmB_6 in which low-energy excitations lack long-range charge transport, as shown by the activated dc electrical conductivity, but display extended character, as shown by quantum oscillations, poses a mystery. A clue might be provided by slow fluctuations between a collectively hybridized insulating state and an unhybridized state in which the conduction electrons form a solely conduction electron Fermi surface, similar to that we observe (2, 35–39). A fluctuation time scale in the range between 10^{-8} and 10^{-11} seconds is suggested by previous x-ray absorption spectroscopy and Mössbauer measurements (40). This time scale is longer or comparable with the inverse cyclotron frequency ($1/\omega_c$), which is on the order of 10^{-11} seconds for the measured cyclotron orbits. Intriguingly, similar slow fluctuations have been invoked to explain quantum critical signatures in the metallic f -electron system $\beta\text{-YbAlB}_4$ (41). SmB_6 may be viewed as being on the border of quantum criticality in the

sense that it transforms from a nonmagnetic insulating phase to a magnetic metallic phase under applied pressures on the order of 10 GPa (42–45), which is in contrast to other metallic rare earth hexaborides in which the *f*-electrons order magnetically in the ambient ground state. Our observation of a large three-dimensional conduction electron Fermi surface revealed by quantum oscillations may be related to reports of a residual density of states at the Fermi energy in SmB₆ through measurements of heat capacity (23, 46), optical conductivity (47), Raman scattering (48), and neutron scattering (49). Another possibility is that quantum oscillations could arise even in a system with a gap in the excitation spectrum at the Fermi energy, provided that the size of the gap is not much larger than the cyclotron energy (50). Within this scenario, the residual density of states observed at the Fermi energy with complementary measurements and the steep upturn in quantum oscillation amplitude we observe at low temperatures appear challenging to explain.

REFERENCES AND NOTES

- A. C. Hewson, *The Kondo Problem to Heavy Fermions* (Cambridge Univ. Press, Cambridge, 1997).
- N. F. Mott, *Philos. Mag.* **30**, 403–416 (1974).
- G. Aeppli, Z. Fisk, *Comments Condens. Matt. Phys.* **16**, 155 (1992).
- A. J. Schofield, *Contemp. Phys.* **40**, 95–115 (1999).
- J. E. Moore, *Nature* **464**, 194–198 (2010).
- S. Paschen *et al.*, *Nature* **432**, 881–885 (2004).
- S. A. Grigera *et al.*, *Science* **306**, 1154–1157 (2004).
- F. Lévy, I. Sheikin, B. Grenier, A. D. Huxley, *Science* **309**, 1343–1346 (2005).
- C. Pfleiderer, S. R. Julian, G. G. Lonzarich, *Nature* **414**, 427–430 (2001).
- N. D. Mathur *et al.*, *Nature* **394**, 39–43 (1998).
- H. Hegger *et al.*, *Phys. Rev. Lett.* **84**, 4986–4989 (2000).
- H. Löhneysen *et al.*, *Phys. Rev. Lett.* **72**, 3262–3265 (1994).
- J. Flouquet, *Prog. Low Temp. Phys.* **15**, 139–281 (2005).
- C. M. Varma, Z. Nussinov, W. van Saarloos, *Phys. Rep.* **361**, 267–417 (2002).
- A. Menth, E. Buehler, T. H. Geballe, *Phys. Rev. Lett.* **22**, 295–297 (1969).
- S. Behler, K. Winzer, *Z. Phys. B Condens. Matter* **82**, 355–361 (1991).
- Y. Ōnuki, T. Komatsubara, P. H. P. Reinders, M. Springford, *J. Phys. Soc. Jpn.* **58**, 3698 (1989).
- D. Shoenberg, *Magnetic Oscillations in Metals* (Cambridge Univ. Press, Cambridge, 1984).
- M. Ciomaga Hatnean, M. R. Lees, D. M. K. Paul, G. Balakrishnan, *Sci. Rep.* **3**, 3071 (2013).
- J. Morillo, C.-H. de Novion, J. Jun, *Solid State Commun.* **48**, 315–319 (1983).
- G. Priztáš, S. Gabáni, K. Flachbart, V. Filipov, N. Shitsevalova, *JPS Conf. Proc.* **3**, 012021 (2014).
- G. Li *et al.*, *Science* **346**, 1208–1212 (2014).
- Materials and methods are available as supplementary materials on Science Online.
- M. Dzero, K. Sun, V. Galitski, P. Coleman, *Phys. Rev. Lett.* **104**, 106408 (2010).
- M. Aono *et al.*, *Surf. Sci.* **86**, 631–637 (1979).
- S. E. Sebastian, N. Harrison, G. G. Lonzarich, *Rep. Prog. Phys.* **75**, 102501 (2012).
- L. Taillefer, G. G. Lonzarich, *Phys. Rev. Lett.* **60**, 1570–1573 (1988).
- P. H. P. Reinders, M. Springford, P. T. Coleridge, R. Boulet, D. Ravot, *Phys. Rev. Lett.* **57**, 1631–1634 (1986).
- H. Shishido, R. Settai, H. Harima, Y. Onuki, *J. Phys. Soc. Jpn.* **74**, 1103–1106 (2005).
- D. C. Tsui, H. L. Stormer, A. C. Gossard, *Phys. Rev. Lett.* **48**, 1559–1562 (1982).
- R. B. Laughlin, *Phys. Rev. Lett.* **50**, 1395–1398 (1983).
- M. Elliott, T. Ellis, M. Springford, *J. Phys. F Met. Phys.* **10**, 2681–2706 (1980).
- A. McCollam, J.-S. Xia, J. Flouquet, D. Aoki, S. R. Julian, *Physica B* **403**, 717–720 (2008).
- P. K. Biswas *et al.*, *Phys. Rev. B* **89**, 161107 (2014).
- R. L. Cohen, M. Eibschütz, K. W. West, *Phys. Rev. Lett.* **24**, 383–386 (1970).
- P. Coleman, E. Miranda, A. Tselik, *Physica B* **186**–**188**, 362–364 (1993).
- H. Kleinert, *Gauge Fields in Condensed Matter* (World Scientific, Singapore, 1988).
- T. Kasuya, *J. Phys. Soc. Jpn.* **63**, 397–400 (1994).
- Q. Si, S. Paschen, *Phys. Status Solidi B* **250**, 425–438 (2013).
- M. Mizumaki, S. Tsutsui, F. Iga, *J. Phys. Conf. Ser.* **176**, 012034 (2009).
- Y. Matsumoto *et al.*, *Science* **331**, 316–319 (2011).
- J. C. Cooley, M. C. Aronson, Z. Fisk, P. C. Canfield, *Phys. Rev. Lett.* **74**, 1629–1632 (1995).
- A. Barla *et al.*, *Phys. Rev. Lett.* **94**, 166401 (2005).
- J. Derr *et al.*, *J. Phys. Condens. Matter* **18**, 2089–2106 (2006).
- J. Derr *et al.*, *Phys. Rev. B* **77**, 193107 (2008).
- K. Flachbart *et al.*, *Physica B* **378**, 610–611 (2006).
- T. Nanba *et al.*, *Physica B* **186**, 440–443 (1993).
- P. Nyhus, S. L. Cooper, Z. Fisk, J. Sarrao, *Phys. Rev. B* **52**, R14308–R14311 (1995).
- P. A. Alekseev, J.-M. Mignot, J. Rossat-Mignod, V. N. Lazukov, I. P. Sadikov, *Physica B* **186**, 384–386 (1993).
- K. Miyake, *Physica B* **186**, 115–117 (1993).

ACKNOWLEDGMENTS

B.S.T., Y.-T.H., M.H., M.K., A.S., and S.E.S. acknowledge support from the Royal Society, the Winton Programme for the Physics of Sustainability, and the European Research Council (ERC) under the European Union's Seventh Framework Programme (grant FP/2007-2013)/ERC Grant Agreement 337425. B.Z. and L.B. acknowledge support from the U.S. Department of Energy (DOE)–Basic Energy Sciences (BES) through award DE-SC0002613. M.C.H. and G.B. acknowledge support from Engineering and Physical

Sciences Research Council (EPSRC) grant EP/L014963/1. N.H. and Z.Z. acknowledge support from the DOE Office of Science, BES–Materials Science and Engineering “Science of 100 Tesla” program. M.D.J. acknowledges support for this project by the Office of Naval Research (ONR) through the Naval Research Laboratory’s Basic Research Program. G.G.L. acknowledges support from EPSRC grant EP/K012894/1. A portion of this work was performed at the National High Magnetic Field Laboratory, which is supported by NSF Cooperative Agreement DMR-1157490 and the state of Florida. We acknowledge valuable inputs from G. Baskaran, D. Benkert, A. K. Cheetham, D. Chowdhury, P. Coleman, N. R. Cooper, M. P. M. Dean, O. Ertem, J. Flouquet, R. H. Friend, R. Golombok, C. Harris, S. A. Hartnoll, T. Kasuya, G. Khalullin, E.-A. Kim, J. Knolle, P. A. Lee, P. B. Littlewood, C. Liu, K. Miyake, J. E. Moore, O. Petrenko, S. Sachdev, A. Shekhter, N. Shitsevalova, Q. Si, A. Thomson, S. Todadri, C. M. Varma, and J. Zaanen. We thank magnet laboratory personnel, including J. Billings, R. Carrier, E. S. Choi, B. L. Dalton, D. Freeman, L. J. Gordon, M. Hicks, C. H. Mielke, J. M. Petty, and J. N. Piotrowski, for their assistance. Data will be made available at the institutional data repository www.data.cam.ac.uk/data-repository.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/287/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S5
Table S1
References (51–57)

28 January 2015; accepted 24 June 2015
Published online 2 July 2015
10.1126/science.aaa7974

NANOPARTICLE IMAGING

3D structure of individual nanocrystals in solution by electron microscopy

Jungwon Park,^{1,2,3*} Hans Elmlund,^{4,5*} Peter Ercius,^{6*} Jong Min Yuk,^{7,8,9} David T. Limmer,¹⁰ Qian Chen,^{1,8,11} Kwanpyo Kim,¹² Sang Hoon Han,¹³ David A. Weitz,^{2,3} A. Zettl,^{7,8,9} A. Paul Alivisatos^{1,8,9,†}

Knowledge about the synthesis, growth mechanisms, and physical properties of colloidal nanoparticles has been limited by technical impediments. We introduce a method for determining three-dimensional (3D) structures of individual nanoparticles in solution. We combine a graphene liquid cell, high-resolution transmission electron microscopy, a direct electron detector, and an algorithm for single-particle 3D reconstruction originally developed for analysis of biological molecules. This method yielded two 3D structures of individual platinum nanocrystals at near-atomic resolution. Because our method derives the 3D structure from images of individual nanoparticles rotating freely in solution, it enables the analysis of heterogeneous populations of potentially unordered nanoparticles that are synthesized in solution, thereby providing a means to understand the structure and stability of defects at the nanoscale.

Colloidal nanoparticles are clusters of hundreds to thousands of inorganic atoms typically surrounded by organic ligands that stabilize them in solution. The atomic arrangement of colloidal nanoparticles determines their chemical and physical properties, which are distinct from bulk materials and can be exploited for many applications in biological imaging, renewable energy, catalysis, and more. The 3D atomic arrangement on the surface and in the core of a nanocrystal influences the electronic struc-

ture, which affects how the nanocrystal functions in catalysis or how it interacts with other components at the atomic scale (1). Introduction of atomic dopants, surface adatoms, defects, and grain boundaries alters the chemical properties of nanocrystals (2). Ensembles of synthesized nanocrystals in solution are structurally inhomogeneous because of the stochastic nature of nanocrystal nucleation and growth (3, 4). Therefore, a method for determination of the 3D atomic arrangement of individual unique nanoparticles in solution is needed.

Electron tomography is routinely used for 3D analysis of materials (5–9). This method cannot be applied to individual particles in a liquid because it relies on acquisition of images of a single object at many different tilt angles over a period of 2 to 5 hours, assuming the object is static during the entire acquisition. Single-particle cryo-electron microscopy (cryo-EM) is a common method for the determination of 3D structures in biological sciences. The average 3D Coulomb potential map (i.e., density) of a protein is reconstructed from tens of thousands of transmission electron microscopy (TEM) images of randomly oriented copies of the same protein embedded in vitreous ice (10). The unknown 3D projection angles of the images are determined by computational methods (11). Single-particle cryo-EM has succeeded in reconstructing biological molecules with nearly 3 Å resolution (10, 12). A similar approach was recently applied to reconstruct the atomic structure of homogeneous ultrasmall gold clusters (13). However, the single-particle method is not readily applicable to 3D reconstruction of colloidal nanoparticles because of their intrinsic structural inhomogeneity at the atomic level.

TEM has undergone technical improvements in the past decades (5, 14–17). The image resolution has been improved with the introduction of electron lens aberration correctors (15). The development of direct electron detectors has led to improvements in image quality and rapid acquisition of movies that allow for compensation for beam-induced specimen motion, thereby providing a substantial enhancement to single-particle cryo-EM (18).

We present a hybrid method for reconstructing the 3D structures of individual nanoparticles in solution. The method represents a combination of three technological advancements from TEM imaging in biological and materials science: (i) the development of graphene as a covering to hold a liquid in vacuum (the so-called graphene liquid cell, or GLC) that allows atomic-resolution imaging of nanoparticles that move and rotate freely in solution by aberration-corrected TEM (3, 19); (ii) the advent of direct electron detectors, producing movies with millisecond

frame-to-frame time resolution of the rotating nanocrystals (18); and (iii) a theory for *ab initio* single-particle 3D reconstruction, used to solve the inverse problem of recovering the unknown 3D orientations of the individual noisy nanocrystal projections (11). The resulting hybrid technique, 3D structure identification of nanoparticles by GLC EM (abbreviated as SINGLE), was used to separately reconstruct the 3D structures of two individual Pt nanocrystals in solution.

Pt nanoparticles were chosen because of their high electron scattering strength, because their detailed atomic structure is important for catalysis, and because earlier graphene liquid cell studies have shown that they grow by nanoparticle aggregation, resulting in complex structures that are not possible to determine by any previously developed method. Pt nanocrystals with sub-2 nm diameter were prepared in solution. Two graphene sheets were grown by the chemical vapor deposition method and used to entrap solvated nanocrystals (3). The graphene provides an ultrathin covering of material to maintain liquid conditions in the TEM vacuum and presents an inert surface onto which the nanoparticles do not adsorb. The translational and rotational motions of the particles in liquid pockets with sub-50 nm diameter were imaged *in situ* using TEAM I, a TEM instrument

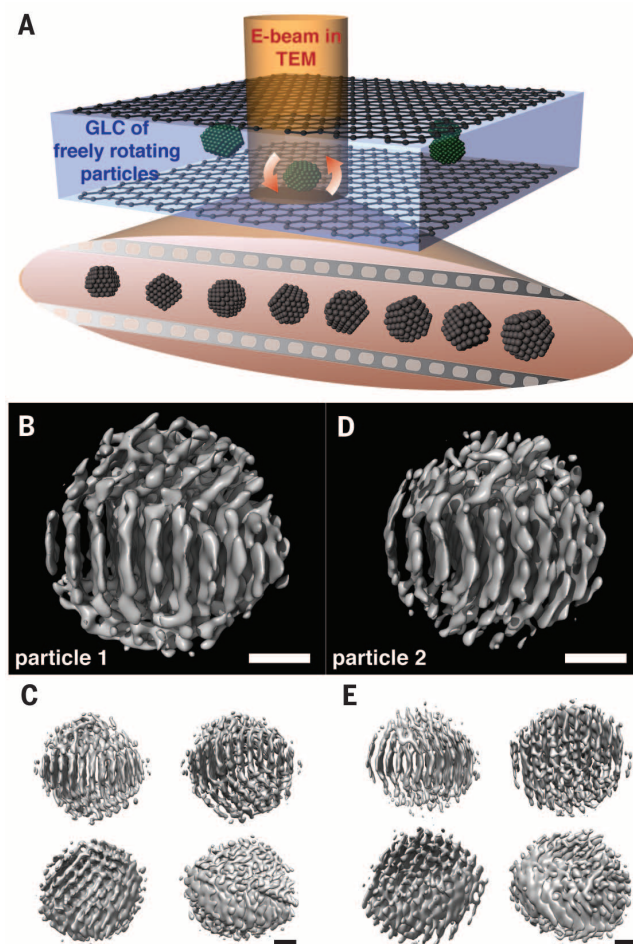
with geometrical and chromatic aberration correction, operated at 300 kV using a direct electron detector (Fig. 1A). The corrector was set to apply a slightly negative spherical aberration coefficient of about $-10\ \mu\text{m}$. Under these imaging conditions, we produced white atom contrast when using a small underfocus value of 30 to 50 Å.

Movies of the moving particles had a temporal resolution of 50 frames/s, a field of view of 1024×1024 pixels, and a Nyquist sampling limit of 0.56 Å. Each movie frame represents a 2D projection of many particles in random orientations. To reconstruct the individual 3D structures, we chose a small region around each single particle of interest in each individual frame to create a set of projections per particle. The 3D orientations of the resulting series of noisy 2D images of a single rotating particle were recovered computationally using an *ab initio* 3D reconstruction algorithm adapted from one originally developed to recover orientations from cryo-EM images of many identical individual particles (11).

Although the TEM movies contain many particles, not all particles could be used for reconstruction because of overlap with other particles and insufficient rotation. Here, we present the two most reliable 3D reconstructions from a 1561-image series (particle 1 in Fig. 1, B and C, and movie S1) and a 1171-image series (particle 2

Fig. 1. Schematic illustration of *in situ* TEM imaging of Pt nanocrystals freely rotating in a graphene liquid cell (GLC) and 3D EM density maps calculated from individual Pt nanoparticles in solution.

(A) A movie of the single rotating Pt nanocrystal provides 2D projected TEM still snapshots in many orientations for *ab initio* particle reconstruction. (B) EM density map obtained from the 3D reconstruction of particle 1. The orientation of the particle is aligned to expose {111} planes of the core domain. Three distinct crystal domains can be identified. (C) 3D EM density map of particle 1 with alternative viewing angles. (D) EM density map obtained from the 3D reconstruction of particle 2. (E) 3D EM density map of particle 2 with alternative viewing angles. Each panel in (C) and (E) presents the two particles with the same angle and direction with respect to the orientations in (B) and (D). Scale bars, 0.5 nm.



¹Department of Chemistry, University of California, Berkeley, CA 94720, USA. ²Department of Applied Physics, Harvard University, Cambridge, MA 02138, USA. ³School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. ⁴Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Monash University, Clayton, VIC 3800, Australia. ⁵ARC Centre of Excellence for Advanced Molecular Imaging, Clayton, VIC 3800, Australia. ⁶Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ⁷Department of Physics, University of California, Berkeley, CA 94720, USA. ⁸Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ⁹Kavli Energy NanoScience Institute, Berkeley, CA 94720, USA. ¹⁰Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08540, USA. ¹¹Miller Institute for Basic Research in Science, University of California, Berkeley, CA 93720, USA. ¹²Department of Physics, Ulsan National Institute of Science and Technology, Ulsan 689-798, South Korea. ¹³Amore-Pacific Co. R&D Center, Yongin 446-829, South Korea.

*These authors contributed equally to this work. †Corresponding author. E-mail: alivis@berkeley.edu

in Fig. 1, D and E, and movie S2). Shown in Fig. 1, B to E, are the EM density maps of the two Pt nanocrystals. The rendered particle volumes are 5300 \AA^3 (diameter 22 \AA) for particle 1 and 4800 \AA^3 (diameter 20 \AA) for particle 2. Along with the direct visualization of the spatial distribution of

Pt atomic planes, external and internal structures of the particle are uncovered. Each reconstruction shows three distinct crystal domains in both of the Pt particles. In Fig. 1, B and D, we show views of the EM maps in an orientation that reveals distinct lattice planes of the core domain.

Differently oriented 3D density maps are shown in Fig. 1, C and E, and movies S1 and S2.

Our 3D reconstruction methodology produced reconstructions at near-atomic resolution from relatively small sets of noisy experimental TEM images of nanocrystals in random orientations.

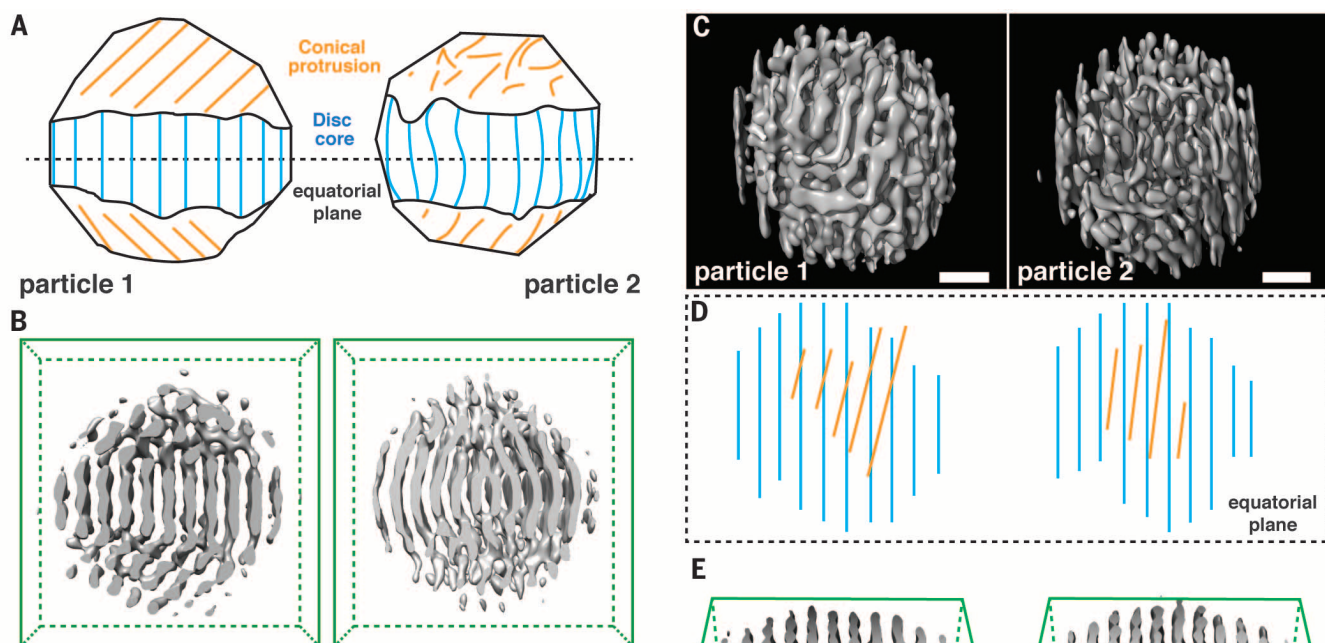
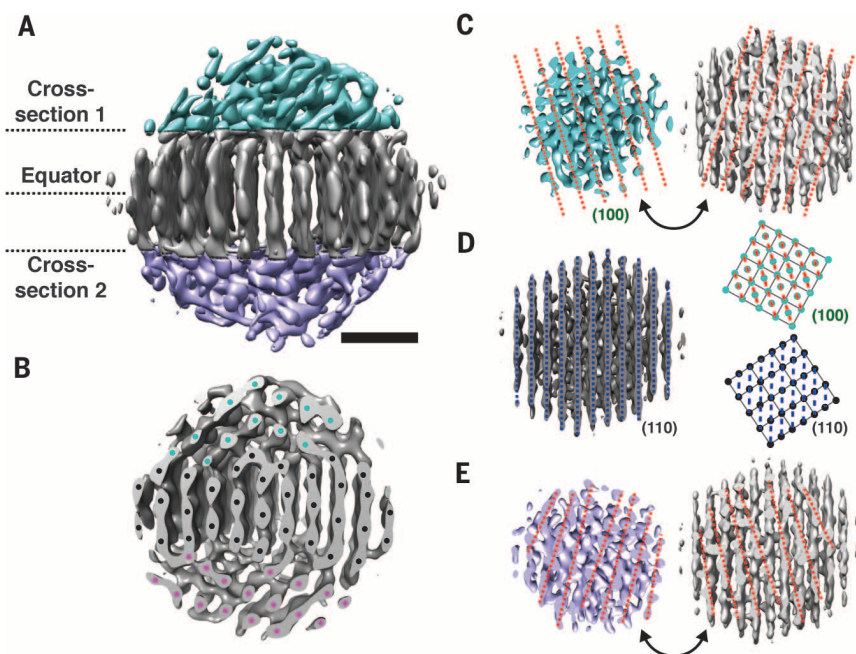


Fig. 2. The underlying structural principle of the small Pt nanoparticles studied here. (A) Schematic illustration of the front view of particles 1 and 2 shown in Fig. 1. Both particles are composed of a dense central disc of atomic planes (blue lines) with conical protrusions (orange lines) anchoring on each side of the disc. (B) Cross-sectional view of the EM density map of particle 1 and 2 along the vertical plane. (C) View orthogonal to (A), showing the overlaid lattices. Scale bars, 0.5 nm . (D) Schematic illustration of the top view of particles 1 and 2 shown in (C). (E) Cross-sectional view of the EM density map of particles 1 and 2 along the equatorial plane.

Fig. 3. Cross-sectional study of particle 1. (A) 3D density map of particle 1 with color coding to highlight the three sections. Cross sections 1 and 2 are in arbitrary positions near crystal domain interfaces. Scale bar, 0.5 nm . (B) Slab through the 3D reconstruction of particle 1 along the vertical plane, with tentative atomic positions indicated. ABC repeats of $\{111\}$ planes are visible. (C) Slab along cross section 1. The exposed (100) surface of the fcc Pt crystal is shown consistently in both exposed surfaces. Intersections with $\{111\}$ planes in the top domain are displayed by red dashed lines. (D) Slab along the equatorial plane of (A) exposes a (110) plane. The intersections with $\{111\}$ planes are shown by blue dashed lines. Pseudo-atomic structure (right) demonstrates the rotation angle (14°) between the (100) and (110) surfaces from the top and bottom domains, respectively. (E) Slab along cross section 2. The disordered (100) surface of the fcc Pt crystal is shown consistently in both exposed surfaces. Intersections with pseudo $\{111\}$ planes in the bottom domain are displayed by red dashed lines.



We tested the validity of SINGLE using 1000 multislice TEM simulations of a randomly oriented Pt nanocrystal with dodecahedral symmetry and corresponding size. Reconstructions were obtained from images with a signal-to-noise ratio (SNR) roughly corresponding to that of the experimental images (see fig. S1). This verified that the experimental images can conform to the projection-slice theorem with ideal microscope conditions (20).

To confirm the existence of a projected lattice in the experimental images, we calculated their power spectra, which showed distinct spots along different crystalline zone axes (fig. S2). Closer examination of the power spectra and corresponding orientation coverage throughout the image series revealed that each particle continuously undergoes small local rotations followed by rapid orientation changes, often accompanied by lateral movement. The noise variance in the individual frames of the movie exceeds the signal variance by approximately a factor of 5, despite the strong scattering from the heavy Pt atoms (movies S3 and S4 show 200 raw TEM images). Although sufficiently thin to observe atomic detail, the liquid between the graphene sheets introduces a

granular background, making it difficult to distinguish the facets of the nanocrystal in the individual frames. To enhance the image contrast and allow accurate 3D orientation determination, we averaged the image series in sets of five consecutive frames, resulting in a time resolution of ~100 ms. After frame averaging, we carefully scanned the image series to remove a small fraction (~20%) of those averages that did not show any lattice contrast because the particles had moved out of the narrow ideal focal plane of the aberration-corrected microscope.

Initial 3D models were obtained using our recently developed framework (PRIME) for *ab initio* single-particle 3D reconstruction (17). The standard technique used in biological single-particle cryo-EM (10) assigns each image the single best matching orientation, as determined by correlation matching to a gallery of reference images, obtained by projecting an *a priori* available 3D reference model. Two fundamental limitations of the standard technique are the bias introduced by the initial 3D model and the lack of mechanisms for modeling the alignment errors when data are noisy and the model is of poor quality. These limitations may be quite substantial when

reconstructing nanocrystals, because every particle is different and the particle population cannot be averaged, as in traditional single-particle EM. The individual frames also have a low SNR because of the relatively low per-frame electron dose and because of the granular background introduced by the liquid. PRIME overcomes these drawbacks by using weighted orientation assignment and stochastic optimization for determination of an optimal orientation weight distribution without any *a priori* information about the nanoparticles.

To initialize the 3D reconstruction process, we assigned the random orientations to images, producing a featureless spherical density map. The random orientations were refined by stochastic optimization of the correlation between the images and reprojections of the density map, using information from 30 Å to 3 Å and a discrete search space of orientations. The resulting initial model had a resolution of 2.5 Å according to the 0.143 Fourier shell correlation (FSC) criterion (21). We extended the PRIME algorithm by introducing a continuous orientation search space and used stochastic optimization to determine a continuous distribution of weights that related the continuous distribution of orientation parameters to the 3D reconstruction. Each round of the PRIME iterative alignment procedure involved determination of orientation weights for all particle images, followed by a weighted 3D reconstruction by direct Fourier reconstruction using a Kaiser-Bessel interpolation kernel. A few hundred iterations were executed, and in every round the FSC was calculated and used to construct a 3D Wiener filter that filtered the map such that the optimal SNR was obtained at the present resolution (22). The resolution of the final refined maps was measured to 2.10 Å (particle 1) and 2.14 Å (particle 2), respectively.

Even though the FSC methodology provides an accurate measure of the resolution when the reference structure used for matching has been appropriately low-pass filtered, it does not ensure against grossly incorrect structures. To validate our structures, we determined the agreement between the individual images used for reconstruction and the corresponding reprojections of the reconstructed 3D map. All image reprojection pairs (fig. S3) were generated for the two reconstructions, and the Fourier ring correlation (FRC) (23) was calculated between all pairs. The average FRC was larger than 0.143 to a resolution of 1 Å and showed a distinct peak spanning the 1 to 2 Å resolution region (fig. S3). This peak is due to the correlation between atomic densities in the reprojections and atomic densities in the images. We concluded that our reconstructions showed excellent agreement with the images. The spatial resolution is higher for particle 1, which has a larger number of frames, indicating that the present resolution can be improved by acquisition of longer movies that cover a wider range of rotational orientations.

We had anticipated that the Pt nanocrystals would have at least twofold rotational symmetry, perhaps even cubic symmetry. Remarkably, the

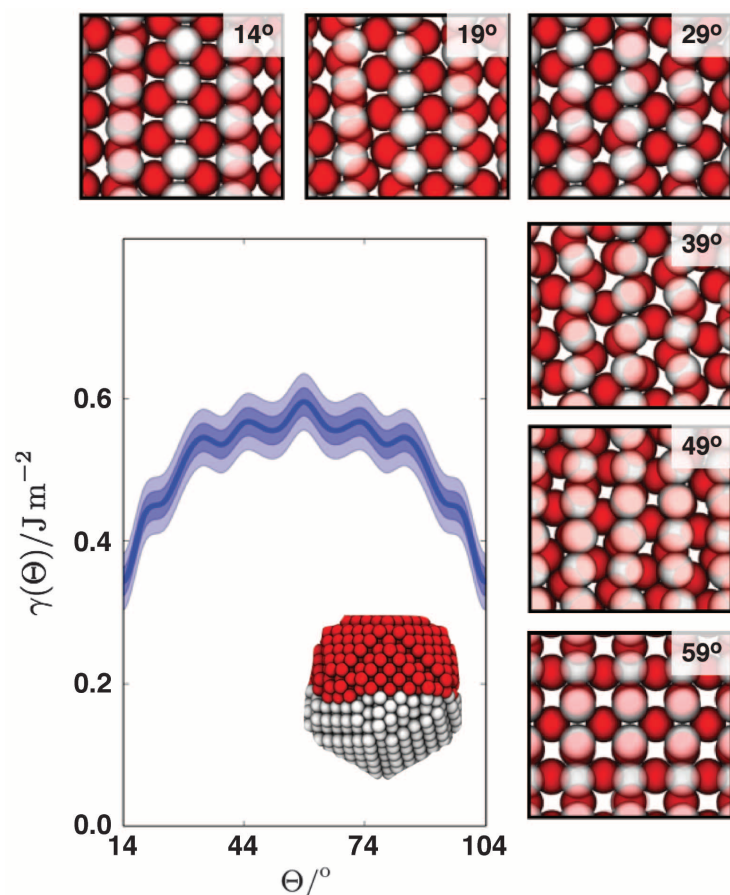


Fig. 4. Twist grain boundary free energies as a function of misalignment angle for the (110)-(100) crystal planes of a nanocrystal with 1135 atoms. Shaded regions indicate error bars of 1 and 2 standard deviations. The inset shows a relaxed nanocrystal with an initial misalignment angle of 14°. The non-flat grain boundary is located in the center of the nanocrystal. Images around the perimeter show the two relaxed planes at the grain boundary from the 100 (red) and 110 (white) grains for different misalignment angles.

reconstructions appeared asymmetrical, and attempts to apply C2, C4, C5, and D2 point-group symmetry by aligning the images to the principal symmetry axis and doing symmetric orientation refinement failed to improve the correlation between the reprojections and the images. This suggests that the Pt nanocrystals do not possess icosahedral, cubic, or pentagonal symmetry but are intrinsically asymmetrical. Reconstruction by the traditional single-particle cryo-EM technique, using an initial model based on a faceted face-centered cubic (fcc) nanoparticle, would suffer severe initial model bias and would not reproduce the true asymmetrical multidomain structure.

Our reconstructions of particles 1 and 2 revealed different asymmetrical crystal structures with the same underlying principle: a dense central disc of atomic planes (the core) with conical extensions anchored on each side, protruding in opposing and orthogonal directions with respect to the equatorial plane (see schematics in Fig. 2A with the lattice in different crystal domains colored blue and orange). Despite this similarity, the reconstructed 3D structures of the two Pt particles show interesting differences. First, the atomic arrangements on the surfaces observed in the EM density maps of particle 1 and 2 are distinct (Fig. 1C and Fig. 2E). Second, the two particles show different degrees of crystallinity in each domain. Whereas straight {111} crystal planes in three domains of particle 1 are shown throughout the cross-sectional images along the mid-vertical plane (Fig. 2B) and vertical planes at different depths (fig. S4), particle 2 shows more disordered internal structures within the domains. Also in particle 1, the conical protrusions have {111} atomic planes tilted with respect to the {111} atomic planes in the core. Particle 2 exhibits protrusions with a larger degree of disorder relative to the well-aligned {111} lattices in the particle 1 protrusions. Views of the reconstructions perpendicular to the equatorial plane (Fig. 2C) and a schematic illustration (Fig. 2D) show the overlaid lattices of the core (blue lines) and protrusions (yellow lines) with different tilting angles for particle 1 (14°) and particle 2 (7°).

The cross-sectional views along the equatorial plane (Fig. 2E) for particles 1 and 2 show similar arrangements of the {111} lattice planes, but cross-sectional images along the horizontal planes below and above the equatorial plane for the two particles (particle 1 in Fig. 3, C to E; particle 2 in fig. S5) indicate that the two particles are assembled by multiple domains but in unique geometries. The multiple domains and twisted grain boundaries that are present in the reconstructions are similar to previous observations of Pt nanocrystal growth trajectories, where small particles were observed to join along surfaces with low ligand coverage (3). The multidomain arrangement is also supported by other tomography reconstructions of larger Pt nanocrystals containing multiply twinned domains with decahedral symmetry and central screw dislocations (3, 5). Multidomain structures are common in many other colloidal metal nanoparticles, which likely evolve as a result of multiple coalescence events

during growth (3, 24–26). Our observation of heterogeneously structured Pt nanoparticles formed in the same solution confirms that individual particles from the same synthesis follow different nucleation and growth trajectories.

Figure 3A presents a 3D density map with colored sections indicating the upper (blue), core (gray), and lower (purple) domains of particle 1. Figure 3B shows a cross section perpendicular to the equatorial plane with the tentative atomic positions (different colored dots in each domain) indicating {111} lattice planes of fcc crystal that have repetition in every three crystal planes in all three crystal domains. Atoms closer to the surface seem to deviate farther from a perfect fcc structure. This is presumably explained by the fact that surface atoms are prone to relax excess free energy because of insufficient coordination and stabilization by surface ligand binding. In addition, the interface between the domains is disordered and not flat—a general consequence of the reduction in surface energies expected for nanoscale crystals relative to their bulk counterparts. The locations of three cross sections along horizontal planes (cross section 1, equator, and cross section 2) are indicated in Fig. 3A. Cross sections 1 and 2 are positioned near the interface between the central disk and the upper and lower conical protrusions, respectively. Cross section 1 exposes two facing surfaces from the upper and core domains (blue and gray densities in Fig. 3C, respectively). Cross section 1 shows a {100} surface with red dashed lines that trace the {111} planes exposed on the {100} surface. The cross section at the equator (Fig. 3D) exhibits a {110} surface, and blue dashed lines indicate the {111} planes exposed to the {110} surface. The red and blue lines are mapped onto the pseudo-atomic illustration of {100} and {110} surfaces with the proper orientation (a 14° rotation angle) in the right image of Fig. 3D. At cross section 2, near the interface between the central disk and the lower conical protrusion, a surface structure with pseudo {100} patterns traced by red dotted lines is exposed from the core and lower domains (gray and purple densities in Fig. 3E, respectively), which deviate from the ideal fcc structure. Figure 3 shows that the conical protrusions and the core join along {100} and {110} surfaces with distortions at the interface. Multiple domains merging along the low-index crystal planes, such as {100} and {110}, are presumably formed by coalescence events between small particles during the particle growth as a route to minimize excess surface energy. We previously observed a similar scenario: Two small Pt particles join along the {111} surfaces during growth (3). Presumably, coalescence along the low-index surfaces and the consequent evolution of the interface structure are mechanisms by which the nanocrystals reduce free energy.

To examine whether there is a thermodynamic rationale for the multidomain structures that we observed, we computed the free energies for the grain boundary formation and for the ligand-exposed surfaces. This was done using the Frenkel-Ladd method (27) for an embedded atom model of Pt (28), using an appropriate thermodynamics

integration path (fig. S6). We found that for low-angle misalignments, like those observed in particle 1 [14° rotation angle between {111} planes exposed on {110} and {100} surfaces from the central disk and upper domain in Fig. 3], the grain boundary free energy for a nanoparticle with 1200 atoms is 0.3 J/m², increasing to 0.55 J/m² for misalignment angles of up to 59° (Fig. 4). Small rotational relaxation of the crystal grains is found to be energetically viable at these interfaces, which may explain the orientation of the side protrusions in particle 1 with respect to its central section.

The surface free energy for the ligand-exposed interfaces was computed to be 2.8 J/m² as averaged over the {100} and {111} surfaces, weighted appropriately for the cuboctahedral shape. The disparity of scales between the grain boundary free energies and the much larger exposed surface free energies confirms that there is a large thermodynamic driving force for coalescence, even when such events result in grain boundary formation. As the free energy gain upon coalescence is much larger than thermal energies, initial aggregations are likely irreversible. The resultant grain boundaries are then kinetically arrested over laboratory time scales (see supplementary materials).

Our results show that the SINGLE methodology can be used to investigate the structural principles underlying the assembly and transient morphology of any stable, small nanoparticle in solution. We envision that SINGLE can be applied directly to in situ 3D structural studies of many other kinds of solvated particles.

REFERENCES AND NOTES

1. D. Mocatta *et al.*, *Science* **332**, 77–81 (2011).
2. H. Zhang, T. Watanabe, M. Okumura, M. Haruta, N. Toshima, *Nat. Mater.* **11**, 49–52 (2012).
3. J. M. Yuk *et al.*, *Science* **336**, 61–64 (2012).
4. H. Zheng *et al.*, *Science* **324**, 1309–1312 (2009).
5. C.-C. Chen *et al.*, *Nature* **496**, 74–77 (2013).
6. P. A. Midgley, R. E. Dunin-Borkowski, *Nat. Mater.* **8**, 271–280 (2009).
7. P. A. Midgley, M. Weyland, *Ultramicroscopy* **96**, 413–431 (2003).
8. C. Zhu *et al.*, *Phys. Rev. B* **88**, 100201 (2013).
9. S. Van Aert, K. J. Batenburg, M. D. Rossell, R. Erni, G. Van Tendeloo, *Nature* **470**, 374–377 (2011).
10. Y. Cheng, T. Walz, *Annu. Rev. Biochem.* **78**, 723–742 (2009).
11. H. Elmlund, D. Elmlund, S. Bengio, *Structure* **21**, 1299–1306 (2013).
12. K. Murakami *et al.*, *Science* **342**, 1238724 (2013).
13. M. Azubel *et al.*, *Science* **345**, 909–912 (2014).
14. B. Goris *et al.*, *Nat. Mater.* **11**, 930–935 (2012).
15. M. C. Scott *et al.*, *Nature* **483**, 444–447 (2012).
16. A. Leis, B. Rockel, L. Andrees, W. Baumeister, *Trends Biochem. Sci.* **34**, 60–70 (2009).
17. C. V. Robinson, A. Sali, W. Baumeister, *Nature* **450**, 973–982 (2007).
18. M. Battaglia *et al.*, *Nucl. Instrum. Methods Phys. Res. A* **598**, 642–649 (2009).
19. Q. Chen *et al.*, *Nano Lett.* **13**, 4556–4561 (2013).
20. R. N. Bracewell, *Aust. J. Phys.* **9**, 198–217 (1956).
21. M. van Heel, M. Schatz, *J. Struct. Biol.* **151**, 250–262 (2005).
22. P. B. Rosenthal, R. Henderson, *J. Mol. Biol.* **333**, 721–745 (2003).
23. W. O. Saxton, W. Baumeister, *J. Microsc.* **127**, 127–138 (1982).
24. M. Takesue *et al.*, *J. Am. Chem. Soc.* **133**, 14164–14167 (2011).
25. J. Polte *et al.*, *ACS Nano* **6**, 5791–5802 (2012).
26. J. Polte *et al.*, *ACS Nano* **4**, 1076–1082 (2010).
27. D. Frenkel, A. J. C. Ladd, *J. Chem. Phys.* **81**, 3188–3193 (1984).
28. H. W. Sheng, M. J. Kramer, A. Cadien, T. Fujita, M. W. Chen, *Phys. Rev. B* **83**, 134118 (2011).

ACKNOWLEDGMENTS

Supported by the Physical Chemistry of Inorganic Nanostructures Program (KC3103), Office of Science, Office of Basic Energy Sciences, U.S. Department of Energy (DOE) under contract

DE-AC02-05CH11231 (J.P. and A.P.A.); NSF grant DMR-1310266, Harvard Materials Research Science and Engineering Center grant DMR-1420570, and Amore Pacific (J.P., S.H.H., and D.A.W.); the Multimodal Australian Sciences Imaging and Visualization Environment (www.massive.org.au) and funds from Monash University (H.E.); the DOE Office of Energy Research, Basic Energy Sciences, Materials Sciences and Engineering Division under contract DE-AC02-05CH11231 within the SP2-Bonded Materials Program and the Molecular Foundry (construction of GLC and TEM characterization), Office of Naval Research grant

NO0014-12-1 (graphene growth), NSF grant DMR-1206512 (graphene transfer methods development), and postdoctoral support from Defense Threat Reduction Agency grant HDTRA1-13-1-0035 (J.M.Y., K.K., and A.Z.); the Princeton Center for Theoretical Science (D.T.L.); a Miller fellowship from Miller Institute for Basic Research in Science at UC Berkeley (Q.C.); and the Basic Science Research Program through the National Research Foundation of Korea funded by Ministry of Education grant NRF-2014R1A1A2058178 (K.K.). Electron microscopy was performed at the Molecular Foundry supported by DOE contract DE-AC02-05CH11231.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/290/suppl/DC1
Molecular Dynamics Simulation
Materials and Methods
Figs. S1 to S9
References (29–38)
Movies S1 to S4

16 March 2015; accepted 4 June 2015
10.1126/science.aab1343

ANIMAL PHYSIOLOGY

Summer declines in activity and body temperature offer polar bears limited energy savings

J. P. Whiteman,^{1,2*} H. J. Harlow,² G. M. Durner,³ R. Anderson-Sprecher,⁴ S. E. Albeke,⁵ E. V. Regehr,⁶ S. C. Amstrup,⁷ M. Ben-David^{1,2}

Polar bears (*Ursus maritimus*) summer on the sea ice or, where it melts, on shore. Although the physiology of “ice” bears in summer is unknown, “shore” bears purportedly minimize energy losses by entering a hibernation-like state when deprived of food. Such a strategy could partially compensate for the loss of on-ice foraging opportunities caused by climate change. However, here we report gradual, moderate declines in activity and body temperature of both shore and ice bears in summer, resembling energy expenditures typical of fasting, nonhibernating mammals. Also, we found that to avoid unsustainable heat loss while swimming, bears employed unusual heterothermy of the body core. Thus, although well adapted to seasonal ice melt, polar bears appear susceptible to deleterious declines in body condition during the lengthening period of summer food deprivation.

The current rate of Arctic sea-ice loss, unprecedented in at least the past several thousand years, is outpacing predictions and accelerating (1). This raises concerns about the persistence of polar bears (*Ursus maritimus*) (2), which hunt on the surface of the sea ice, most successfully between April and July (3), when ringed seals (*Pusa hispida*) use this substrate for rearing pups and molting (4). Between August and October, hunting can be poor (5) as seals reduce ice surface time (4). Additionally, in about two-thirds of the polar bear range (6), seals become largely pelagic as ice retreats from the continental shelf (7, 8). Some polar bears spend this period on shore, where foraging is also usually limited (9).

To reduce the loss of body condition during summer food deprivation, shore bears purportedly enter a state of lowered activity and resting metabolic rate similar to winter hibernation but without denning (10). This “walking hibernation” could partially compensate for the nega-

tive impacts of extended ice melt (11). However, in western Hudson Bay, Canada, shore bears lose body mass at a rate indicative of typical, rather

than hibernation-like, metabolism (12). The physiological state of bears that follow the retreating sea ice into the central Arctic basin in summer is unknown. In addition, recent sea-ice loss may be increasing the frequency of long-distance swims by polar bears (13), during which they risk losing over 10 times more heat than they produce (supplementary text) because their fur loses 90% of its insulation value when wet (14), and their subcutaneous fat does not provide blubber-like insulation (15).

To understand polar bear responses to these challenges of summer ice melt, we investigated activity on shore (2008 and 2009) and on ice (2009) in the Beaufort Sea (Fig. 1) by affixing telemetry transmitters and activity loggers (16) to 25 females (mean age = 10 years \pm 1 SE, age range = 4 to 20 years) and one male (age 3). We recorded temperatures of the body core (an index of metabolic rate) (17) and periphery by implanting loggers into the abdomens (core) of 10 bears (nine females, mean age = 11 years \pm 2 SE, age range = 3 to 23 years; one male, age 6) and the rumps (periphery) of seven other individuals (six females, mean age = 9 years \pm 2 SE, age range = 5 to 20 years; one male, age 2).

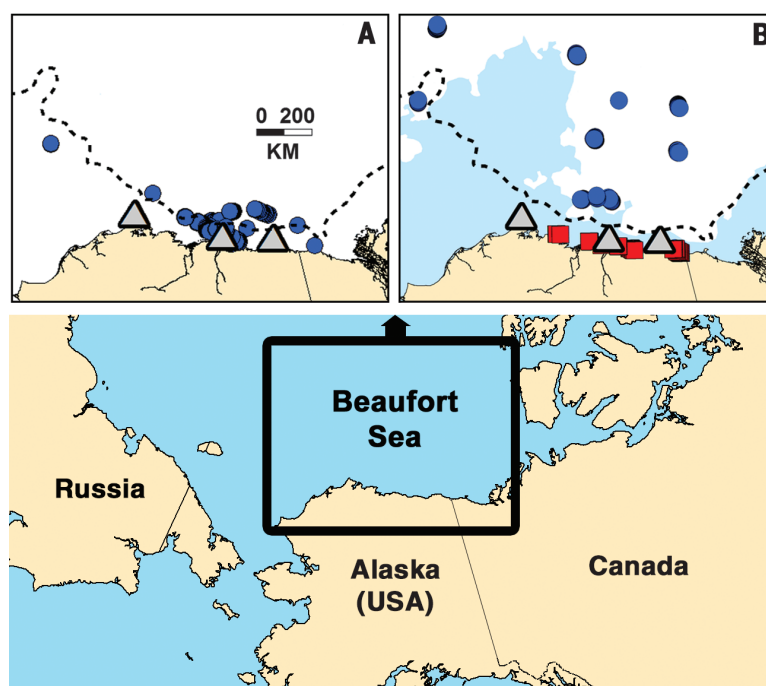


Fig. 1. The western Arctic on (A) 11 May 2009 and (B) 31 August 2009. Locations are shown for ice polar bears (blue circles), shore polar bears (red squares), and whale carcasses (triangles). The 300-m depth contour is shown as a dashed line. Sea ice is shown in white.

¹Program in Ecology, University of Wyoming, Laramie, WY 82071, USA. ²Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA. ³U.S. Geological Survey, Alaska Science Center, Anchorage, AK 99508, USA. ⁴Department of Statistics, University of Wyoming, Laramie, WY 82071, USA. ⁵Wyoming Geographic Information Science Center, University of Wyoming, Laramie, WY 82071, USA. ⁶Marine Mammals Management, U.S. Fish and Wildlife Service, Anchorage, AK 99503, USA. ⁷Polar Bears International, Bozeman, MT 59772, USA.

*Corresponding author. E-mail: jwhitema@uwyo.edu

Bears on shore and ice exhibited similar activity patterns (Fig. 2A and tables S1 and S2; in table S1, shore data from 2008 and 2009 were pooled for lack of difference). Their time spent active peaked at ~25% between May and July, then fell to 12 to 22% between August and October (Fig. 2A), remaining greater than the values of 1 to 2% previously observed during winter hibernation (18). The maximum activity level we measured (~25%) is approximately half that observed in all other bear species (19), perhaps reflecting polar bears' specialization for hunting large vertebrate prey (3). The high energy costs of finding and subduing such prey can be reduced by ambush tactics (20), such as still-hunting at seal breathing holes (3).

Although bears on shore and ice were similarly active between August and October, movement rates were higher for ice bears (Fig. 2B and table S1), potentially because the sea ice underfoot was drifting at 0.30 to 0.60 m/s (21). The proportion of rapid movement rates (>0.33 m/s) recorded when ice bears were resting (i.e., motionless for ≥98% of the previous half hour) (16) increased in late summer (Fig. 2C), suggesting that as bears reduced their activity, their movement rates increasingly reflected ice drift.

We did not find support for the hypothesis that lowered activity of shore and ice bears is a response to decreased food availability. For ice bears, location relative to primary seal habitat over the continental shelf (7, 8) was a poor predictor of activity (mean model coefficients overlapped with zero; $n = 23$ bears) (table S3). This result suggests that additional factors determine seal distribution (e.g., fine-scale variation in productivity) (8) or their availability as prey (e.g., time spent on the ice surface) (4). Similarly, access to concentrated food resources for shore bears [i.e., locations within 500 m of bowhead whale (*Balaena mysticetus*) carcasses from Inuit subsistence harvest (table S4)] was not associated with activity levels ($n = 7$ bears) (table S3).

Bears may reduce activity to avoid heat stress in summer, because their large body size and low ratio of surface area to volume hinder heat dissipation. In a previous study, captive polar bears became hyperthermic while walking on treadmills at speeds ≥1.6 m/s when air temperature was ≥-5°C (22). However, air temperature in our study (daily means: -14.8° to 15.1°C) was unrelated to activity of shore bears ($n = 12$ bears) (table S3). Also, only 31 of our 61,882 measurements of movement rate (shore and ice combined; $n = 25$ bears) were ≥1.6 m/s, indicating that free-ranging bears seldom walk that rapidly. Thus, locomotion-induced heat stress is probably rare and insufficient to explain the reduction in summer activity.

Core body temperatures (monthly means of smoothed data) (16) did not differ between shore and ice bears during summer (Fig. 2D, table S1, and fig. S1), suggesting that all bears had similar nonhibernating, resting metabolic rates (fig. S2A). Mean core temperatures of ice bears gradually declined from May (37.3°C) to September (36.6°C)

(Fig. 2D). It is unclear whether a similar trend occurred in shore bears, because they were implanted with loggers in August.

These gradual temperature declines correlated with activity (mean $r = 0.31$, $n = 9$ bears) (table S5) (16) and may have been associated with fasting (5). Fasting can cause progressive reduction in body temperature of ~1°C in mammals (23, 24), corresponding with up to ~20% decreases in whole-body metabolic rate (23–25). However, the reduction in mass-specific metabolic rate is smaller and sometimes nonexistent after the loss of metabolically active tissue is considered (24). Unfortunately, we were unable to assess mass changes because we captured bears in spring before they reached peak mass (table S6). Reduced insulation from thinning of fur and subcutaneous fat could also cause temperature declines, although warm summer conditions could counteract insulation loss. Hence, gradual declines in the summer core temperature of polar bears suggest reductions in energy expenditure typical of food-deprived mammals (24).

Data from one pregnant female that retained her logger through January provide evidence that polar bears in maternal dens exhibit hibernation core temperatures in winter. In contrast to the gradual declines observed in summer, her core temperature abruptly fell to ~35°C after 28 November (Fig. 2D and fig. S2B), as would be expected after parturition (3, 26). Such low temperature presumably reflects a 50 to 80% reduction in metabolic rate during winter hibernation, similar to that seen in other ursids (fig. S2A) (26, 27).

Core temperatures, like movement rates, indicated that polar bears did not experience heat stress in summer. When walking on a treadmill at ≥1.6 m/s, captive polar bears frequently exhibited temperatures >39.0°C, including uncontrolled rises to >40.0°C, leading to the conclusion that they store excess heat and are inefficient walkers (22). However, only 18 of our 27,843 measurements ($n = 10$ bears) were >39.0°C and none were >40.0°C. This suggests that polar bear heat storage and locomotion efficiency should be reassessed and that polar bears thermoregulate effectively during summer.

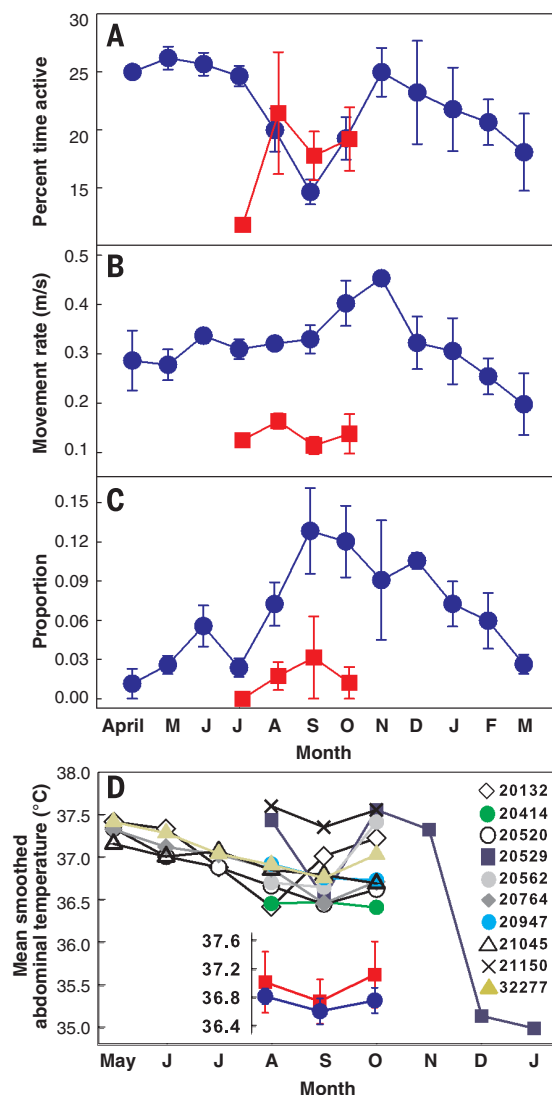


Fig. 2. Polar bear activity and abdominal temperature. (A and B) Grand means (±SE) of activity and movements for bears on sea ice (circles) and on shore (squares), April to March (2008–2010). **(C)** Proportion of movement rates >0.33 m/s recorded when bears were inactive. **(D)** Individual means (±95% CI) from May to October for ice bears (except location unknown for bear 20132) and from August to October for shore bears. The inset depicts grand monthly means from August to October on ice (circles) and on shore (squares). Sample sizes are in table S7 ($n = 1$ for July on shore); raw data are in fig. S1.

We also observed brief bouts of cold core temperatures ($<35.0^{\circ}\text{C}$, typically ≤ 12 hours) as low as 22.3°C . These temperature changes were too rapid (up to $\pm 5.0^{\circ}\text{C}$ per hour) to represent fluctuations in whole-body metabolic rate. Such bouts occurred in five of five shore bears and one of four ice bears. Similar peripheral temperature changes were recorded by rump loggers in six shore bears and one ice bear, although maximum hourly swings were greater for the periphery (means: $+11.8^{\circ}\text{C}$ and -10.4°C) than for the core (means: $+5.0^{\circ}\text{C}$ and -5.0°C) (supplementary text).

It is unlikely that cold bouts of peripheral and core temperatures reflected consumption of ice or lying on it, as some occurred during above-freezing air temperature when sea ice was absent. Instead, cold bouts appeared to be caused by two modes of regional heterothermy: (i) cool-

ing the body periphery during inactivity and (ii) cooling the periphery and part of the core while swimming (Fig. 3, A to D and fig. S3). Supporting this distinction, peripheral temperatures fell below 35.0°C during both inactivity (49 of 800 measurements, $n = 3$ bears) and swimming (572 of 741 measurements, $n = 3$ bears). In contrast, core temperatures fell below 35.0°C only during swimming (6 of 11 measurements, $n = 1$ bear) (supplementary text) and never during inactivity (0 of 353 measurements, $n = 1$ bear). Furthermore, cold peripheral temperatures ($<35.0^{\circ}\text{C}$) were associated with lower activity and warmer collar temperatures (i.e., when the bear curled up, warming the collar sensor) than were cold core temperatures (Fig. 3, E and F).

Regional heterothermy of the body core is unusual (17) and may minimize heat loss while

swimming. Immersed polar bears probably reduce skin temperature to several degrees above the surrounding water (supplementary text) (15), similar to seals (28). Our data suggest that bears maintain an internal thermal gradient by temporarily cooling the outermost tissues of their core to form an insulating shell. A similar process occurs in diving king penguins (*Aptenodytes patagonicus*), another polar endotherm without blubber (29). Control of this process is probably active (e.g., via vasoconstriction), because abdominal loggers cooled more quickly than is possible passively, and temperatures were subsequently regulated (Fig. 4 and supplementary text). This regional heterothermy may represent an adaptation to long-distance swims (13), although its limits remain unknown. One of our bears survived a 9-day swim, but when recaptured 7 weeks later, she had lost 22% of her body mass and her cub (30).

Sea-ice loss (1) increasingly limits spring and summer hunting opportunities for polar bears in parts of their range (2). In the Beaufort Sea and elsewhere (2, 31), this has reduced the energy stores available for bears during subsequent food deprivation (2, 5). We found that both core temperature and activity remained above values observed during winter hibernation. The gradual declines in core temperature during summer suggest a typical mammalian response to fasting, which offers limited to no energy savings based on mass-specific metabolic rates (24). Thus, our data indicate that bears cannot use a hibernation-like metabolism to meaningfully prolong their summer period of fasting and reliance on energy stores. In conjunction with theoretical models linking normal metabolic rate to depletion of stored energy and mortality (32), our findings suggest that bears are unlikely to avoid deleterious declines in body condition, and ultimately survival, that are expected with continued ice loss and lengthening of the ice melt period (2).

Fig. 3. Body temperature of a swimming polar bear. (A)

Hourly abdominal (core) temperature of bear 20414 in 2009. A bout of cold temperatures (arrow) is expanded in (B).

(C) Temperatures from the collar and a weather station 110 km away during the same cold bout. (D) Collar acceleration scores during the same cold bout. (B) through (D) are divided into time periods 1 to 5.

In period 1, stationary locations and high collar temperatures indicate resting on shore and covering the collar. In 2, locations were offshore, collar temperatures were consistent with immersion, and acceleration suggests swimming. In 3 through 5, locations indicate walking along the coast, resting, then walking again. (E) Mean activity ($\pm 95\%$ CI) measured during bouts of cold temperatures ($<35.0^{\circ}\text{C}$) recorded from abdominal ($n = 69$, three bears pooled) and rump ($n = 259$, one bear) loggers. (F) Mean collar temperature ($\pm 95\%$ CI) measured during bouts of cold temperatures ($<35.0^{\circ}\text{C}$) recorded from abdominal ($n = 68$, three bears pooled) and rump ($n = 829$, six bears pooled) loggers.

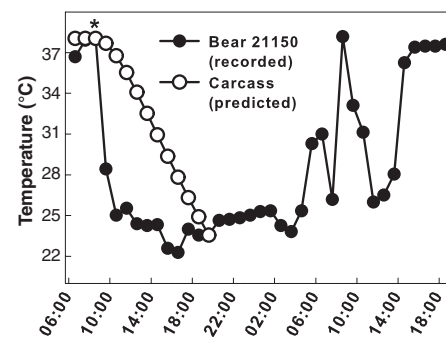
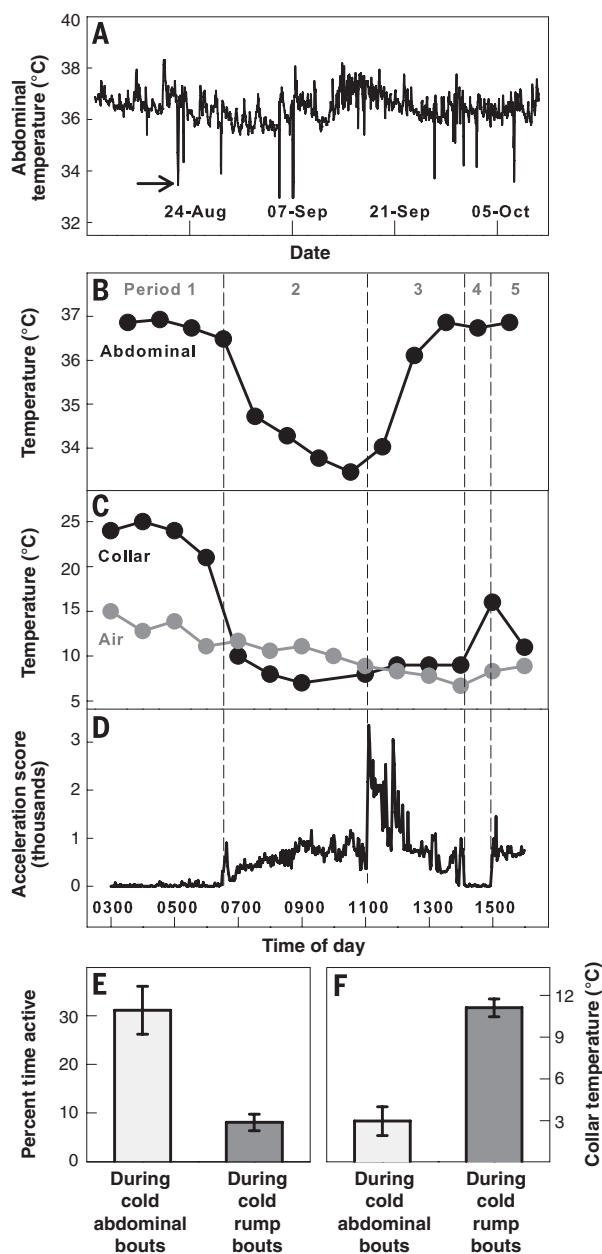


Fig. 4. Cooling curves of a polar bear. Hourly abdominal (core) temperatures [06:00 (UTC -08:00) on 2 October to 18:00 on 3 October 2009] represent swimming based on sparse location data (solid circles). Predicted intraperitoneal temperatures (open circles) represent cessation of heat production (death) at the asterisk and subsequent immersion in 4°C water (supplementary text).

REFERENCES AND NOTES

- W. N. Meier et al., *Rev. Geophys.* **52**, 185–217 (2014).
- I. Stirling, A. E. Derocher, *Glob. Chang. Biol.* **18**, 2694–2706 (2012).
- S. C. Amstrup, in *Wild Mammals of North America: Biology, Management, and Conservation*, G. A. Feldhamer, B. C. Thompson, J. A. Chapman, Eds. (The Johns Hopkins Univ. Press, Baltimore, ed. 2, 2003), pp. 587–610.
- B. P. Kelly et al., *Polar Biol.* **33**, 1095–1109 (2010).
- J. Whiteman, thesis, University of Wyoming, Laramie, WY (2014).
- S. C. Amstrup, B. G. Marcot, D. C. Douglas, in *Arctic Sea Ice Decline: Observations, Projections, Mechanisms, and Implications*, E. T. Deweaver, C. M. Bitz, L. B. Tremblay, Eds. (Geophysical Monograph 180, American Geophysical Union, Washington, DC, 2008), pp. 213–268.
- L. A. Harwood, T. G. Smith, J. C. Auld, *Arctic* **65**, 35–44 (2012).
- L. A. Harwood, I. Stirling, *Can. J. Zool.* **70**, 891–900 (1992).
- K. D. Rode, C. T. Robbins, L. Nelson, S. C. Amstrup, *Front. Ecol. Environ.* **13**, 138–145 (2015).
- R. A. Nelson et al., *Ursus* **5**, 284–290 (1983).
- M. G. Dyck et al., *Ecol. Complex.* **4**, 73–84 (2007).
- C. T. Robbins, C. Lopez-Alfaro, K. D. Rode, Ø. Tøien, O. L. Nelson, *J. Mammal.* **93**, 1493–1503 (2012).
- A. M. Pagano, G. M. Durner, S. C. Amstrup, K. S. Simac, G. S. York, *Can. J. Zool.* **90**, 663–676 (2012).
- P. F. Scholander, V. Walters, R. Hock, L. Irving, *Biol. Bull.* **99**, 225–236 (1950).
- C. M. Pond, C. A. Mattacks, R. H. Colby, M. A. Ramsay, *Can. J. Zool.* **70**, 326–341 (1992).
- Materials and methods are available as supplementary materials on Science Online.
- B. K. McNab, *The Physiological Ecology of Vertebrates: A View from Energetics* (Cornell Univ. Press, Ithaca, NY, 2002).
- F. Messier, M. K. Taylor, M. A. Ramsay, *J. Zool. (London)* **226**, 219–229 (1992).
- S. Paisley, D. L. Garshelis, *J. Zool. (London)* **268**, 25–34 (2006).
- T. M. Williams et al., *Science* **346**, 81–85 (2014).
- R. J. Galley, B. G. T. Else, S. J. Prinsenberg, D. Babb, D. G. Barber, *Arctic* **66**, 105–116 (2013).
- R. C. Best, *J. Comp. Physiol. B* **146**, 63–73 (1982).
- O. E. Owen, G. A. Reichard Jr., M. S. Patel, G. Boden, *Adv. Exp. Med. Biol.* **111**, 169–188 (1979).
- M. D. McCue, *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **156**, 1–18 (2010).
- E. W. Pfeiffer, L. N. Reinking, J. D. Hamilton, *Comp. Biochem. Physiol. A Physiol.* **63**, 19–22 (1979).
- A. Friebe et al., *PLOS ONE* **9**, e101410 (2014).
- Ø. Tøien et al., *Science* **331**, 906–909 (2011).
- I. L. Boyd, *J. Exp. Biol.* **203**, 1907–1914 (2000).
- Y. Handrich et al., *Nature* **388**, 64–67 (1997).
- G. M. Durner et al., *Polar Biol.* **34**, 975–984 (2011).
- J. F. Bromaghin et al., *Ecol. Appl.* **25**, 634–651 (2015).
- P. K. Molnár, A. E. Derocher, G. W. Thiemann, M. A. Lewis, *Biol. Conserv.* **143**, 1612–1622 (2010).

ACKNOWLEDGMENTS

This study was funded by NSF (OPP 0732713), the U.S. Geological Survey (USGS) Climate and Land Use Change Research and Development Program, U.S. Fish and Wildlife Service Marine Mammals Management, Wyoming NASA Space Grant (NNG05G165H), the University of Wyoming, and the Environmental Protection Agency (EPA) Science To Achieve Results program (F91737301). This report was approved under USGS Fundamental Science Practices but not by the EPA. Views are solely those of the authors. Data are archived by the National Center for Atmospheric Research (<http://www.eol.ucar.edu/projects/arcss/>).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/295/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S5
Tables S1 to S7
References (33–64)

5 February 2015; accepted 11 June 2015
10.1126/science.aaa8623

THERMAL PHYSIOLOGY

Keeping cool: Enhanced optical reflection and radiative heat dissipation in Saharan silver ants

Norman Nan Shi,¹ Cheng-Chia Tsai,¹ Fernando Camino,² Gary D. Bernard,³ Nanfang Yu,^{1*} Rüdiger Wehner^{4**}

Saharan silver ants, *Cataglyphis bombycina*, forage under extreme temperature conditions in the African desert. We show that the ants' conspicuous silvery appearance is created by a dense array of triangular hairs with two thermoregulatory effects. They enhance not only the reflectivity of the ant's body surface in the visible and near-infrared range of the spectrum, where solar radiation culminates, but also the emissivity of the ant in the mid-infrared. The latter effect enables the animals to efficiently dissipate heat back to the surroundings via blackbody radiation under full daylight conditions. This biological solution for a thermoregulatory problem may lead to the development of biomimetic coatings for passive radiative cooling of objects.

The silver ants of the Sahara desert, *Cataglyphis bombycina*, inhabit one of the hottest terrestrial environments on Earth, where they occupy the ecological niche of a “thermophilic scavenger” (1). In wide-ranging foraging journeys, they search for corpses of insects and other arthropods that have succumbed to the thermally harsh desert conditions, which they themselves are able to withstand more successfully. On hot summer days, they may reach maximal foraging activities when temperatures of the desert surface are as high as 60° to 70°C and their body temperatures measured as “operative environmental temperatures” are in the range of 48° to 51°C (2, 3). In order to survive under these conditions, occasionally the ants must unload excess heat by pausing on top of stones or dry vegetation, where, because of the steep temperature gradient above the sand surface, they encounter considerably lower temperatures. Under the midday sun of a summer day, the ants may resort to this thermal respite (cooling off) up to 70% of their entire foraging time (3). In keeping their body temperature below their critical thermal maximum of 53.6°C (4), they need not only to reduce heat absorption from the environment but also to be able to efficiently dissipate excess heat, so that they can minimize the amount of time spent in thermal refuges.

As we showed, through a series of optical and thermodynamic measurements, full-wave simulations, and heat-transfer modeling, a dense array of triangularly shaped hairs, characteristic of *Cataglyphis bombycina*, enables the ants to main-

tain lower body temperatures by (i) reflecting a large portion of the solar radiation in the visible and near-infrared (NIR) range of the spectrum and (ii) radiating heat to the surrounding environment by enhancing the emissivity in the mid-infrared (MIR), where the blackbody radiation spectrum of the ant's body culminates. The thermoregulatory solutions that the silver ants have evolved to cope with thermally stressful conditions show that these animals are able to control electromagnetic waves over an extremely broad range of the electromagnetic spectrum (from the visible to the MIR) and that different physical mechanisms are employed in different spectral ranges to realize an important biological function.

Specimens of *Cataglyphis bombycina* collected in Tunisia (34°10'N, 08°18'E) were used for all of the optical and thermodynamic measurements. In these ants, the dorsal and lateral sides of the body have a silvery glare (Fig. 1A) and are covered by dense and uniform arrays of hairs (Fig. 1B and fig. S4). As scanning electron microscopy (SEM) images show, the hairs, which gradually taper off at the tip, are locally aligned in the same direction (Fig. 1C). Their most remarkable structural feature is the triangular cross-section characterized by two corrugated top facets and a flat bottom facet facing the ant's body (Fig. 1, D and E). Cross-sectional views obtained by focused ion beam (FIB) milling show that the gap between the bottom hair facet and the cuticular surface also varies but is generally larger than a few hundred nanometers.

Optical reflectivity measurements of ant specimens were obtained with two Fourier transform spectrometers, one collecting spectra in the visible and NIR (wavelengths from 0.45 to 1.7 μm) and the other in the MIR (wavelengths from 2.5 to 16 μm). The visible and NIR measurements showed that hemispherical reflection [i.e., the sum of specular and diffuse reflection collected through an integrating sphere (2)] is substantially enhanced

¹Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA. ²Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY, USA. ³Department of Electrical Engineering, University of Washington, Seattle, WA, USA. ⁴Brain Research Institute, University of Zürich, Zürich, Switzerland.

*Corresponding author. E-mail: ny2214@columbia.edu (N.Y.); rwehner@zool.uzh.ch (R.W.)

in regions with intact hair coverage as compared to regions from which the hairs had been removed (Fig. 2A and fig. S1). The hair-covered region reflects 67% of the incoming solar radiation rather than only 41%, as is the case after their removal. This enhancement is due to scattering within the triangular hairs (Mie scattering), where light gets trapped and then reradiates out in all directions (5–7). Individual hairs of given cross-sectional dimensions generate enhanced reflection due to scattering at specific wavelengths where fundamental and higher-order Mie resonance modes are supported (Fig. 2C and fig. S5). Due to the variation in cross-sectional areas, resonance peaks from individual hairs are averaged out, so that the hair cover effectively acts as a coating with enhanced broadband reflection.

Because of the ellipsoidal shape of the ant's body, about two-thirds of the dorsolateral surface is obliquely hit by solar radiation (fig. S4). This prompted us to examine the reflectivity as a function of the incidence angle of radiation, which was varied from 0° to 80°, with 0° representing the direction normal to the surface. As the results show, Mie scattering enhances reflectivity over all

angles when regions with intact hair cover are compared to those with hairs removed (Fig. 2B). With increasing angle of incidence, reflectivity enhancement becomes particularly strong at beyond 30°. This is the critical angle at which total internal reflection starts to occur at the bottom facets of the hairs (Fig. 2D, II). At angles approaching 90°, reflectivity drops off when total internal reflection at one of the top facets starts to direct more of the radiation toward the ant's body (Fig. 2D, III).

Next, we performed finite-difference time-domain (FDTD) simulations in order to demonstrate the functional significance of the triangular cross section of the hairs in enhancing reflectivity in the visible and NIR ranges (figs. S5 to S8). These simulations compared the reflective properties of triangular and circular hairs of the same cross sectional area. Even though the enhancement of reflectivity at normal incidence is comparable in both cases, triangular hairs produce an extra enhancement at higher angles of incidence (Fig. 2B). The reason is that although Mie scattering of similar strength occurs in both circular and triangular hairs, in the latter the total

internal reflection at the bottom facets of the hairs enhances reflectivity substantially further. The high reflectivity disappeared when the specimens were wetted by an ethanol-water solution (fig. S3), which removed the refractive index contrast between air and hairs and thus destroyed the conditions required for Mie scattering and total internal reflection.

Reflectivity measurements performed in the MIR range revealed a second important point in the silver ant thermotolerance story. When proceeding from lower to higher wavelengths, at about 8 μm , the enhanced reflectivity of regions with hair cover as compared to those without hairs reverses to reduced reflectivity (Fig. 2E). As Kirchhoff's law of thermal radiation states, reduced reflectivity corresponds to enhanced emissivity (8). At a body temperature of 50°C, which the silver ants may reach when foraging at peak activity times, the blackbody radiation of the ant's surface would lie in the range of 6 to 16 μm (peaked at ~9 μm) and thus allow the silver ants to offload heat more efficiently through radiative heat transfer. The latter decreases the steady-state body temperature and reduces the respite time.

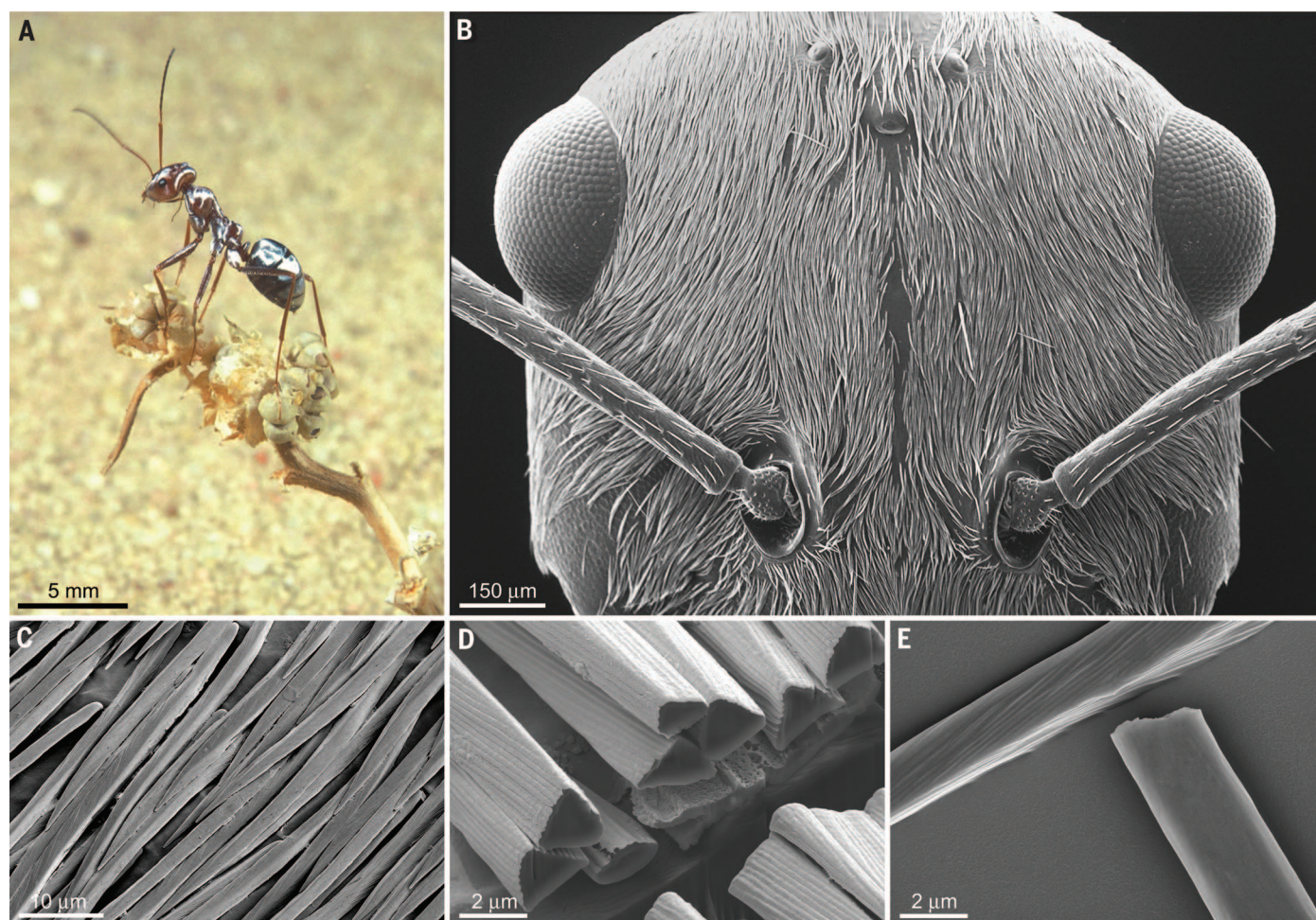
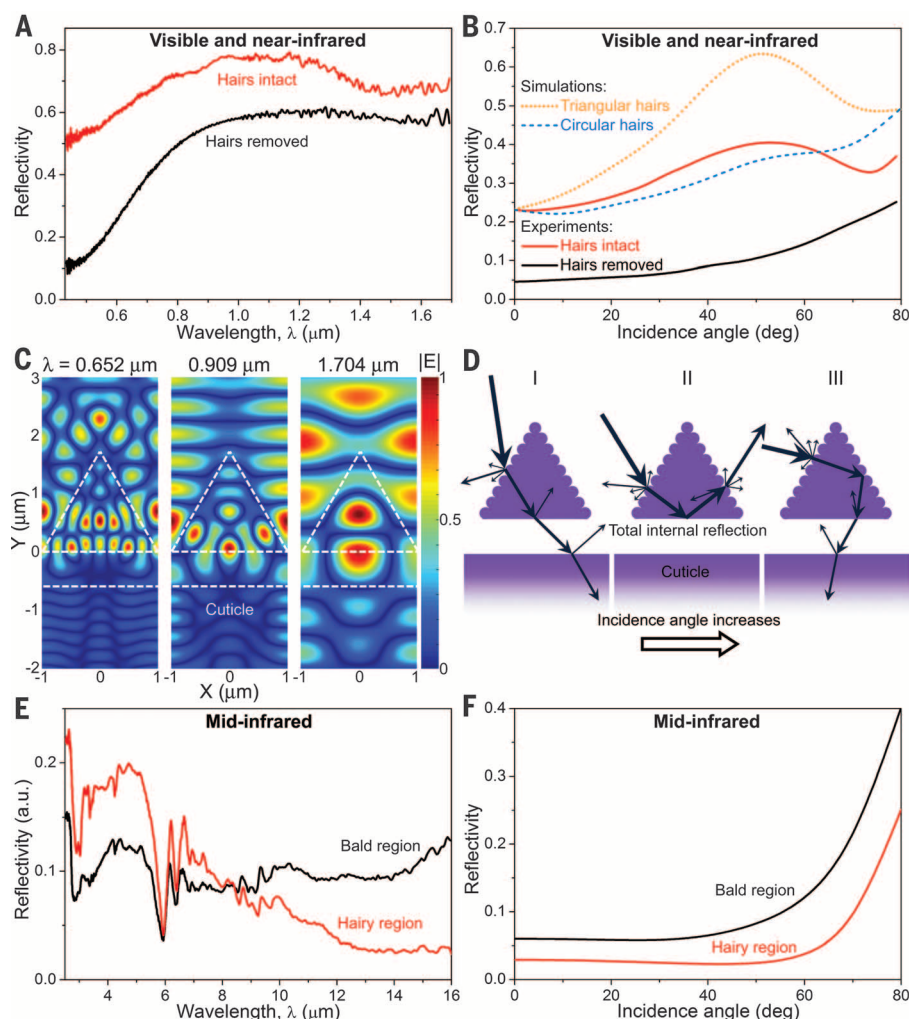


Fig. 1. The bright glare of the silver ant and its structural basis. (A) Silver ant offloading heat on top of dry vegetation (18). (B) SEM frontal view of the head densely covered by hairs. (C) SEM image of the hairs gradually tapering off toward the tip. (D) Cross-sectional view of the hairs milled with FIB. (E) SEM image of two hairs with one flipped upside down to exhibit the flat bottom facet.

Fig. 2. Reflectivity of the silver ant's body surface from the visible to the MIR range of the spectrum. (A) Hemispherical reflectivity measured in the visible and NIR. **(B)** Measurement and simulation results showing visible and NIR reflectivity as a function of incidence angle. **(C)** Cross-sectional view of a two-dimensional distribution of a light field (magnitude of electric field component of light, or $|E|$) around a triangular hair for three exemplary Mie resonances. **(D)** Schematic diagram showing the interaction between visible and NIR light and a hair at small (I), intermediate (II), and large (III) incidence angles. The corrugated upper two facets may enhance diffuse reflection in the ultraviolet and visible ranges. **(E)** Reflectivity measured in the MIR at normal incidence. **(F)** Simulated MIR reflectivity as a function of incidence angle.



How does the hair cover with its enhanced reflectivity in the visible and NIR and its enhanced emissivity in the MIR affect the radiative heat transfer between the ant's body and the environment? To investigate this question, we performed thermodynamic experiments, which mimicked all radiative heat transfer effects in the silver ants' natural foraging environment (fig. S2). To accomplish this task, we used a high-power xenon lamp to simulate the solar spectral distribution at the desert sand surface (9) and a thermoelectrically cooled high-emissivity metal plate to simulate the clear sky with its low level of blackbody radiation (10). The ant specimens were suspended on thin threads to minimize thermal conduction. Thermodynamic experiments were conducted in vacuum to study thermal radiation, as well as in still air to study the interplay of thermal radiation and convection. Under both conditions, the specimens with their natural hair covers were able to maintain significantly lower steady-state body temperatures than the same specimens with the hairs removed (Fig. 3).

The thermodynamic experiments conducted in vacuum comparing specimens before and after hair removal further revealed that the hair

cover decreases the time constants of temperature change (Fig. 3, B and E). The shortened time constants indicate an increased rate of radiative heat transfer and are a direct confirmation of the effect of the hairs in enhancing the MIR emissivity. By using the time constants and the heat transfer model, we computed that the hair cover enhances emissivity by about 15% (table S1). This enhanced emissivity is due to the fact that at large MIR wavelengths (i.e., at wavelengths much larger than the dimensions of the cross sections of the hairs), the hair structure acts as a gradient refractive index layer (fig. S9) (11–13), which provides the surface with broadband, broad-angle antireflective properties in the MIR (Fig. 2, E and F). Because of the influence of convection, the time constants of temperature change decreased by a factor of about 3 when the specimens were brought from vacuum into air (Fig. 3). This indicates that radiative heat dissipation amounted to about one-half of convection and, therefore, still played a significant role in the presence of natural convection.

Applying experimentally extracted parameters to the heat transfer model revealed that the enhanced visible and NIR reflectivity and enhanced

MIR emissivity make comparable contributions to reducing the steady-state temperature in the presence of natural convection (figs. S10 and S11). On hot summer days in the Sahara, the foraging activities of silver ants often occur under low wind or even still air conditions, when the ants have to rely on enhanced visible and NIR reflectivity and enhanced MIR emissivity equally heavily to reduce their body temperature during the respite behavior.

It is interesting to note that the hairs cover only the top and the sides of the ant's body, where they are responsible for the effects described above. The absence of hairs on the bottom surface reduces the radiative energy transfer between the hot sand and the cooler ant body, so that the animals can reduce the absorption of blackbody radiation from the desert floor.

In conclusion, Saharan silver ants are covered with a dense array of triangular hairs on the top and sides of their bodies. These silvery hairs protect the ants against getting overheated in at least three ways. First, as a result of Mie scattering and total internal reflection, the hairs enhance reflectivity in the visible and NIR, where solar radiation culminates. Second, in the MIR,

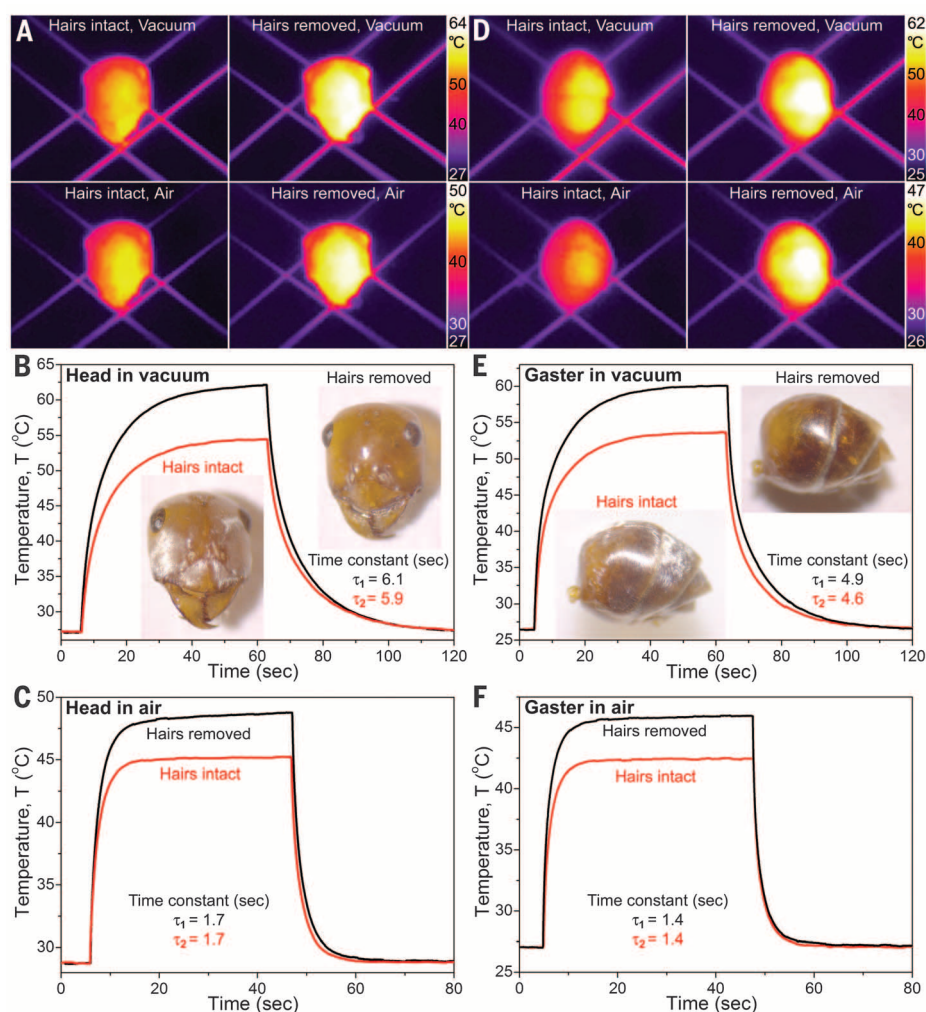


Fig. 3. Results of thermodynamic experiments.

(A) Thermal camera images showing the head of an ant specimen at the thermal steady state under different conditions. Temporal temperature profiles measured for the head before and after hair removal in vacuum (B) and in still air (C) are shown. (D to F) Results obtained for the hind part (gaster) of an ant specimen. Insets in (B) and (E) are photos of specimens before and after hair removal. In the “hairs intact” pictures of head and gaster, because of the limited solid angle of illumination, the silvery glance is not shown all over the body surface portrayed in the figures.

where solar radiation becomes negligible for wavelengths $>2.5 \mu\text{m}$, the hairs acting as an antireflection layer enhance emissivity and thus increase the ants' ability to offload excess heat via blackbody radiation. Third, the ants' bare bottom surface reflects MIR radiation from the hot desert floor more efficiently than if it were covered by hairs. Taken together, these effects result in decreasing the ants' steady-state body temperature and thus enable these thermophilic scavengers to forage at exceedingly high environmental temperatures. Finally, the present interdisciplinary account on the silver ants could have a significant technological impact by inspiring the development of biomimetic coatings for the passive cooling of objects (14–16). A recent article reported the demonstration of a multilayered film that can cool down an object by using essentially the same mechanisms as the silver ants: high reflectivity in the solar spectrum and high emissivity in the MIR (17).

REFERENCES AND NOTES

1. R. Wehner, S. Wehner, *Physiol. Entomol.* **36**, 271–281 (2011).
2. Materials and methods are available as supplementary materials on Science Online.
3. R. Wehner, A. C. Marsh, S. Wehner, *Nature* **357**, 586–587 (1992).
4. W. J. Gehring, R. Wehner, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 2994–2998 (1995).
5. C. F. Bohren, D. R. Huffman, *Absorption and Scattering of Light by Small Particles* (Wiley, New York, 1998).
6. J. Schuller, T. Taubner, M. L. Brongersma, *Nat. Photonics* **3**, 658–661 (2009).
7. D. Lin, P. Fan, E. Hasman, M. L. Brongersma, *Science* **345**, 298–302 (2014).
8. J. R. Howell, R. Siegel, M. P. Mengüç, *Thermal Radiation Heat Transfer* (CRC Press, Boca Raton, FL, ed. 5, 2010).
9. K. L. Coulson, *Solar and Terrestrial Radiation: Methods and Measurements* (Academic Press, New York, 1975).
10. J. Monteith, M. Unsworth, *Principles of Environmental Physics* (Academic Press, Oxford, ed. 3, 2007).
11. C.-H. Sun, P. Jiang, B. Jiang, *Appl. Phys. Lett.* **92**, 061112 (2008).
12. J.-Q. Xi et al., *Nat. Photon.* **1**, 176–179 (2007).
13. M. J. Minot, *J. Opt. Soc. Am.* **66**, 515–519 (1976).
14. C. G. Granqvist, A. Hjortsberg, *Appl. Phys. Lett.* **36**, 139–141 (1980).
15. P. Berdahl, *Appl. Opt.* **23**, 370–372 (1984).
16. E. Rephaeli, A. Raman, S. Fan, *Nano Lett.* **13**, 1457–1461 (2013).
17. A. P. Raman, M. A. Anoma, L. Zhu, E. Rephaeli, S. Fan, *Nature* **515**, 540–544 (2014).
18. R. Wehner, *Jahrb. Akad. Wiss. Lit. Mainz* **89**, 101–112 (1989).

ACKNOWLEDGMENTS

We acknowledge intriguing discussions with N. Pierce; help with experiments from Z. Li, M.-H. Kim, B. Patterson, and M. Y. Sfeir; and R. F. Foelix for kindly preparing and providing Fig. 1B. The work was supported by NSF (grants PHY-1411445 and ECCS-1307948) and the Air Force Office of Scientific Research, Multidisciplinary University Research Initiative program (grant FA9550-14-1-0389). Research was carried out in part at the Center for Functional Nanomaterials, Brookhaven National Laboratory, which is supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under contract no. DE-SC0012704. Data reported in this paper are archived at <http://datadryad.org/resource/doi:10.5061/dryad.2bm50>. Author contributions are as follows: R.W. and G.D.B. initiated the study, N.N.S., C.-C. Tsai, and N.Y. designed the study and conducted the analyses, and all authors contributed to developing the study and to drafting the manuscript.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/298/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S11
References (19–29)
Movie S1

15 April 2015; accepted 11 June 2015
Published online 18 June 2015
10.1126/science.aab3564

PLANT ECOLOGY

Worldwide evidence of a unimodal relationship between productivity and plant species richness

Lauchlan H. Fraser,^{1*} Jason Pither,² Anke Jentsch,³ Marcelo Sternberg,⁴ Martin Zobel,⁵ Diana Askarizadeh,⁶ Sandor Bartha,⁷ Carl Beierkuhnlein,⁸ Jonathan A. Bennett,⁹ Alex Bittel,¹⁰ Bazartseren Boldgiv,¹¹ Ilsi I. Boldrini,¹² Edward Bork,¹³ Leslie Brown,¹⁴ Marcelo Cabido,¹⁵ James Cahill,⁹ Cameron N. Carlyle,¹³ Giandiego Campetella,¹⁶ Stefano Chelli,¹⁶ Ofer Cohen,⁴ Anna-Maria Csergo,¹⁷ Sandra Díaz,¹⁵ Lucas Enrico,¹⁵ David Ensing,² Alessandra Fidelis,¹⁸ Jason D. Fridley,¹⁹ Bryan Foster,¹⁰ Heath Garriss,²⁰ Jacob R. Goheen,²¹ Hugh A. L. Henry,²² Maria Hohn,²³ Mohammad Hassan Jouri,²⁴ John Klironomos,² Kadri Koorem,⁵ Rachael Lawrence-Lodge,²⁵ Ruijun Long,²⁶ Pete Manning,²⁷ Randall Mitchell,²⁰ Mari Moora,⁵ Sandra C. Müller,²⁸ Carlos Nabinger,²⁹ Kamal Naseri,³⁰ Gerhard E. Overbeck,¹² Todd M. Palmer,³¹ Sheena Parsons,¹⁰ Mari Pesek,¹⁰ Valério D. Pillar,²⁸ Robert M. Pringle,³² Kathy Roccaforte,¹⁰ Amanda Schmidt,¹ Zhanhuan Shang,²⁶ Reinhold Stahlmann,⁸ Gisela C. Stotz,⁹ Shu-ichi Sugiyama,³³ Szilárd Szentes,³⁴ Don Thompson,³⁵ Radnaakhand Tungalag,¹¹ Sainbileg Undrakhbold,¹¹ Margaretha van Rooyen,³⁶ Camilla Wellstein,³⁷ J. Bastow Wilson,^{25,38} Talita Zupo¹⁸

The search for predictions of species diversity across environmental gradients has challenged ecologists for decades. The humped-back model (HBM) suggests that plant diversity peaks at intermediate productivity; at low productivity few species can tolerate the environmental stresses, and at high productivity a few highly competitive species dominate. Over time the HBM has become increasingly controversial, and recent studies claim to have refuted it. Here, by using data from coordinated surveys conducted throughout grasslands worldwide and comprising a wide range of site productivities, we provide evidence in support of the HBM pattern at both global and regional extents. The relationships described here provide a foundation for further research into the local, landscape, and historical factors that maintain biodiversity.

Despite a long history of research, the nature of basic patterns between environmental factors and biological diversity remain poorly defined. A notable example is the relationship between plant diversity and productivity, which has stimulated a long-running debate (1–6). A classic hypothesis, the humped-back model (HBM) (7), states that plant species richness peaks at intermediate productivity, taking above-ground biomass as a proxy for annual net primary productivity (7–9). This diversity peak is driven by two opposing processes. In unproductive ecosystems with low plant biomass, species richness is limited by abiotic stress, such as insufficient water and mineral nutrients, which few species are able to tolerate. In contrast, in the productive conditions that generate high plant biomass, competitive exclusion by a small number of highly competitive species is hypothesized to constrain species richness (7–9). Other mechanisms that may explain the unimodal relationship between species richness and biomass include disturbance (7, 10), evolutionary history and dispersal limitation (11, 12), and the reduction of total plant density in productive communities (13).

Since its initial proposal, a range of studies have both supported and rejected the HBM, and three separate meta-analyses reached different

conclusions (14–17). Although this inconsistency may indicate a lack of generality of the HBM, it may instead reflect a sensitivity to study methodology, including the plant community types considered, the taxonomic scope, the range of site productivities sampled, the spatial grain and extent of analyses (17, 18), and the particular measure of net primary productivity used (19). The questions therefore remain open as to what the form of the relationship between diversity and productivity is, and whether the HBM serves as a useful and general model for grassland ecosystem theory and management.

We quantified the form and the strength of the richness-productivity relationship by using globally coordinated surveys (20), which yielded scale-standardized data and were distributed across 30 sites in 19 countries and six continents (Fig. 1). Collectively, our samples spanned a broad range of biomass production (from 2 to 5711 g m⁻²) and grassland community types, including natural and managed (pastures and meadows) grasslands over a wide range of climatic zones (temperate, Mediterranean, and tropical), and altitudes (lowland to alpine) (table S1). Our protocol involved sampling 64 1-m² quadrats within 8-m-by-8-m grids (18, 21). At each site, between 2 and 14 grids were sampled, thus resulting in 128 to 896 quad-

rats per site. In each 1-m² quadrat, we identified and counted all plant species and harvested above-ground biomass and plant litter. Litter production is a function of annual net primary productivity in grasslands and can have profound effects on the structure and functioning of communities, from altering nutrient cycling to impeding vegetative growth and seedling recruitment (22, 23), thereby also playing a major role in driving community structure. Indeed, the HBM was originally defined in terms of live biomass plus litter material (7, 8). Most of the sites in our survey were subject to some form of management, usually livestock grazing or mowing. In this respect, our sites are representative of most of the world's grasslands. Our sampling was conducted at

¹Department of Natural Resource Sciences, Thompson Rivers University, Kamloops, BC, Canada. ²Department of Biology, University of British Columbia, Okanagan campus, Kelowna, BC, Canada. ³Department of Disturbance Ecology, BayCEER, University of Bayreuth, Bayreuth, Germany. ⁴Department of Molecular Biology and Ecology of Plants, Tel Aviv University, Tel-Aviv, Israel. ⁵Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia. ⁶Faculty of Natural Resources College of Agriculture and Natural Resources, University of Tehran, Iran. ⁷MTA Centre for Ecological Research, Institute of Ecology and Botany, Vácraót, Hungary, and School of Plant Biology, University of Western Australia, Crawley, Australia. ⁸Department of Biogeography, BayCEER, University of Bayreuth, Bayreuth, Germany. ⁹Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada. ¹⁰Department of Ecology and Evolutionary Biology, University of Kansas, Manhattan, KS 66047, USA. ¹¹Department of Biology, National University of Mongolia, Ulaanbaatar, Mongolia. ¹²Department of Botany, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. ¹³Department of Agricultural, Food and Nutritional Sciences, University of Alberta, Edmonton, AB, Canada. ¹⁴Applied Behavioural Ecology and Ecosystem Research Unit, University of South Africa, Johannesburg, South Africa. ¹⁵Instituto Multidisciplinario de Biología Vegetal (IMBIV-CONICET) and Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba, España. ¹⁶School of Biosciences and Veterinary Medicine, University of Camerino, Camerino, Italy. ¹⁷School of Natural Sciences, Trinity College Dublin, Dublin, Ireland. ¹⁸Departamento de Botânica, UNESP - Univ. Estadual Paulista, Rio Claro, Brazil. ¹⁹Department of Biology, Syracuse University, Syracuse, NY 13210, USA. ²⁰Department of Biology, University of Akron, Akron, OH 44325, USA. ²¹Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA. ²²Department of Biology, University of Western Ontario, London, ON, Canada. ²³Department of Botany, Corvinus University of Budapest, Budapest, Hungary. ²⁴Department of Natural Resources, Islamic Azad University, Nour Branch, Iran. ²⁵Department of Botany, University of Otago, Dunedin, New Zealand. ²⁶International Centre for Tibetan Plateau Ecosystem Management, Lanzhou University, Lanzhou, China. ²⁷Institute of Plant Sciences, University of Bern, Altenbergrain 21, CH-3013, Bern, Switzerland. ²⁸Department of Ecology, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. ²⁹Faculty of Agronomy, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. ³⁰Department of Range and Watershed Management, Ferdowsi University of Mashhad, Iran. ³¹Department of Biology, University of Florida, Gainesville, FL 32611, USA. ³²Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. ³³Laboratory of Plant Ecology, Hirotsaki University, Hirotsaki, Japan. ³⁴Institute of Plant Production, Szent István University, Gödöllő, Hungary. ³⁵Agriculture and Agri-Food Canada, Lethbridge Research Centre, Lethbridge, AB, Canada. ³⁶Department of Plant Science, University of Pretoria, Pretoria, South Africa. ³⁷Faculty of Science and Technology, Free University of Bozen-Bolzano, Bolzano, Italy. ³⁸Landcare Research, Dunedin, New Zealand.

*Corresponding author. E-mail: lfraser@tru.ca

least 3 months after the last grazing, mowing, or burning event and at the annual peak of live biomass, which, when coupled with litter, constitutes a reliable measure of annual net aboveground production in herbaceous plant communities (24).

Our results strongly support the HBM of the plant richness-productivity relationship. By using a global-extent regression model ($N = 9631$ 1-m² quadrats) (21), we found that plant richness formed a unimodal relationship with productivity (Fig. 2A) that is characterized by a highly significant concave-down quadratic regression [negative binomial generalized linear model (GLM); Table 1]. This relationship was not sensitive to the statistical model used; the hump-backed relationship was also evident when we used a negative binomial generalized linear mixed model (GLMM) that accommodated the hierarchical structure of our sampling design (grids nested within sites; Table 1 and fig. S1).

At a sampling grain of 1 m², 19 of 28 site level analyses (68%) yielded significant concave-down relationships (table S2 and Fig. 2A). This contrasts markedly with the results of Adler *et al.* (1), who found only 1 of their 48 within-site analyses to be significantly concave-down. We also found the form of the productivity-diversity relationship to be robust to sampling grain: by using grains of 1 m² up to 64 m², each time maintaining a global extent, we consistently found a significant concave-down relationship, although the proportion of variation explained tended to decrease with increasing grain (fig. S2).

The HBM predicts a boundary condition or upper limit to diversity that, in any given site,

may not be realized for a variety of reasons (18). Consistent with this view, our global-extent association is characterized by a significant concave-down quantile regression (95th percentile) (Table 1), below which considerable scatter exists (Fig. 2A). This pattern was also insensitive to the statistical method used; a hierarchical Bayesian analysis that accommodated the nested sampling design and that enabled both the mean and the variance of species richness to be modeled more accurately against (log-transformed) biomass also revealed a significant 95th percentile quantile regression (fig. S3). Likewise, we found a significant, concave-down quantile regression (95th percentile) between the maximum (quadrat-scale) richness found within a grid and the total biomass of the same quadrat (Table 1 and fig. S4). Each of these approaches to characterizing boundary conditions suggests the existence of a “forbidden space,” wherein high productivity precludes high local diversity. Furthermore, they suggest that extremely low-productivity sites rarely accommodate high diversity.

Why do our data show a hump-backed relationship, whereas those of Adler *et al.* (1) and related studies (4, 6), do not? One possibility is that data limitations can thwart detection of the HBM (18). For example, the data used by Adler *et al.* differed from ours in the following potentially important ways: (i) They exhibited a maximum live biomass of only 1535 g⁻² (ours was 3374 g⁻²), (ii) litter was not included within the calculation of biomass, and (iii) sample size was limited to 30 quadrats per site (ours ranged from 128 to 894 quadrats per site; table S1). We conducted a form

of sensitivity analysis in which we reran our statistical analyses using random subsets of our data that were constrained to exhibit similar properties to those of the Adler *et al.* data set. Specifically, after limiting the overall data set to less than 1535 g⁻² and excluding litter, we randomly selected 30 quadrats per site 500 times, each time conducting the within-site regression analyses ($N = 30$ for each of the 28 site-level GLMs conducted per subsampling iteration). For each iteration, we also calculated the average range of biomass spanned by the 28 site-level relationships. Across the 500 iterations (one example set of outcomes is shown in Fig. 2B), the average proportion of significant concave-down, within-site regressions was 0.31 ± 0.003 (SEM), significantly less than our observed proportion of 0.68 (fig. S5). Moreover, when significant concave-down relationships were detected, they tended to span a broader range of biomass than the remaining forms (including nonsignificant relationships). Specifically, in 458 of the 500 iterations (92%), the mean biomass range of the concave-down regressions was larger than the mean of the remaining forms' biomass ranges (binomial test: $P < 2.2 \times 10^{-16}$). Last, the 48 within-site analyses of Adler *et al.* spanned, on average, a live biomass range of $428.7 \text{ g}^{-2} \pm 38.36$ (range of 89 to 1217 g⁻²). This is (i) less than half of the average range encompassed by our 28 site-level analyses shown in Fig. 2A (mean = $1067.5 \text{ g}^{-2} \pm 140.63$; range of 286 to 3256 g⁻²) and (ii) almost 50% narrower than the smallest average biomass range encompassed by our 500 random subset analyses (627.4 g^{-2}) (fig. S6). Taken together, these findings

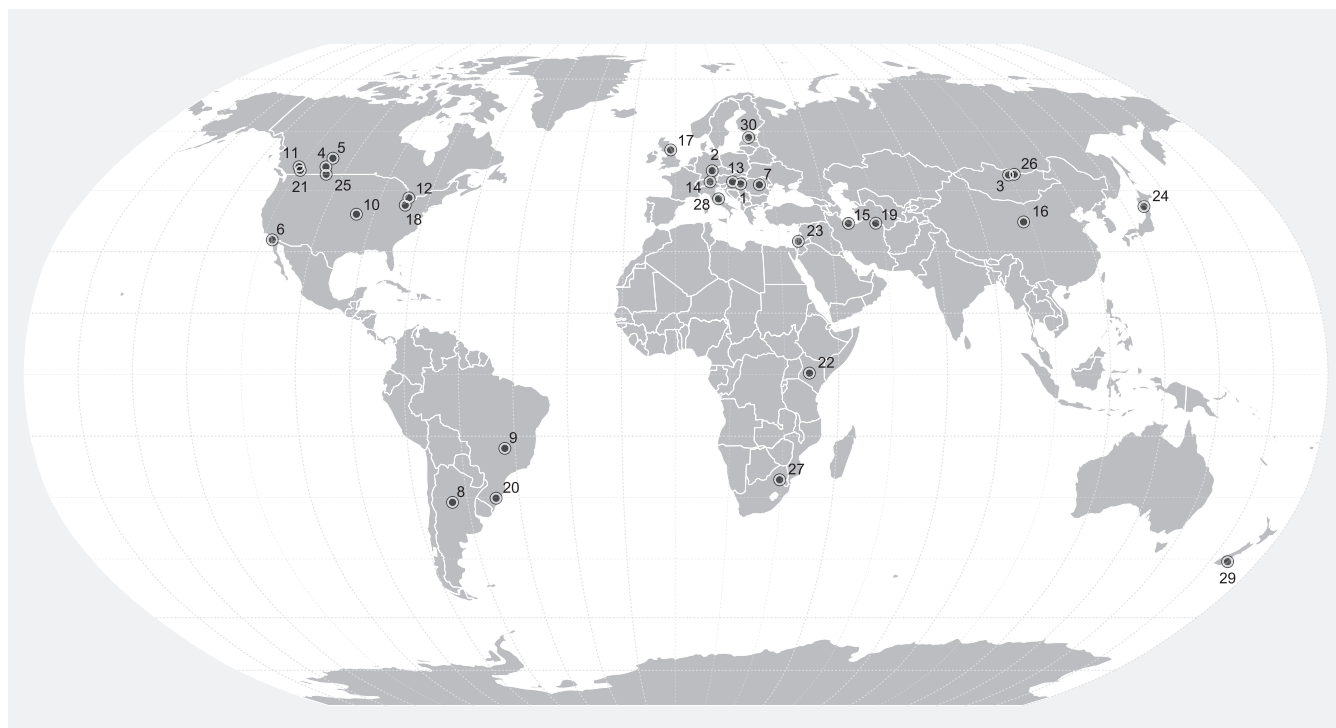


Fig. 1. Site locations. Locations of the geographic centroids of the 30 study sites, which include 151 sampling grids. Some points overlap and are therefore indistinguishable. Additional site details are provided in table S1. Map is displayed using the Robinson projection.

strongly suggest that we were able to detect more concave-down relationships because of the greater sample sizes and biomass ranges in our analysis.

It has been suggested (2) that some previous studies, including Adler *et al.* (1), failed to support the HBM because they excluded litter. Although we do find a significant concave-down relationship at the global extent using only live biomass (Table 1), a comparison of models using biomass versus biomass and litter (both $N = 9,631$) shows

total biomass to provide a far better fit [residual deviance = 10,105 (live) versus 10,037 (total); Vuong z -statistic for comparing non-nested models: -13.4 ; $P < 0.001$]. It has also been suggested that previous surveys failed to adequately represent high-productivity communities. Indeed, our high-biomass quadrats (1011 samples were over 1000 g^{-2} , $\sim 10\%$ of the 9631 samples; maximum of 5711 g^{-2}) contributed considerably to the right-hand part of the fitted humped-back regression. This could be a reason why the data set of Adler *et al.* (1) (in

which only 0.5% of samples were over 1000 g^{-2} with a maximum of 1534 g^{-2}) failed to support the HBM. Our results therefore show that a test of the HBM in herbaceous plant communities yields the expected pattern when it is robust and comprehensive, spans a wide range of biomass production (from 1 to at least $3000\text{ dry g}^{-2}\text{ year}^{-1}$), and provides sufficient replication of quadrats along the productivity gradient.

Competitive exclusion has been cited as the primary factor driving low species numbers at

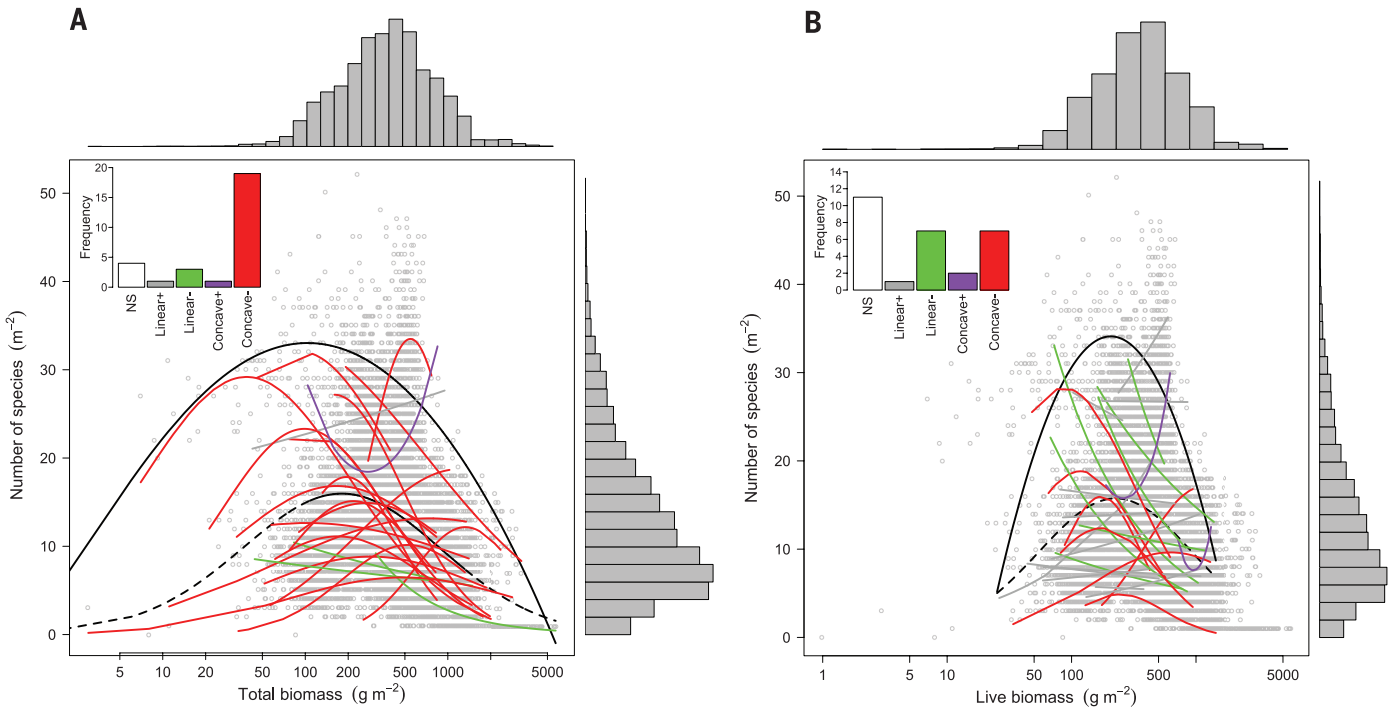


Fig. 2. Biomass production as a function of species richness. (A) Biomass production-species richness relationships for 28 study sites. Solid black line indicates significant quantile regression (95th percentile) of overall relationship (quadratic coefficient $P < 0.001$; $N = 9631$ quadrats). Dashed black line, significant negative binomial GLM (quadratic coefficient $P < 0.001$; $N = 9631$). Colored lines indicate significant GLM regressions (Poisson or quasi-Poisson), with N ranging from 128 to 894 quadrats.

(Inset) The frequencies of each form of relationship observed across study regions. NS, not significant. (B) Same as (A) but the results are derived from the analysis of an example, random subsample of the complete data set that satisfies the following criteria: litter biomass excluded, quadrats with biomass $>1534\text{ g}^{-2}$ excluded, and including 30 (randomly selected) quadrats per site (total $N = 840$). These criteria match the characteristics of the data set used by Adler *et al.* (1).

Table 1. Regression results. Results of regression analyses of the relationship between productivity and species richness, measured at a global extent and a sampling grain of 1-m ² quadrat. Total biomass = live biomass + litter biomass. All linear and quadratic term coefficients were highly significant ($P < 0.001$).						
Productivity measure	Type of regression	Sample size	Test of model fit	Intercept estimate \pm SEM	Linear term coefficient \pm SEM	Quadratic term coefficient \pm SEM
Total biomass	negative binomial GLM (log-link function)	9631 quadrats	likelihood ratio stat. = 1602.2	-2.52 ± 0.235	4.69 ± 0.186	-1.04 ± 0.037
Total biomass	negative binomial GLMM (log-link function) random effects: grid nested in site	9631 quadrats 151 grids 28 sites	likelihood ratio stat. = 114.0	0.91 ± 0.191	1.33 ± 0.133	-0.29 ± 0.028
Total biomass	quantile (95th percentile)	9631 quadrats	pseudo- F statistic = 179.1	-12.9 ± 7.159	45.6 ± 5.833	-11.3 ± 1.173
Live biomass	negative binomial GLM (log-link function)	9644 quadrats	likelihood ratio stat. = 950.3	-2.03 ± 0.212	4.27 ± 0.178	-0.96 ± 0.037

high plant biomass (7, 8, 25). However, in the case of nitrogen addition the negative relationship between productivity and species richness has been shown to diminish over time [(26), but see (27, 28)]. It may be that low species richness in high-productivity conditions arises in part because most such habitats are anthropogenic, and there are few species in the local pool adapted to these conditions (11, 12). If so, it is possible that species will eventually immigrate from distant pools, so that the right-hand part of the hump will then flatten out.

We have shown a global-scale concave-down unimodal relationship between biomass production and richness in herbaceous grassland communities. However, the original HBM (7) is vaguely articulated by the standards of modern ecological theory, and it is clear that more work is needed to determine the underlying causal mechanisms that drive the unimodal pattern (1, 6, 17, 18). We recognize that, in our study and many others, productivity accounts for a fairly low proportion of the overall variation in richness and that many other drivers of species richness exist (28–30). Accordingly, we echo the call of Adler *et al.* (1) for additional efforts to understand the multivariate drivers of species richness.

REFERENCES AND NOTES

- P. B. Adler *et al.*, *Science* **333**, 1750–1753 (2011).
- J. D. Fridley *et al.*, *Science* **335**, 1441 (2012).
- X. Pan, F. Liu, M. Zhang, *Science* **335**, 1441 (2012).
- J. B. Grace *et al.*, *Science* **335**, 1441 (2012).
- S. Pierce, *Funct. Ecol.* **28**, 253–257 (2014).
- J. B. Grace, P. B. Adler, W. S. Harpole, E. T. Borer, E. W. Seabloom, *Funct. Ecol.* **28**, 787–798 (2014).
- J. P. Grime, *J. Environ. Manage.* **1**, 151–167 (1973).
- M. M. Al-Mufti, C. L. Sydes, S. B. Furness, J. P. Grime, S. R. Band, *J. Ecol.* **65**, 759–791 (1977).
- Q. Guo, W. L. Berry, *Ecology* **79**, 2555–2559 (1998).
- J. H. Connell, *Science* **199**, 1302–1310 (1978).
- M. Zobel, M. Pärtel, *Glob. Ecol. Biogeogr.* **17**, 679–684 (2008).
- D. R. Taylor, L. W. Aarssen, C. Loehle, *Oikos* **58**, 239–250 (1990).
- J. Oksanen, *J. Ecol.* **84**, 293–295 (1996).
- G. G. Mittelbach *et al.*, *Ecology* **82**, 2381–2396 (2001).
- L. N. Gillman, S. D. Wright, *Ecology* **87**, 1234–1243 (2006).
- M. Pärtel, L. Laanisto, M. Zobel, *Ecology* **88**, 1091–1097 (2007).
- R. J. Whittaker, *Ecology* **91**, 2522–2533 (2010).
- L. H. Fraser, A. Jentsch, M. Sternberg, *J. Veg. Sci.* **25**, 1160–1166 (2014).
- B. J. Cardinale, H. Hillebrand, W. S. Harpole, K. Gross, R. Ptacnik, *Ecol. Lett.* **12**, 475–487 (2009).
- L. H. Fraser *et al.*, *Front. Ecol. Environ.* **11**, 147–155 (2013).
- Materials and methods are available as supplementary materials on Science Online.
- A. K. Knapp, T. R. Seastedt, *Bioscience* **36**, 662–668 (1986).
- B. L. Foster, K. L. Gross, *Ecology* **79**, 2593–2602 (1998).
- M. Oosterheld, S. J. McNaughton, in *Methods in Ecosystem Science*, O. E. Sala, R. B. Jackson, H. A. Mooney, R. Howarth, Eds. (Springer-Verlag, New York, 2000), chap. 2, pp. 151–157.
- T. K. Rajaniemi, *J. Ecol.* **90**, 316–324 (2002).
- F. Isbell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11911–11916 (2013).
- K. N. Suding *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 4387–4392 (2005).
- T. L. Dickson, K. L. Gross, *Oecologia* **173**, 1513–1520 (2013).
- P. Chesson, *Annu. Rev. Ecol. Syst.* **31**, 343–366 (2000).
- K. J. Gaston, *Nature* **405**, 220–227 (2000).
- G. Crowder, A. I. Csathó, L. Demeter, M. Demski, M. Deutschlandler, S. Donnelly, A. L. P. Dresseno, S. Enkhjin, O. Enkhmandal, T. Erdenebold, L. Erdenechimeg, B. Erdenetsseg, J. K. Fedrigo, A. C. Ferreira, Z. Foldvari, L. Fourie, B. Fraser, J. Galdi-Rosa, E. Gorgone-Barbosa, R. Greuel, A. Guido, É. György, D. Hall, A. Hassan, J. Hazi, R. Henkin, S. Hoffmann, T. Jairus, M. Jankju, Ü. Jögar, T. Jongbloets, M. Juhász, C. F. Jurinitz, V. R. Kakroudi, A. Kelemen, T. Khandarmaa, E. Khash-Erdene, C. Koch, C. Komoly, S. Kurukura, P. Liancourt, S. Lima, A. Lkhagva, M. Lucrecia Lipoma, D. Lkhagvasuren, J. Lombardi, M. Eugenia Marcotti, J. McPhee, B. McWhirter, L. Menezes, J. McCulloch, M. Mesdaghi, I. Máthé, M. Messini, M. Mistral, C. Moffat, M. Mohamed, L. Mugwedi, J. Padgham, P. Padilha, S. Paetz, S. Pagmadulam, G. Pec, C. Peconi, G. Péter, S. Piro, V. C. Pistóia, L. Pyle, M. Randall, M. Ninno Rissi, R. G. Rolim, M. Ross, T. Salarian, S. Sandagdorj, S. Sangasuren, C. Santinelli, C. Scherer, G. H. M. Silva, M. G. Silva, T. Smith, S. Solongo, F. Spada, R. Stahlmann, J. Steel, M. Sulyok, A. Sywenky, G. Szabó, L. Szabules, V. Tomlinson, J. Tremblay-Gravel, G. Ungvari, O. Urango, M. Uuganbayar, M. S. Viera, C. E. Vogel, D. Wallach, R. Wellstein, J. I. Withworth Hulse, and Z. Zimmermann. This work was supported in part by the Canada Research Chair Program, Canadian Foundation for Innovation (CFI), and a Natural Sciences and Engineering Research Council Discovery Grant (NSERC-DG) of Canada awarded to L.H.F., and Thompson Rivers University; a CFI and NSERC-DG awarded to J.P.; the University of Tartu, Estonia, and a European Regional Development Fund: Centre of Excellence FIBIR awarded to M.Z. and M.M.; a Hungarian National Science Foundation (OTKA K 105608) awarded to S.B.; Taylor Family-Asia Foundation Endowed Chair in Ecology and Conservation Biology and University of Mongolia's Support for High Impact Research program awarded to B.B.; the Rangeland Research Institute, University of Alberta, Canada; CONICET, Universidad Nacional de Córdoba, FONCYT, and the Inter-American Institute for Global Change Research (with support of NSF) awarded to S.D., L.E., and M.C.; a NSERC-DG awarded to J.C.; State Nature Reserve "Montagna di Torricchio" and University of Camerino, Italy; Hungarian University of Transylvania, Romania; a Fundação Grupo Boticário, Brazil (0153_2011_PR) awarded to A.F.; NSF DEB-1021158 and DEB-0950100 awarded to B.F.; UHURU: NSERC and CFI awarded to J.R.G. and the University of Wyoming; an NSERC-DG awarded to H.A.L.H.; an NSERC-DG awarded to J.K.; a National Natural Science Foundation of China grant (No. 41171417) awarded to R.L.; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (n. 307719/2012-0) awarded to S.M.; CNPq, Brazil (grants 403750/2012-1 and 307689/2014-0) awarded to V.P.; University of Florida and a NSF DEB 1149980 awarded to T.P.; Princeton Environmental Institute and a NSF DEB 1355122 awarded to R.M.P.; a CONYIT Becas-Chile Scholarship awarded to G.C.S. Data and R scripts associated with this paper are deposited in the Dryad repository (<http://datadryad.org/>).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/302/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S6
Tables S1 and S2
References (31–36)

29 April 2015; accepted 15 June 2015
10.1126/science.aab3916

ICE SHEETS

Reverse glacier motion during iceberg calving and the cause of glacial earthquakes

T. Murray,^{1*} M. Nettles,² N. Selmes,¹ L. M. Cathles,³ J. C. Burton,⁴ T. D. James,¹ S. Edwards,⁵ I. Martin,⁵ T. O'Farrell,⁶ R. Aspey,⁶ I. Rutt,¹ T. Baugé⁷

Nearly half of Greenland's mass loss occurs through iceberg calving, but the physical mechanisms operating during calving are poorly known and in situ observations are sparse. We show that calving at Greenland's Helheim Glacier causes a minutes-long reversal of the glacier's horizontal flow and a downward deflection of its terminus. The reverse motion results from the horizontal force caused by iceberg capsizing and acceleration away from the glacier front. The downward motion results from a hydrodynamic pressure drop behind the capsizing berg, which also causes an upward force on the solid Earth. These forces are the source of glacial earthquakes, globally detectable seismic events whose proper interpretation will allow remote sensing of calving processes occurring at increasing numbers of outlet glaciers in Greenland and Antarctica.

One-third to one-half of Greenland's total mass loss occurs through iceberg calving at the margins of tidewater-terminating glaciers (1, 2). Recent rapid changes in glacier dynamics are associated with increased calving rates (3–5) and increased rates of glacial earthquakes (6). At large glaciers with near-grounded termini, calving typically occurs when buoyancy forces cause icebergs that are the full thickness of the glacier to capsize against the calving front (6–9). This type of calving is associated with glacial earthquakes (6, 7, 10), long-period seismic emissions of magnitude ~5 that are observed globally (11). These earthquakes have expanded northward and increased sevenfold in number during

the past two decades (6, 12, 13), tracking changes in glacier dynamics, the retreat of glacier fronts, and increased mass loss (6, 14). Buoyancy-driven calving represents an increasingly important source of dynamic mass loss (6–8) as glacier fronts throughout Greenland have retreated to positions near their grounding lines (15). However, because of the difficulty of instrumenting the immediate near-terminus region of these highly active glaciers, few direct observations of the calving process are available, limiting development of the deterministic calving models required for improved understanding of controls on dynamic ice-mass loss. Detailed knowledge of the glacial earthquake source would allow quantification of calving processes for a large

ACKNOWLEDGMENTS

We are grateful to all of the people who helped in the collection and processing of the samples, including L. Alabiso-Cahill, D. Arunzaya, M. R. Ávila, J. C. R. Azambuja, L. Bachinger, I. Badamnyambuu, K. Baethke, J. Batbaatar, S. Ballelli, K. Bayarkhuu, G. Bertone, V. Besnyoi, C. L. Bonilha, G. Boorman, R. A. X. Borges, T. Broadbent, R. Canullo, J. Carding, B. Casper, K. Castillioni, M. Cervellini, G. Charles, G. Chiara, E. Cleland, R. Cornfoot,

class of Greenland glaciers, as well as for glaciers in several regions of Antarctica (13).

Agreement on the source mechanism of glacial earthquakes is limited. Analysis of long-period seismic data shows that a sub-horizontal force acts approximately perpendicular to the glacier calving front during the earthquakes (6, 13). The observed seismic signal is generated over a period of 1 min or more (6, 11, 16), much longer than the source duration for tectonic earthquakes of similar size (17). Some authors favor a model in which momentum transfer produces a force acting in the upglacier and then downglacier directions as a newly calved iceberg overturns, accelerates away from the calving front, and subsequently decelerates (6, 10, 13, 18). Others suggest that the seismic signal arises from the iceberg scraping along the calving front or fjord bottom (7) or colliding with the glacier terminus (19). Hydrodynamic interactions with fjord water may be important (20) but are little explored. Analytical investigations admit more than one possible mechanism for the earthquakes (18), and no persuasive explanation has been presented for the vertical component of the earthquake force. We combined geodetic, seismic, and laboratory data to identify the forces acting during calving at large glaciers and to document the source of the associated seismic signals.

We recorded geodetic data at the calving margin of Helheim Glacier (Fig. 1) (9), a major outlet of the Greenland Ice Sheet, during 55 days in July–September 2013. A wireless network of on-ice Global Positioning System (GPS) sensors (21) captured glacier motion with centimeter-level accuracy at a high temporal sampling rate in positions very close to the calving front (22). Hourly images from two cameras located ~4 km down-fjord from and looking at the calving front were used in stereo configuration to obtain the three-dimensional geometry of the calving front and calved icebergs (8, 22). Data from the global seismographic network were analyzed for the same time period to identify glacial earthquakes (13, 23) and obtain source parameters (11), including the orientation of the force active during the earthquake and the amplitude and centroid time, t_c , of a centroid-single-force (CSF) history of prescribed shape (22).

The glacier retreated ~1.5 km in a series of calving events during the observing period. We identified 10 large calving events from the camera images. All coincided with glacial earthquakes; in two cases, two earthquakes occurred between subsequent images. During the earthquakes, the region near the calving front showed a dramatic reversal of flow, moving upglacier for several minutes

while simultaneously moving downward (Fig. 2 and fig. S1). The horizontal and vertical motion then rebounded rapidly.

Observations from a glacial earthquake occurring on day of year (DOY) 206 at 03:13:47 UTC are shown in Fig. 2, A and C. Analysis of camera images indicates ice loss of $0.461 \pm 0.009 \text{ km}^2$ (Fig. 1) at a location of ice thickness 0.79 km, yielding an iceberg volume of 0.36 km^3 with an aspect ratio of 0.23. The earthquake had a CSF amplitude of $0.24 \times 10^{14} \text{ kg-m}$, with the force oriented 64°W (Fig. 1) and 9° above horizontal. GPS sensor 1 (Fig. 1) showed a pre-earthquake flow speed of 29 m/day. Immediately before the earthquake centroid time, the sensor reversed its direction and moved upglacier at ~40 m/day (displacement = 9 cm) and downward (displacement = 10 cm). The reversed motion was sustained for ~200 s and was followed by a downglacier rebound at ~190 m/day (displacement = 20 cm) and upward movement (displacement = 16 cm) for ~90 s. Similar temporally coincident signals were detected by nearby sensors 6 and 15 (Fig. 1 and fig. S1).

Glacier deflection for a calving event on DOY 212 (Fig. 1) is shown in Fig. 2, B and D. We observed similar responses for all glacial earthquake–iceberg calving events during which GPS sensors recording data of adequate quality were located within 500 m of the calved block (a total of nine glacial earthquakes and eight image pairs). These events occurred on DOY 205, 206 (three events), 207, 211, 212, and 226 and were detected by multiple GPS sensors (further examples in fig. S1).

The earthquake centroid times occurred at or near the end of the glacier's rapid rebound phase, such that the upglacier earthquake force aligned in time with the reverse motion of the glacier. The horizontal glacier deflection is consistent with a model in which the reaction force on the glacier caused by seaward acceleration of the newly calved iceberg compresses the glacier front elastically. The front then rebounds as the force decreases and reverses polarity during iceberg deceleration. The glacier front thus acts as a spring, compressing and re-extending in phase with the applied force, which is the horizontal component of the seismic source.

The downward deflection of the glacier front occurred in a region where vertical motion of the GPS sensors at tidal frequencies showed that the glacier is ungrounded and seawater is present beneath it. Iceberg rotation is likely to cause a low-pressure zone in the opening cavity between the iceberg and the glacier front. This pressure decrease would lower the load on the bedrock, resulting in an upward force acting on the solid Earth, as observed in our seismic analysis. A pressure decrease near the calving front would apply a net downward force on the glacier terminus, lowering the glacier surface in a manner similar to that occurring twice each day when the ocean tides draw down the water level. At sensors experiencing earthquake deflections, we observed tidal variations in the glacier's vertical position of ~0.1 m per 1 m of tidal amplitude. The calving-related deflection of the glacier surface was ~0.1 to 0.16 m, suggesting a change in water pressure

equivalent to a water-height change of ~1 to 1.6 m, or roughly $1 \text{ to } 2 \times 10^4 \text{ Pa}$.

No observations of pressure or water-level variations are available from the region in the fjord immediately in front of the glacier, where thick ice mélange (Fig. 1) prohibits instrumentation. However, results from analog laboratory experiments allowed us to evaluate our inferences (22). A model glacier “terminus” was secured at one end of a water-filled tank, and plastic “icebergs” made from low-density polyethylene were placed flush against the terminus and allowed to capsize spontaneously under the influence of gravitational and buoyancy forces (24) (Fig. 3). Sensors embedded in the model glacier terminus monitored pressure in the water column and the force exerted on the terminus during iceberg capsize.

The measured force on the terminus as the icebergs began to capsize was oriented in the upglacier direction and slowly increased as the icebergs rotated. As the icebergs neared horizontal, the force decreased rapidly. Pressure at the terminus decreased as the icebergs rotated, increasing again as the icebergs neared horizontal. Once the icebergs lost contact with the terminus, the measured force and pressure began to oscillate as a result of induced wave action in the tank.

We scaled up the measured forces and pressures to match the dimensions of icebergs calved at Helheim Glacier (Fig. 3). The laboratory data scale by powers of the ratio of the iceberg height in the field to the iceberg height in the laboratory (20, 24). The scaled peak force agreed well with typical values inferred from earthquake analysis ($\sim 10^{11} \text{ N}$). The scaled peak pressure drop ($\sim 5 \times 10^4 \text{ Pa}$) applied over an area corresponding to the iceberg's map-view dimensions yielded an upward-directed force consistent with the seismically inferred vertical force component, such that the total force acting on the solid Earth was oriented $\sim 10^\circ$ above horizontal. Computation and inversion of synthetic seismograms from the scaled force and pressure data confirmed the consistency of the laboratory model with real-world data.

We used the scaled force and pressure to predict the deformation of the terminus region (22). The total force (F_{tot}) per unit area (A_F) acting on the calving region produces a horizontal, linear deflection orthogonal to the calving front, such that $F_{\text{tot}}/A_F = E\Delta L/L$, where E is the Young's modulus of glacial ice. The value of L is chosen to provide the best match to the glacier position data. This length-scale probably represents the distance from the terminus to the grounding zone. We modeled the ungrounded section of the glacier as an elastic beam of length L loaded by the vertical force created by the pressure drop. The inferred distances L are a few kilometers, consistent with values estimated from GPS data.

Glacier displacements predicted from the scaled laboratory data for iceberg dimensions corresponding to a calving event on DOY 206 (Fig. 1 and Fig. 2A) are shown in Fig. 3. Agreement with the observed glacier displacement was very good, particularly during the time over which the force acted in the upglacier direction (until the earthquake centroid time). After this time, the

¹Glaciology Group, Department of Geography, College of Science, Swansea University, Swansea SA2 8PP, UK.

²Lamont-Doherty Earth Observatory, Columbia University, New York, NY 10964, USA. ³Department of Atmospheric, Oceanic and Space Sciences, University of Michigan, Ann Arbor, MI 48109, USA. ⁴Department of Physics, Emory University, Atlanta, GA 30322, USA. ⁵School of Civil Engineering and Geosciences, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. ⁶Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield S1 3JD, UK. ⁷Thales UK, Research and Technology, Worton Drive, Reading, Berkshire RG2 0SB, UK.

*Corresponding author. E-mail: t.murray@swansea.ac.uk

laboratory-derived prediction was dominated by oscillations of the water column in the tank, which did not contain the thick layer of ice mélange present in Helheim Fjord that would be expected to dampen such high-frequency oscillations.

We conclude that as large icebergs rotate and accelerate away from the glacier calving front (Fig. 4), the reaction force—which is the horizontal component of the earthquake force—compresses the glacier front elastically, overcoming normal

downglacier flow and temporarily reversing the motion of the glacier. Hydrodynamic interaction of the iceberg with the fjord water rapidly reduces pressure behind the rotating iceberg, resulting in an upward force on the solid Earth that is the

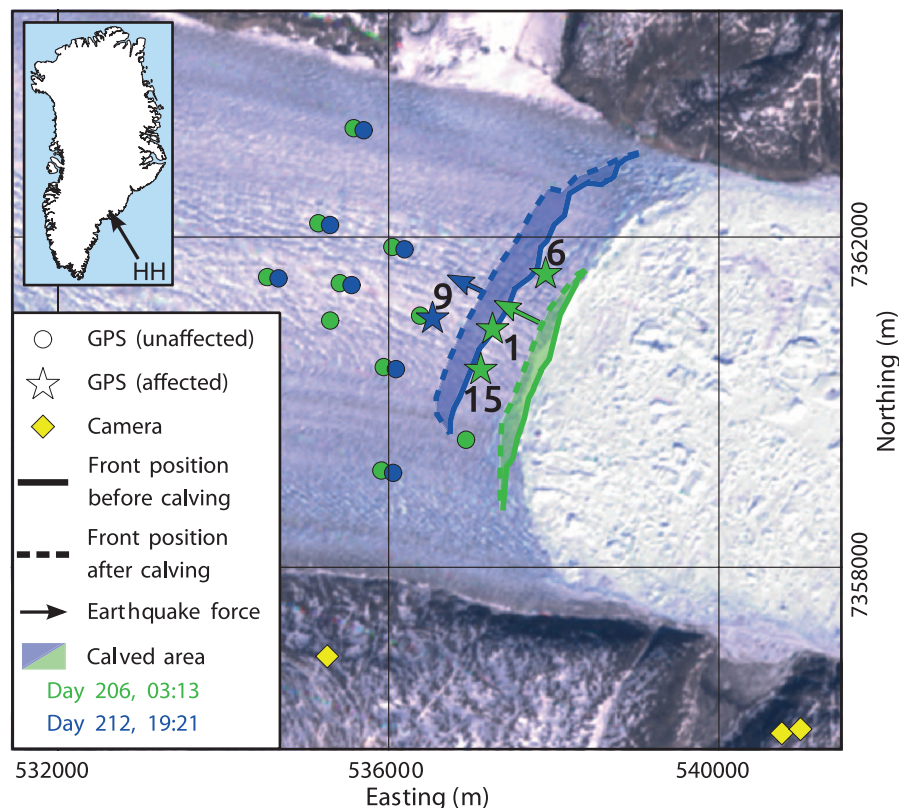


Fig. 1. Helheim Glacier, position of sensors, and seismic force directions. The location of GPS sensors and icebergs calved at Helheim Glacier (HH) for glacial earthquake events at 03:13 UTC on DOY 206 2013 and 19:21 UTC on DOY 212 2013 are superimposed on a Landsat 7 image from DOY 167 2013. “Affected” sensors exhibit earthquake-related deflections. Scan-line-corrector failure stripes have been removed for clarity. Glacier flow is from left to right; bright white mélange (a mix of iceberg fragments and sea ice) can be seen in front of the calving margin. Calving-front positions were obtained from photogrammetric digital elevation models derived from cameras. Positions are meters in Universal Transverse Mercator zone 24N.

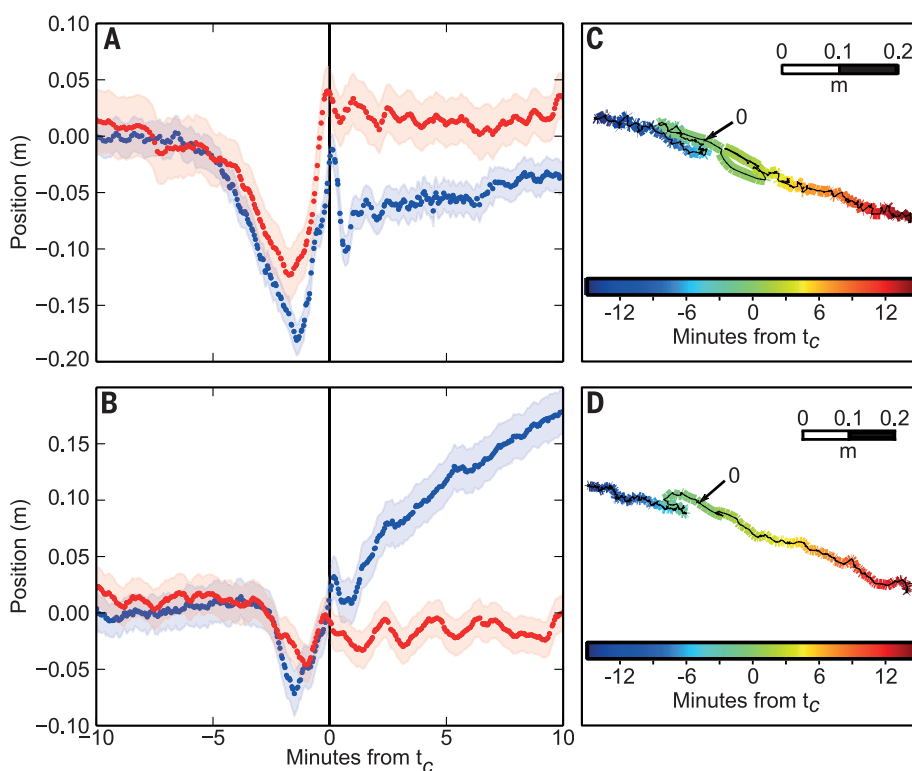


Fig. 2. Response of GPS sensors on glacier at the time of glacial earthquakes. (A) Sensor 1 at 03:13 UTC on DOY 206 2013. (B) Sensor 9 at 19:21 UTC on DOY 212 2013. Blue dots show detrended along-flow displacement; red dots show height. Shading shows 1σ position errors. Horizontal displacement has trends from 30 to 10 min before t_c removed (A = 28.9 m/day, B = 24.6 m/day). Height has mean removed. (C and D) Plan view of GPS traces shown in (A) and (B) during the 30 min surrounding t_c (marked as 0).

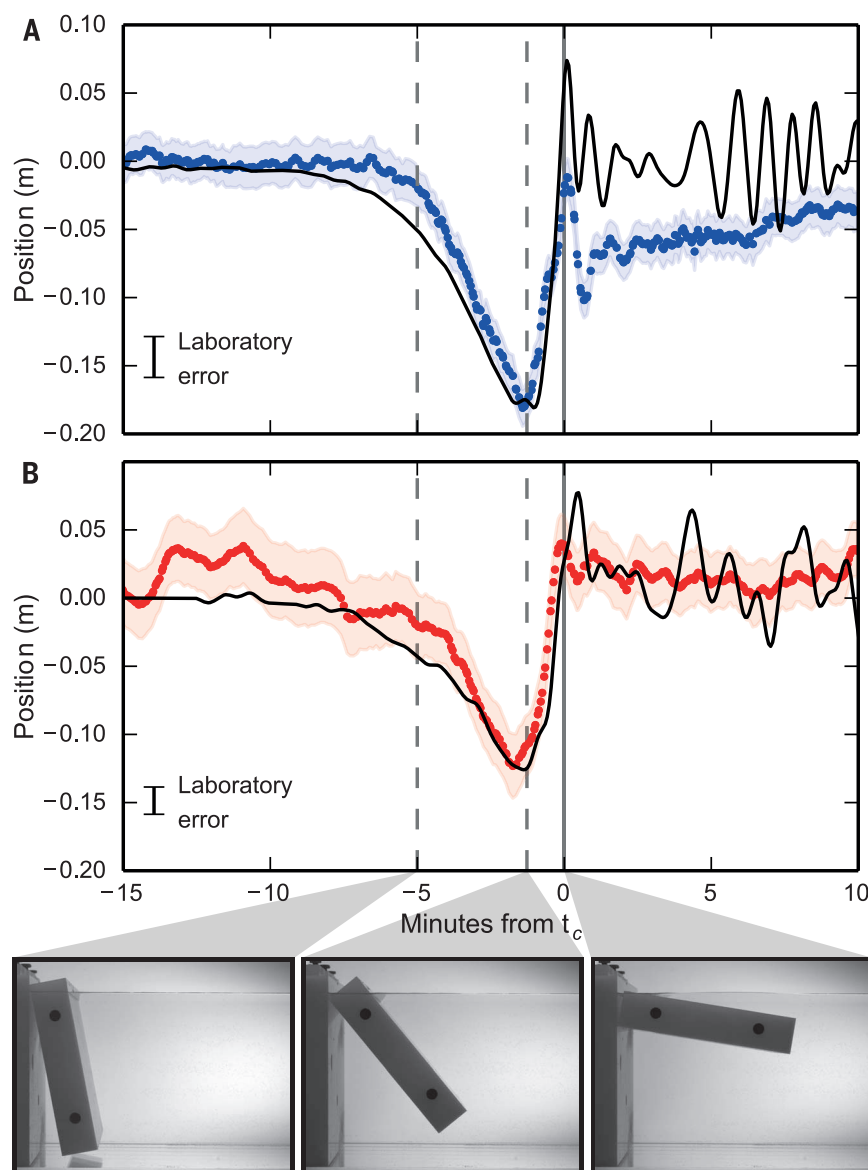
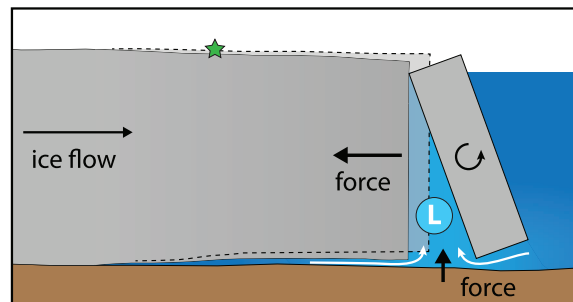


Fig. 3. Scaled laboratory data from glacier “terminus” during “iceberg” capsize event, compared with field observations. (A) Horizontal displacement scaled from force (black line) compared with downflow GPS data (blue). (B) Vertical displacement scaled from pressure (black line) compared with vertical GPS data (red). Errors in laboratory data are standard deviation calculated from repeated capsize events. GPS data shown are as in Fig. 2A. Photographs show stages of capsize at times marked by dashed lines and (solid gray line) t_c . The aspect ratio of the model iceberg is 0.22.

Fig. 4. Cartoon of glacier terminus during calving event.

Glacier deflection caused by a capsizing iceberg is shown relative to the initial glacier position (dotted line). Acceleration of the iceberg to the right exerts a force in the upglacier direction (left), leading to reverse motion of the GPS sensors (green star). Reduced pressure behind the iceberg (L) draws water from beneath the glacier and from the proglacial fjord, pulling the floating portion of the glacier downward and exerting an upward force on the solid Earth.



vertical force observed in the earthquake. The lowered water pressure draws down the ungrounded glacier margin, pulling the glacier surface downward during the earthquake.

Our results document the forces active during an increasingly important class of calving events and definitively identify the processes that cause glacial earthquakes. This understanding of glacier calving and glacial earthquakes opens the potential for remote quantitative characterization of iceberg calving and calving rates, as well as improved models for ice-ocean interaction.

REFERENCES AND NOTES

1. M. van den Broeke *et al.*, *Science* **326**, 984–986 (2009).
2. E. M. Enderlin *et al.*, *Geophys. Res. Lett.* **41**, 866–872 (2014).
3. I. Joughin, W. Abdalati, M. Fahnestock, *Nature* **432**, 608–610 (2004).
4. A. Luckman, T. Murray, R. de Lange, E. Hanna, *Geophys. Res. Lett.* **33**, L03503 (2006).
5. I. M. Howat, I. Joughin, T. A. Scambos, *Science* **315**, 1559–1561 (2007).
6. S. A. Veitch, M. Nettles, *J. Geophys. Res.* **117**, F04007 (2012).
7. J. M. Amundson *et al.*, *Geophys. Res. Lett.* **35**, L22501 (2008).
8. T. D. James, T. Murray, N. Selmes, K. Scharrer, M. E. O’Leary, *Nat. Geosci.* **7**, 593–596 (2014).
9. T. Murray *et al.*, *J. Geophys. Res. Earth Surf.* **10.1002/2015JF003531** (2015).
10. M. Nettles *et al.*, *Geophys. Res. Lett.* **35**, L24503 (2008).
11. G. Ekström, M. Nettles, G. A. Abers, *Science* **302**, 622–624 (2003).
12. G. Ekström, M. Nettles, V. C. Tsai, *Science* **311**, 1756–1758 (2006).
13. M. Nettles, G. Ekström, *Annu. Rev. Earth Planet. Sci.* **38**, 467–491 (2010).
14. I. Joughin *et al.*, *J. Geophys. Res.* **113**, F01004 (2008).
15. I. M. Howat, A. Eddy, *J. Glaciol.* **57**, 389–396 (2011).
16. V. C. Tsai, G. Ekström, *J. Geophys. Res.* **112**, F03S22 (2007).
17. G. Ekström, E. R. Engdahl, *J. Geophys. Res.* **94**, 15,499–15,519 (1989).
18. V. C. Tsai, J. R. Rice, M. Fahnestock, *J. Geophys. Res.* **113**, F03014 (2008).
19. F. Walter *et al.*, *J. Geophys. Res.* **117**, F01036 (2012).
20. J. M. Amundson, J. C. Burton, S. Correa-Legidos, *Ann. Glaciol.* **53**, 106–112 (2012).
21. I. Martin *et al.*, *IEEE Sens. J.* **14**, 3926–3931 (2014).
22. Materials and methods are available as supplementary materials on Science Online.
23. G. Ekström, *Bull. Seismol. Soc. Am.* **96**, 1201–1212 (2006).
24. J. C. Burton *et al.*, *J. Geophys. Res.* **117**, F01007 (2012).

ACKNOWLEDGMENTS

This work was supported by the Natural Environment Research Council UK grant NE/I007148/1. T.M. is currently supported by a Royal Society Leverhulme Trust Senior Research Fellowship. T.D.J. was supported by the Climate Change Consortium of Wales (C3W). M.N. was supported by U.S. NSF grant EAR-1249167. L.M.C. is currently supported by the Michigan Society of Fellows. J.B. and L.M.C. were supported and the laboratory equipment was developed with support from NSF grant ANT-0944193. A. Everett is thanked for assistance in the field and L. Kaluzienski for assistance with laboratory data. We thank the staff of the Civil Engineering and Geosciences workshop, Newcastle University, for GPS sensor construction. We acknowledge the use of bed data from the Center for Remote Sensing of Ice Sheets, generated with support from NSF grant ANT-0424589 and NASA grant NNX10AT68G, and the use of seismic data from the Incorporated Research Institutions for Seismology (IRIS)—U.S. Geological Survey Global Seismographic Network, Geoscope, Geofon, Mednet, and the Greenland Ice Sheet Monitoring Network. A 2013 lidar survey flown by the Natural Environment Research Council Airborne Remote Sensing Facility was used in the processing of photographs. Seismic waveforms are available from the IRIS Data Management Center (NSF EAR-1261681); GPS data are available from UNAVCO (NSF EAR-1261833).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/305/suppl/DC1

Materials and Methods

Fig. S1

References (25–41)

Data Table S1

10 March 2015; accepted 12 June 2015

Published online 25 June 2015

10.1126/science.aab0460

PLANT SCIENCE

Morphinan biosynthesis in opium poppy requires a P450-oxidoreductase fusion protein

Thilo Winzer,¹ Marcelo Kern,¹ Andrew J. King,¹ Tony R. Larson,¹ Roxana I. Teodor,¹ Samantha L. Donninger,¹ Yi Li,¹ Adam A. Dowle,² Jared Cartwright,² Rachel Bates,² David Ashford,² Jerry Thomas,² Carol Walker,³ Tim A. Bowser,³ Ian A. Graham^{1*}

Morphinan alkaloids from the opium poppy are used for pain relief. The direction of metabolites to morphinan biosynthesis requires isomerization of (S)- to (R)-reticuline. Characterization of high-reticuline poppy mutants revealed a genetic locus, designated *STORR* [(S)- to (R)-reticuline] that encodes both cytochrome P450 and oxidoreductase modules, the latter belonging to the aldo-keto reductase family. Metabolite analysis of mutant alleles and heterologous expression demonstrate that the P450 module is responsible for the conversion of (S)-reticuline to 1,2-dehydroreticuline, whereas the oxidoreductase module converts 1,2-dehydroreticuline to (R)-reticuline rather than functioning as a P450 redox partner. Proteomic analysis confirmed that these two modules are contained on a single polypeptide in vivo. This modular assembly implies a selection pressure favoring substrate channeling. The fusion protein *STORR* may enable microbial-based morphinan production.

The naturally occurring opiates of the morphinan subclass of benzylisoquinoline alkaloids (BIAs) include morphine, codeine, and thebaine. Morphine and codeine can be directly used as analgesic painkillers, and thebaine is widely used as a feedstock for the synthesis of a number of semisynthetic opiates, including hydrocodone, hydromorphone, oxycodone, and oxymorphone, as well as the opioid antagonist naloxone. The discovery and isolation of morphine from the opium poppy (*Papaver somniferum* L.) by Friedrich Sertürner in 1806 (1) are a milestone in the history of pharmacy. More than 200 years later, opiate alkaloid-based pharmaceutical formulations remain the most potent treatment for severe pain, with sales totaling \$US1.6 billion in 2013 (2).

BIAs are found in a number of species in the Papaveraceae family, but morphine production has only been reported in the opium poppy and the closely related *P. setigerum* (3). The morphinan backbone contains five asymmetric carbon centers. Total chemical synthesis, although possible (4), is not an economically viable means of production. Consequently, morphinan alkaloids are still exclusively sourced from the opium poppy plant. Much effort has gone into the elucidation of the morphinan branch of BIA metabolism over the past 25 years, resulting in the identi-

fication of genes for all but the gateway step involving the epimerization of (S)- to (R)-reticuline (3, 5–14). Thus, although it has been possible to produce both (S)- and racemic mixtures of (R,S)-reticuline and morphinans in metabolically engineered microbial systems (15–18), the clean enzymatic conversion of (S)- to (R)-reticuline remains the goal.

(S)-reticuline is the central intermediate of BIA metabolism (fig. S1), and conversion to its R epimer is believed to be a two-step process (19, 20). The S epimer is first oxidized to the quaternary positively charged amine 1,2-dehydroreticuline, followed by reduction to (R)-reticuline (Fig. 1A). Activities for each step have been reported, but the identity of the corresponding proteins has not been established (19, 20). We have combined a candidate gene approach with genetic analyses of F₂ populations of *P. somniferum* segregating for mutations that are deficient in (S)- to (R)-reticuline conversion and discovered that a fusion protein is responsible for sequentially catalyzing both steps of the epimerization.

RNA interference (RNAi) knockdown of codeinone reductase in the opium poppy was reported to cause accumulation of (S)-reticuline, which is eight steps upstream of the codeinone reductase substrate (21), a result the authors attributed to metabolite channeling. We considered an alternative hypothesis to be the off-target co-silencing of a closely related oxidoreductase involved in the conversion of (S)- to (R)-reticuline. Using the sequence of the RNAi silencing construct to query an in-house expressed sequence tag (EST) library from stem and capsule tissue of opium poppies (22), we identified a contiguous assembly comprising a cytochrome P450 mono-

oxygenase that is 3'-linked to an oxidoreductase. Sequencing cDNA clones from opium poppy stems confirmed the in-frame fusion transcript (Fig. 1A and fig. S2). The P450 module was designated CYP82Y2.

To investigate whether the corresponding gene is a candidate for one or both steps in the epimerization of (S)- to (R)-reticuline, we sequenced corresponding cDNA clones from three independent mutants identified from an ethyl methanesulfonate-mutagenized population of a high-morphine cultivar, HM2 (23). All three mutants have lost the ability to produce morphinan alkaloids and instead accumulate high levels of (S)-reticuline as well as the (S)-reticuline-derived alkaloids laudanone and laudanone (Fig. 1, B and D). We found that all three mutant lines carry mutations in the corresponding gene locus (Fig. 1A), which we name *STORR* [(S)- to (R)-reticuline]. The *storr-1* allele carries a premature stop codon corresponding to amino acid position W668 in the oxidoreductase module of the predicted fusion protein. *storr-1* plants also contain low but significant levels of 1,2-dehydroreticuline (Fig. 1C), which suggests that the oxidoreductase module catalyzes the second step of the epimerization, the reduction of 1,2-dehydroreticuline to (R)-reticuline. *storr-2* and *storr-3* are both disrupted in the CYP82Y2 module: *storr-2* contains a premature stop at codon position W278, and *storr-3* contains a missense mutation causing a glycine-to-arginine substitution at position 550 (Fig. 1A). Dried capsules of *storr-2* and *storr-3* accumulate (S)-reticuline but not 1,2-dehydroreticuline, suggesting that the CYP82Y2 module is responsible for the first epimerization step, the oxidation of (S)-reticuline to 1,2-dehydroreticuline. Complementation tests and F₂ segregation analysis confirmed that a single genetic locus is responsible for the high-reticuline phenotype, association of the three recessive *storr* alleles with the high-reticuline phenotype, and the consecutive roles of the CYP82Y2 and oxidoreductase modules in the epimerization of (S)- to (R)-reticuline (tables S2 and S3).

To establish whether the *STORR* locus is not only transcribed but also translated as a fusion protein, we used a quantitative mass spectrometry approach after gel fractionation of crude protein extract from HM2. Peptides from across the entire *STORR* protein were found to be most abundant in the gel regions covering the 100.65-kD predicted size of the fusion protein, confirming this as the in vivo form (Fig. 2). For direct functional characterization, the *STORR* fusion protein and the separate modules were expressed in *Saccharomyces cerevisiae*, and enzyme assays were performed on soluble extracts and microsomal preparations (Fig. 3). We found that 1,2-dehydroreticuline is converted to (R)-reticuline with 100% conversion efficiency by both the *STORR* fusion protein and the oxidoreductase module, but not by the CYP82Y2 module plus its redox partner (Fig. 3A). In contrast, the CYP82Y2 module plus its redox partner catalyzed 97% conversion of (S)-reticuline to 1,2-dehydroreticuline,

¹Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5DD, UK. ²Bioscience Technology Facility, Department of Biology, University of York, York YO10 5DD, UK. ³GlaxoSmithKline, 1061 Mountain Highway, Post Office Box 168, Boronia, Victoria 3155, Australia.

*Corresponding author. E-mail: ian.graham@york.ac.uk

Fig. 1. Characterization of opium poppy mutants disrupted in the conversion of (S)- to (R)-reticuline. (A) Schematic showing epimerization of (S)- to (R)-reticuline and position of the *storr-1*, *storr-2*, and *storr-3* mutations in the predicted fusion protein. (B) Mean ± SD capsule reticuline content in the HM2 wild-type cultivar and *storr* mutants (HM2, *n* = 5; *storr-1*, *n* = 12; *storr-2*, *n* = 17; *storr-3*, *n* = 15). DW, dry weight. Reticuline content was verified as >99.2% (S)-reticuline in all mutants by chiral high-performance liquid chromatography (HPLC) (table S1). (C) 1,2-dehydroreticuline. (D) All compounds >1% total alkaloids (*n* = 10) are individually identified, with minor peaks (*n* = 379), grouped as “Other.”

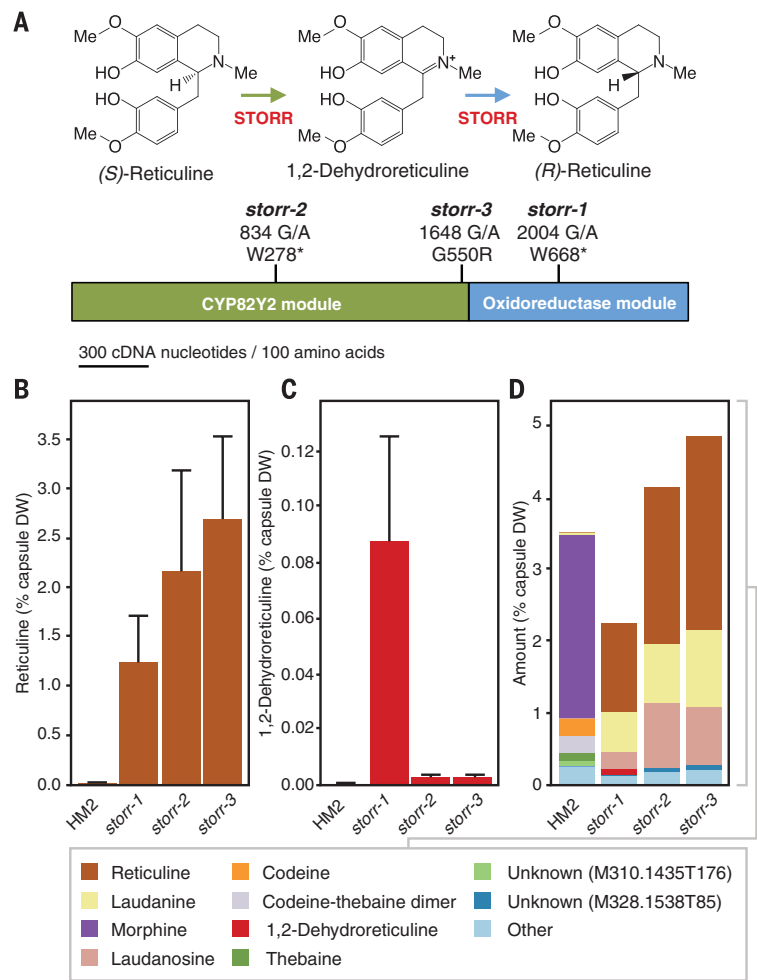
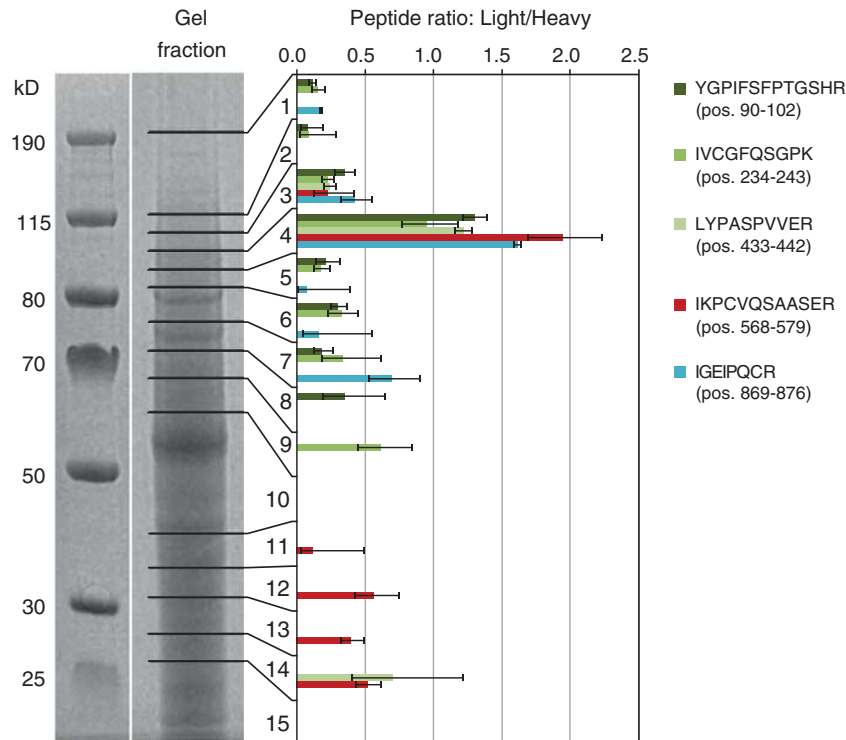


Fig. 2. Size determination of STORR protein in the opium poppy. Protein extracts from three stem samples of HM2 wild type were fractionated, together with size markers, by SDS–polyacrylamide gel electrophoresis (SDS–PAGE), and the three lanes were each cut into 15 fractions (see horizontal lines on one representative lane) to resolve the predicted fusion protein and putative individual CYP450 and oxidoreductase modules (101, 65, and 36 kD, respectively). For relative quantification, an equal amount of a tryptic digest of ¹⁵N-labeled recombinant STORR protein was spiked into the in-gel digest of each SDS–PAGE fraction before HPLC mass spectrometry (HPLC–MS). Ratios of peak areas from extracted-ion chromatograms of light (endogenous) to heavy (labeled) versions of five peptides from across the STORR protein sequence were compared. Ratios of normalized peak areas from extracted-ion chromatograms were converted to binary logarithms for the calculation of means and standard errors of the mean. Only measurements where the respective peptides were found in all three biological replicates are shown.



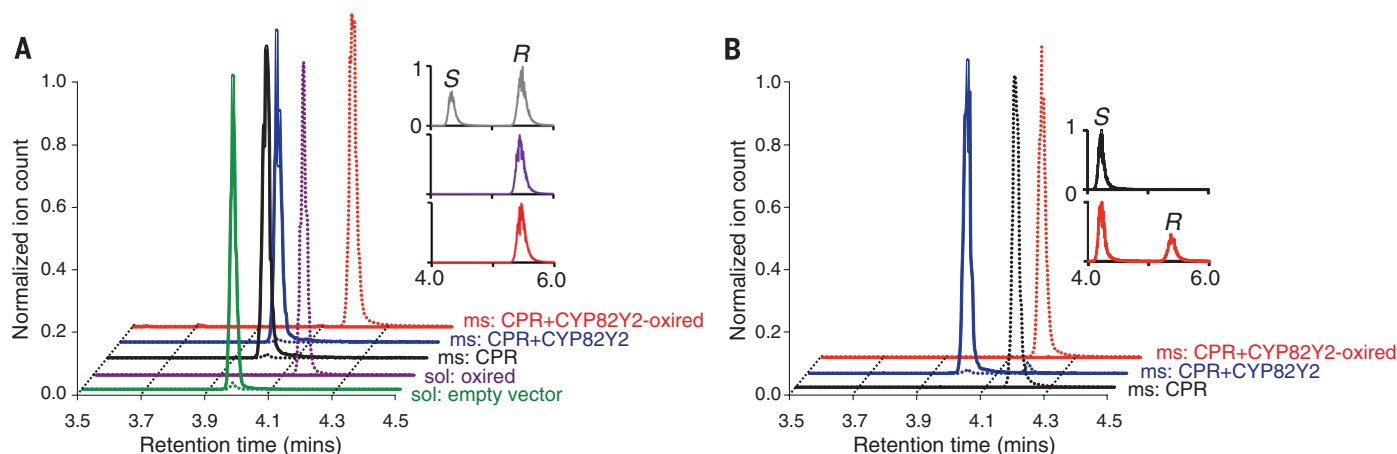


Fig. 3. Functional characterization of the STORR fusion protein by heterologous expression in *S. cerevisiae*. (A) HPLC-MS analysis of the *in vitro* conversion of 1,2-dehydroreticuline to (*R*)-reticuline. Crude soluble (sol) or microsomal (ms) preparations harboring the empty pESC-TRP vector (green), vector containing the oxidoreductase module (purple), an opium poppy cytochrome P450 reductase (CPR) redox partner (black), CPR + CYP82Y2 (blue), or CPR + the CYP82Y2-oxidoreductase fusion (red) were assayed (21). The solid lines of the HPLC-MS chromatograms show the normalized total ion count at a mass/charge ratio (*m/z*) of 328, corresponding to 1,2-dehydroreticuline (substrate), whereas the dotted lines show the normalized total ion count at *m/z* 330, corresponding to reticuline (product). The inset panel shows the chiral analysis of reticuline: The gray

trace is for an (*S*)- and (*R*)-reticuline standard, and the purple and red traces correspond to reticuline derived by activity of the oxidoreductase and CYP82Y2-oxidoreductase fusion, respectively. (B) HPLC-MS analysis of the conversion of (*S*)-reticuline into 1,2-dehydroreticuline and (*R*)-reticuline. Crude microsomal preparations obtained from *S. cerevisiae* harboring expression vector pESC-TRP containing CPR only (black), CPR + a CYP82Y2 module (blue), or CPR + CYP82Y2-oxidoreductase fusion (red) were assayed (21). The solid lines of the HPLC-MS chromatograms show the normalized total ion count at *m/z* 328, corresponding to 1,2-dehydroreticuline, and the dotted lines show the normalized total ion count at *m/z* 330, corresponding to reticuline. The inset panel shows the chiral analysis of reticuline, with the same line colors as in the main panel.

demonstrating that it acts as a 1,2-dehydroreticuline synthase (Fig. 3B). Microsomal preparations harboring the entire STORR fusion protein converted about 20% of the added (*S*)-reticuline to (*R*)-reticuline, confirming the bifunctional role of the protein in performing sequential reactions in the epimerization of reticuline. Kinetic analysis revealed that the microsomal CYP82Y2 module alone and the CYP82Y2-oxidoreductase fusion had similar Michaelis constant values of 13 and 14 μ M for (*S*)-reticuline and 1,2-dehydroreticuline, respectively (fig. S3). Consistent with the plant mutant phenotypes, we found that microsomally expressed STORR carrying the *storr-2* mutation lacks both the P450 and oxidoreductase activities, the *storr-3* mutation lacks the P450 activity but does still exhibit the oxidoreductase activity, and the *storr-1* mutation has lost the oxidoreductase activity but maintains very low levels of P450 activity (fig. S4).

P450-redox systems where the P450 enzyme is covalently linked to redox partner reductase components are well known in both prokaryotes and lower eukaryotes (24). In the STORR fusion protein, the P450 module is linked to a reductase, but rather than functioning as a redox partner for the P450, this reductase catalyzes the product of the P450 to complete a two-step epimerization of (*S*)- to (*R*)-reticuline. Other forms of bifunctional P450 fusions with oxygenase/peroxidase, hydrolase, and dioxygenase modules have been reported to occur in ascomycetes, and all of these also appear to catalyze sequential

reactions (25–27). A possible explanation as to why such fusion proteins evolve is that they facilitate efficient channeling of highly unstable or reactive intermediates. Evidence for efficient substrate channeling in the case of the STORR fusion protein comes from the observation that microsomal fractions harboring the fusion protein directly convert (*S*)- to (*R*)-reticuline, with no detectable accumulation of 1,2-dehydroreticuline (Fig. 3B).

Phylogenetic analysis suggests that the STORR fusion occurred after the split of the CYP82X and the CYP82Y subfamilies and that the oxidoreductase module falls into subfamily 4 of the aldo-keto reductases, with closest homology to the codeinone reductase family from *P. somniferum* (fig. S5). A query of the 1K plant transcriptome resource (28) identified similar predicted module arrangements in EST collections from two other morphinan-producing *Papaver* species, *P. bracteatum* and *P. setigerum*, but not in *P. rhoeas*, which does not make morphinans (table S4). We hypothesize that the STORR fusion was the key step in the evolution of the morphinan branch of BIA metabolism. After this step, other enzymes were recruited and adapted, including dioxygenases and reductases, ultimately giving rise to codeine and morphine in *P. somniferum* and *P. setigerum* (29). Thus, this morphinan biosynthetic pathway, and probably other plant secondary metabolic pathways, depends on the organization of both individual gene structure and genome rearrangement (22).

REFERENCES AND NOTES

1. F. Sertürner, *J. Pharmacie* **14**, 47–93 (1806).
2. IMS Health Database, Formulation sales by opiate molecule, www.imshealth.com (2013).
3. J. Ziegler et al., *Plant J.* **48**, 177–192 (2006).
4. M. Gates, G. Tschudi, *J. Am. Chem. Soc.* **78**, 1380–1393 (1956).
5. N. Samanani, D. K. Liscombe, P. J. Facchini, *Plant J.* **40**, 302–313 (2004).
6. A. Ounaro, G. Decker, J. Schmidt, F. Lottspeich, T. M. Kutchan, *Plant J.* **36**, 808–819 (2003).
7. K. B. Choi, T. Morishige, F. Sato, *Phytochemistry* **56**, 649–655 (2001).
8. H. H. Pauli, T. M. Kutchan, *Plant J.* **13**, 793–801 (1998).
9. T. Morishige, T. Tsujita, Y. Yamada, F. Sato, *J. Biol. Chem.* **275**, 23398–23405 (2000).
10. A. Gesell et al., *J. Biol. Chem.* **284**, 24432–24442 (2009).
11. R. Lenz, M. H. Zenk, *J. Biol. Chem.* **270**, 31091–31096 (1995).
12. T. Grothe, R. Lenz, T. M. Kutchan, *J. Biol. Chem.* **276**, 30717–30723 (2001).
13. J. M. Hagel, P. J. Facchini, *Nat. Chem. Biol.* **6**, 273–275 (2010).
14. B. Unterlinner, R. Lenz, T. M. Kutchan, *Plant J.* **18**, 465–475 (1999).
15. A. Nakagawa et al., *Nat. Commun.* **2**, 326 (2011).
16. A. Nakagawa et al., *Sci. Rep.* **4**, 6695 (2014).
17. K. M. Hawkins, C. D. Smolke, *Nat. Chem. Biol.* **4**, 564–573 (2008).
18. K. Thodey, S. Galanie, C. D. Smolke, *Nat. Chem. Biol.* **10**, 837–844 (2014).
19. W. De-Eknamkul, M. H. Zenk, *Phytochemistry* **31**, 813–821 (1992).
20. K. Hirata, C. Poeaknapo, J. Schmidt, M. H. Zenk, *Phytochemistry* **65**, 1039–1046 (2004).
21. R. S. Allen et al., *Nat. Biotechnol.* **22**, 1559–1566 (2004).
22. T. Winzer et al., *Science* **336**, 1704–1708 (2012).
23. Materials and methods are available as supplementary materials on Science Online.
24. F. P. Guengerich, A. W. Munro, *J. Biol. Chem.* **288**, 17065–17073 (2013).

25. F. Brodhun, C. Göbel, E. Hornung, I. Feussner, *J. Biol. Chem.* **284**, 11792–11805 (2009).
 26. B. G. Hansen *et al.*, *Appl. Environ. Microbiol.* **78**, 4908–4913 (2012).
 27. I. Hoffmann, F. Jernerén, E. H. Olm, *J. Lipid Res.* **55**, 2113–2123 (2014).
 28. The thousand plant transcriptomes project, www.onekp.com.
 29. S. C. Farrow, P. J. Facchini, *J. Biol. Chem.* **288**, 28997–29012 (2013).

ACKNOWLEDGMENTS

We thank the laboratory and horticultural staff at GlaxoSmithKline Australia and the University of York for valuable technical assistance;

J. Mitchell for administrative support; T. Davis for providing high-(S)-reticuline poppy straw; A. Bridges for providing DNA constructs; and D. Nelson for naming the P450 module. We acknowledge financial support from the UK Biotechnology and Biological Sciences Research Council (grant BB/K018809/1) and The Garfield Weston Foundation. The STORR cDNA sequence is available through the National Centre for Biotechnology Information under GenBank accession number KP998574. The University of York and GlaxoSmithKline have filed a patent application relating to this work. The supplementary materials contain additional data.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/309/suppl/DC1
 Materials and Methods
 Figs. S1 to S9
 Tables S1 to S4
 References (30–42)

25 March 2015; accepted 18 June 2015
 Published online 25 June 2015;
 10.1126/science.aab1852

CIRCADIAN RHYTHMS

Atomic-scale origins of slowness in the cyanobacterial circadian clock

Jun Abe,^{1*} Takuya B. Hiyama,^{1*} Atsushi Mukaiyama,^{1,2*} Seyoung Son,^{3*†} Toshifumi Mori,^{2,4*} Shinji Saito,^{1,2,4} Masato Osako,³ Julie Wolanin,^{1,5} Eiki Yamashita,⁶ Takao Kondo,³ Shuji Akiyama^{1,2,†}

Circadian clocks generate slow and ordered cellular dynamics but consist of fast-moving bio-macromolecules; consequently, the origins of the overall slowness remain unclear. We identified the adenosine triphosphate (ATP) catalytic region [adenosine triphosphatase (ATPase)] in the amino-terminal half of the clock protein KaiC as the minimal pacemaker that controls the in vivo frequency of the cyanobacterial clock. Crystal structures of the ATPase revealed that the slowness of this ATPase arises from sequestration of a lytic water molecule in an unfavorable position and coupling of ATP hydrolysis to a peptide isomerization with high activation energy. The slow ATPase is coupled with another ATPase catalyzing autodephosphorylation in the carboxyl-terminal half of KaiC, yielding the circadian response frequency of intermolecular interactions with other clock-related proteins that influences the transcription and translation cycle.

Circadian clocks comprise suites of biological processes that oscillate with a 24-hour period (1). Clock genes and clock proteins are present in prokaryotes and eukaryotes (2, 3); together, they constitute feedback loops that effect transcriptional and translational oscillations (TTOs). The origin of the slow circadian time scale is thought to be the time delay between clock gene transcription and feedback signals that regulate it; however, the transcriptional and translational events can occur quickly (i.e., within minutes) (4). Posttranslational oscillations (PTOs) (5–7) in biochemical modifications of clock proteins occur even without transcriptional and translational regulation. Proteins generally exhibit dynamics within pico-

seconds or seconds (8), much faster than the circadian time scale. Thus, both TTO and PTO circadian systems are assembled from building blocks with intrinsically fast dynamics, raising questions about how and why the systems are so slow and stable overall (9).

The cyanobacterium *Synechococcus elongatus* PCC7942 is the simplest organism known to have both TTOs (10) and PTOs (5). The *S. elongatus* PTOs can be reconstructed in vitro by incubating a core clock protein, KaiC, with two other clock proteins, KaiA and KaiB, and adenosine triphosphate (ATP) (6). The rhythmic behaviors of the Kai oscillator have been confirmed in many functional and structural analyses, which have probed the ATP hydrolysis [adenosine triphosphatase (ATPase)] activity of KaiC (11, 12), autophosphorylation and autodephosphorylation activities of KaiC (6, 13, 14), conformational transitions of the proteins (12, 15, 16), and assembly or disassembly of Kai complexes (17–20). Because the PTO period in *S. elongatus* is firmly correlated to the TTO period during the day (5, 6), the in vitro Kai oscillator should enable us to identify the mechanisms underlying the slowness of the circadian clock.

We searched the Kai oscillator for a minimal slow reaction whose efficiency correlated with in vivo TTO frequency (Fig. 1) (see supplementary materials and methods). The ATPase activity of full-length wild-type KaiC (KaiC-WT), con-

sisting of the N-terminal C1 and C-terminal C2 domains, has been proposed as the basic timing cue (11, 12). We identified the steady-state ATPase activity of the C1 domain (C1-ATPase) as a suitable slow reaction. A truncated version of KaiC consisting solely of the N-terminal domain, KaiC1-WT, exhibited much slower ATP hydrolysis (11 ± 1 ATP per day per KaiC at 30°C) than well-known motor proteins (10^3 to 10^7 day⁻¹) such as myosin, kinesin, and F₁-ATPase (table S1). We confirmed the correlation of this activity with the in vivo TTO frequency using a series of period-modulating KaiC mutations in the C1 domain [S¹⁵⁷→P¹⁵⁷ (S157P), S157C, T42S, S48T] (21) (Fig. 1D), in which higher steady-state C1-ATPase in vitro resulted in higher-frequency TTOs in vivo. Thus, KaiC should experience a certain number of hydrolysis events per cycle in vivo, and the absolute rate of ATPase activity is connected to the cellular clock's overall slowness (11, 22). Therefore, we examined the structural origin of the slow C1-ATPase and its coupling with TTOs via spatiotemporally distinct events, including the intramolecular KaiC ATPase and its phosphorylation cycles, as well as intermolecular interactions with KaiA and KaiB (Fig. 1).

To this end, we crystallized KaiC1-WT and five period-modulating variants in the pre- or posthydrolysis states, or both. All resultant crystals were in the P2₁2₁2₁ space group (Fig. 2, A and B, fig. S1, and table S2). The prehydrolysis states (Fig. 2A, blue subunits) exhibited common features: assembly of six subunits into a hexamer and incorporation of one molecule of the slowly hydrolyzed ATP analog adenosine 5'-(γ -thiotriphosphate) (ATP- γ -S) into every subunit-subunit interface. ATP- γ -S existed in a complex with a Mg²⁺ ion (Mg-ATP- γ -S) in the ordinary octahedral-coordination geometry (fig. S2A). We observed the posthydrolysis state of the long-period variant KaiC1-S48T (Fig. 2B, orange and green subunits), which crystallized as an asymmetrically ATP- or adenosine diphosphate (ADP)-bound hexamer (fig. S2B). The ring-shaped hexamer was deformed asymmetrically due to steric constraints resulting from close juxtaposition of three types of subunits, creating both tight and loose intersubunit interfaces (Fig. 2B, fig. S3, and supplementary text).

We identified two structural sources of slow ATPase activity. The first is the regulatory influence of the protein moiety on a lytic water molecule (W1) near the phosphorus atom (P _{γ}) of the γ -phosphate group of an ATP. In the prehydrolysis state (blue dotted box in Fig. 2C), W1 was sequestered (with a W1-P _{γ} distance of 3.8 to 3.9 Å and a W1-P _{γ} -O_{3 β} angle of 154° to 158°) by

¹Research Center of Integrative Molecular Systems (CIMoS), Institute for Molecular Science, 38 Nishigo-Naka, Myodaiji, Okazaki 444-8585, Japan. ²Department of Functional Molecular Science, SOKENDAI (The Graduate University for Advanced Studies), 38 Nishigo-Naka, Myodaiji, Okazaki 444-8585, Japan. ³Division of Biological Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan. ⁴Department of Theoretical and Computational Molecular Science, Institute for Molecular Science, 38 Nishigo-Naka, Myodaiji, Okazaki 444-8585, Japan. ⁵PSL Research University, Chimie ParisTech, 75005 Paris, France. ⁶Institute for Protein Research, Osaka University, 3-2 Yamada-oka, Suita 565-0871, Japan.
 *These authors contributed equally to this work. †Present address: College of Pharmacy, Chungbuk National University, 410 Seongbong-ro, Heungdeok-gu, Cheongju, Chungbuk 361-763, Korea. ‡Corresponding author. E-mail: akiyamas@ims.ajp

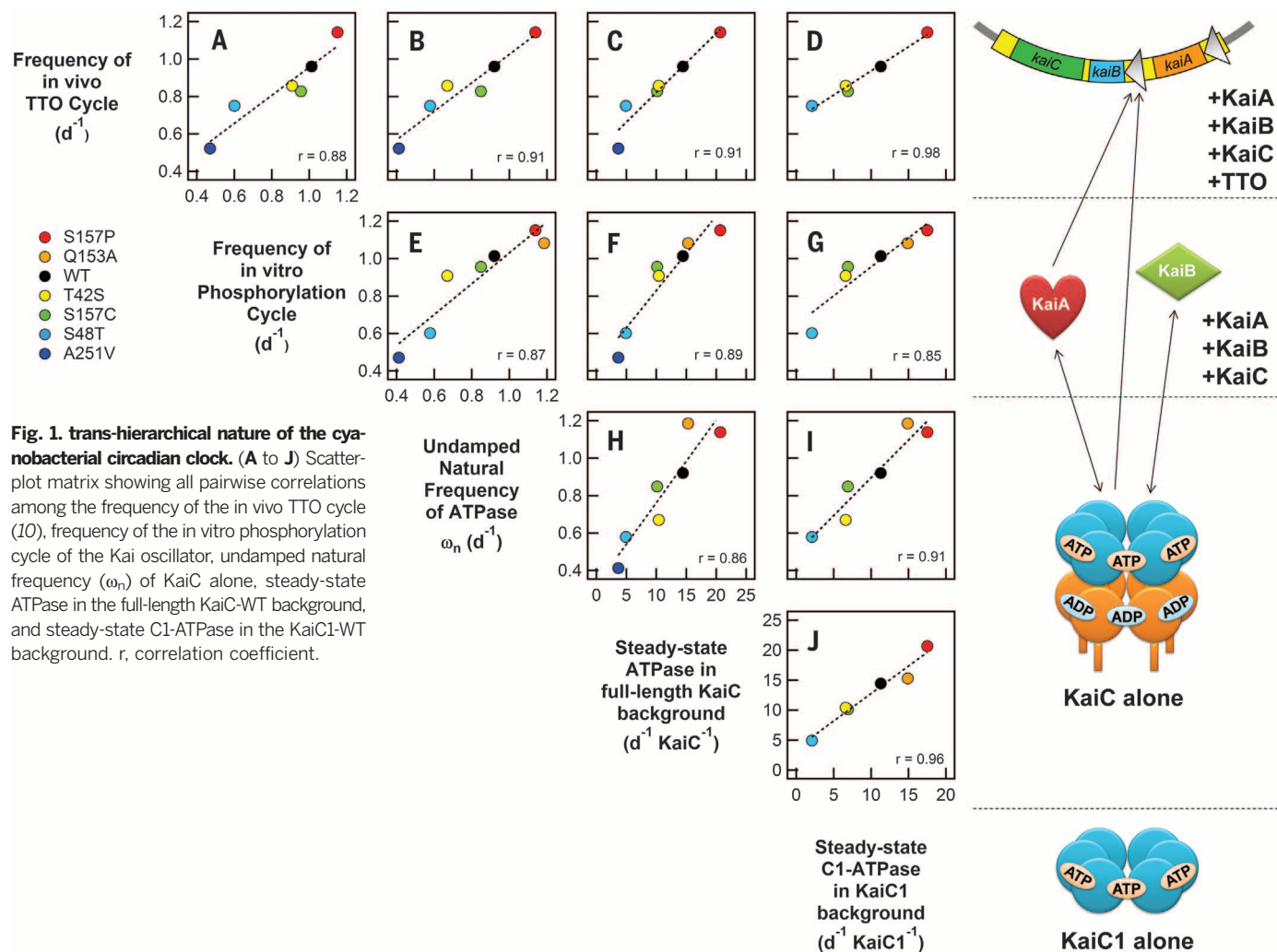
H-bonding to the F199 carbonyl oxygen, the nitrogen atom of R226 side chain (N_n), and another water molecule (W2). This position of W1 was much farther from a near-in-line configuration, with respect to the P_γ - $O_{3\beta}$ bond (3.0 \AA , 180°), than analogous water molecules in other motor proteins that hydrolyze ATP efficiently—e.g., kinesin (3.3 \AA , 164° to 167°) (23), myosin (2.3 \AA , 170°) (24), and F_1 -ATPase (3.1 to 3.8 \AA , 168° to 172°) (25) (Fig. 2D and table S1). Molecular dynamics simulation suggested that W1 was only marginally stabilized. Although a water molecule might occasionally migrate into and out of the W1 position, the near-in-line position was completely inaccessible due to steric hindrance from the F199 carbonyl oxygen and the R226 N_n (Fig. 2E and movies S1 to S3). Thus, the sequestration of W1 from the near-in-line position is one atomic-scale origin of slow hydrolysis. In another crystal form of KaiC1-WT ($P3_121$ space group; other subunits in Fig. 2F and fig. S1A), W1 was even farther from the P_γ of the γ -phosphate group, which should be even more unfavorable to hydrolysis than the W1 position that we observed in the $P2_12_1$ space group. In yet another crystal form, more or less unfavorable

positioning of W1 occurred asymmetrically within the hexamer (fig. S1B). On the basis of nine structures (Fig. 2F, fig. S1, and supplementary text), we identified S157P and Q153A as mutations that indirectly stabilized the less unfavorable positioning of W1 (blue subunit), resulting in higher C1-ATPase activity (Fig. 1J). By contrast, S157C stabilized the more unfavorable positioning of W1 (other subunit) through N-terminal fray of $\alpha 7$ (fig. S4) and repositioning of $\alpha 6$, resulting in lower C1-ATPase activity (Fig. 1J).

The second origin of slowness involves coupling of ATP hydrolysis to slow cis-trans isomerization of the peptide chain. In the prehydrolysis state, the peptide bond between D145 and S146 ($D^{145}S^{146}$ peptide) mainly adopted the cis conformation (Fig. 3A). By contrast, in the posthydrolysis state, the $D^{145}S^{146}$ peptide was entirely in the trans conformation (Fig. 3A and fig. S5A). This trans selectivity was achieved through hydrolysis of ATP bound in the counterclockwise (CCW) interface (Fig. 2C, dotted boxes: blue \rightarrow orange \rightarrow green; see also supplementary text), along with repositioning of helices $\alpha 6$ and $\alpha 7$ on the clockwise (CW) side (Fig. 3A). These structural observations support the idea that hydrolysis of ATP

bound in the CCW interface results in a conformational change that forces the $D^{145}S^{146}$ peptide to adopt the trans conformation.

According to our computational results, the prehydrolysis state containing the cis- $D^{145}S^{146}$ peptide must overcome a barrier of 14 to 16 kcal mol $^{-1}$ (Fig. 3B) for cis-trans isomerization and a barrier of 11 to 17 kcal mol $^{-1}$ (table S1) to disrupt the P_γ - $O_{3\beta}$ bond. These two events are potentially related to each other through repositioning of helices $\alpha 6$ to $\alpha 8$ (fig. S5G and supplementary text). Given a frequency factor of $\sim 10^{12} \text{ s}^{-1}$, the rate of a reaction that must overcome a barrier of $\sim 22 \text{ kcal mol}^{-1}$ is $\sim 0.5 \text{ hours}^{-1}$ ($\sim 12 \text{ ATP day}^{-1}$) at 30°C . The impact of cis-trans flipping was confirmed using the KaiC1-S146P mutant, in which the $D^{145}S^{146}$ peptide is forced to adopt the cis conformation (fig. S5B). Moreover, full-length KaiC-S146P exhibited a notable decrease in steady-state ATPase activity (4.8 ATP per day per KaiC), as well as a nearly 50% increase in the phosphorylation cycle period (34 hours). Thus, cis-trans flipping of the $D^{145}S^{146}$ peptide could impose a substantial energy barrier on the hydrolysis-coupled transition of the subunits.



In full-length KaiC, the slow C1-ATPase is integrated with another ATPase in the C2 domain (C2-ATPase), forming a coupled C1-C2-ATPase system. At steady state, the C1-ATPase predominated over the C2-ATPase (Fig. 1J). The pacemaking role of C1-ATPase became obvious during the approach to the steady state (pre-steady state). Immediately after the addition of excess ATP to full-length apo-KaiC protein (26), nascent KaiC-WT hydrolyzed 35 ATP per day per KaiC at 30°C (Fig. 4A and fig. S6A). The activity suddenly decreased to ~8 ATP per day per KaiC during the first 6 hours and then stably recovered to and remained at 13 ± 1 ATP per day per KaiC, the activity required for steady-state KaiC-WT (11). The biphasic-exponential phases reflected the pre-steady-state relaxation of the coupled C1-C2-ATPase (supplementary text). Minimal dependencies on ATP, ADP, and KaiC protein concentrations (fig. S6B) further supported the hydrolysis-related maturation of ATP- or ADP-bound KaiC-WT hexamer (supplementary text). The phosphorylation

state of the C2 domain was only weakly associated with the slow process, as the phosphorylation-mimic KaiC-S431D/T432E exhibited detectable relaxation (fig. S6C).

Response speed of the C1-C2-ATPase is controlled by the C1-ATPase. Long-period C1 mutants (T42S, S157C, S48T, and A251V) with suppressed C1-ATPase activity exhibited slower relaxations of the C1-C2-ATPase than KaiC-WT, whereas short-period C1 mutants (Q153A and S157P) with elevated C1-ATPase activity exhibited more rapid relaxation with a deeper undershoot (Fig. 4B). These observations suggested that steady-state C1-ATPase activity governs the speed of pre-steady-state relaxation of the C1-C2-ATPase and oscillation period in vitro (Fig. 1G) and in vivo (Fig. 1D). The undamped natural frequency (ω_n) is another measure of the response speed of the coupled C1-C2-ATPase, as determined by fitting each transient curve of ATPase activity (Fig. 4B) to the initial-state response curve of the quasi-second-order system (supplementary text). Furthermore, the ω_n

value of KaiC-WT alone in vitro was 0.91 ± 0.01 day⁻¹, matching the circadian oscillatory frequency. We confirmed fine correlations among steady-state C1-ATPase activity, ω_n value, and in vivo frequency of TTOs for a series of the period-modulating C1 mutations (Fig. 1, B, D, and I). Thus, the slow C1-ATPase contributes to the circadian pacemaker for both the C1-C2-ATPase system and TTOs.

Slow relaxation of the C1-C2-ATPase was related to emergence of temperature compensation in full-length KaiC. Early ATPase activity (<2 hours in Fig. 4C) decayed in a temperature-dependent manner (thermal sensitivity $Q_{10}^{t=0} = 1.4$) (inset of Fig. 4C), whereas steady-state activity after the later decay (>24 hours) was almost independent of temperature ($Q_{10}^{t=\infty} \approx 1.1$). Because the contribution of C2-ATPase was negligibly small at steady state (Fig. 1J), the activity that did not vary with temperature can be attributed to the C1-ATPase. Thus, steady-state C1-ATPase activity may be kept constant in each hexamer

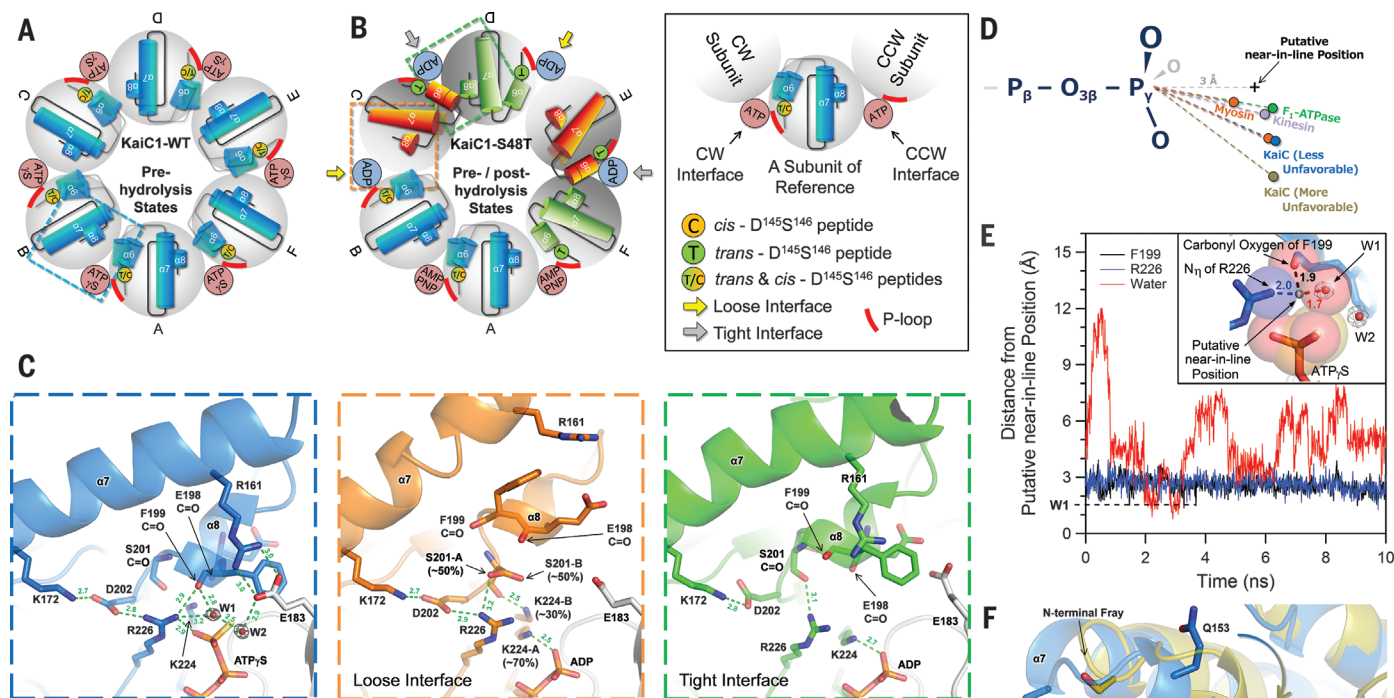


Fig. 2. Reaction cycle of the C1-ATPase, a slow but stable ATPase that serves as the basic timing cue.

(A) Schematic overview of the crystal structure of ATP- γ -S-bound KaiC1-WT ($P_{21}2_12_1$). The drawing in the box to the right of (B) summarizes the nomenclature used to describe the structures. For the seven other crystal structures, refer to details in fig. S1. (B) Schematic overview of the crystal structure of adenylyl imidodiphosphate (AMP-PNP), or ADP-bound KaiC1-S48T ($P_{21}2_12_1$). (C) Zoomed-in views of the prehydrolysis A-B interface (blue dotted box) of Fig. 2A, posthydrolysis B-C interface (orange dotted box) of Fig. 2B, and posthydrolysis C-D interface (green dotted box) of Fig. 2B. The mesh indicates the $F_{\text{obs}} - F_{\text{calc}}$ omit map of a potential lytic water molecule (W1) and another water molecule (W2), contoured at 4σ . (D) Schematic drawing of W1 positioning (filled circles) in KaiC1-WT, along with motor proteins that hydrolyze ATP efficiently (table S1). (E) Distances of a potential lytic water, the carbonyl oxygen atom of F199, and N_{η} of R226 from the putative near-in-line position during a 10-ns molecular dynamics simulation. The inset describes the crystal structure of the prehydrolysis state of KaiC1-WT, with van der Waals radii depicting surfaces for the oxygen atom of W1, the carbonyl oxygen atom of F199, N_{η} of R226, and the γ -phosphate group. (F) Less- and more-unfavorable positioning of W1 confirmed in the $P_{21}2_12_1$ (S146G, blue) and $P_{31}2_1$ (WT, ochre) crystal structures, respectively. The N-terminal fray of the α_7 helix at S157, followed by the repositioning of the α_6 helix (Q153), caused the rearrangement of the E77-E78 pair, through which the γ -phosphate group was rotated by $\sim 25^\circ$ and W1 moved away from P_{γ} .

by adjusting the ratio of prehydrolysis blue and other subunits and by scrambling their spatial arrangement (Fig. 2A and fig. S1, A and B). In this respect, oscillation of ATPase activity (Fig. 4A) is realized by perturbing the intramolecular

homeostasis of KaiC ATPase via intermolecular interactions with KaiA and KaiB.

The C1-C2-ATPase is coupled indirectly to interactions with KaiA and KaiB. Binding affinities of KaiA and KaiB for pre-steady-state

KaiC-WT were modulated in a biphasic and parabolic manner on the same time scale as relaxation of the C1-C2-ATPase (fig. S7, A and B). However, such time evolution was lost in KaiC-S431D/T432E (fig. S7, A and B), which retained

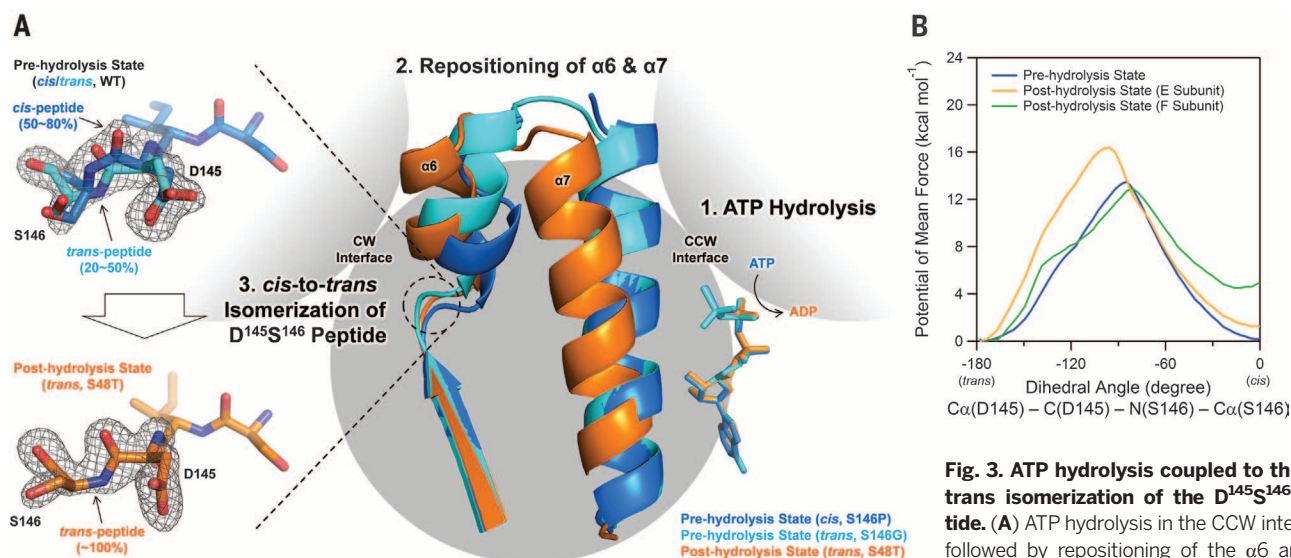


Fig. 3. ATP hydrolysis coupled to the cis-trans isomerization of the D¹⁴⁵S¹⁴⁶ peptide.

tion of the D¹⁴⁵S¹⁴⁶ peptide. The prehydrolysis states possessing the cis- D¹⁴⁵S¹⁴⁶ peptide (KaiC-S146P, dark blue) (fig. S5B) and the trans-D¹⁴⁵S¹⁴⁶ peptide (KaiC-S146G, light blue) (fig. S5C) are superimposed on the posthydrolysis state (S48T, orange). The location of the D¹⁴⁵S¹⁴⁶ peptide is highlighted by a dotted circle. The meshes represent the $F_{\text{obs}} - F_{\text{calc}}$ omit maps contoured at 3σ . In the prehydrolysis state, the fraction of the cis-D¹⁴⁵S¹⁴⁶ peptide ranged from 0.5 to 0.8, depending on the subunit position and the crystal forms (Fig. 2 and fig. S1). **(B)** Potential of mean force (kilocalories per mole) for cis-trans isomerization of the D¹⁴⁵S¹⁴⁶ peptide as a function of the dihedral angle $\text{Ca}(\text{D145})-\text{N}(\text{D145})-\text{C}(\text{S146})-\text{Ca}(\text{S146})$.

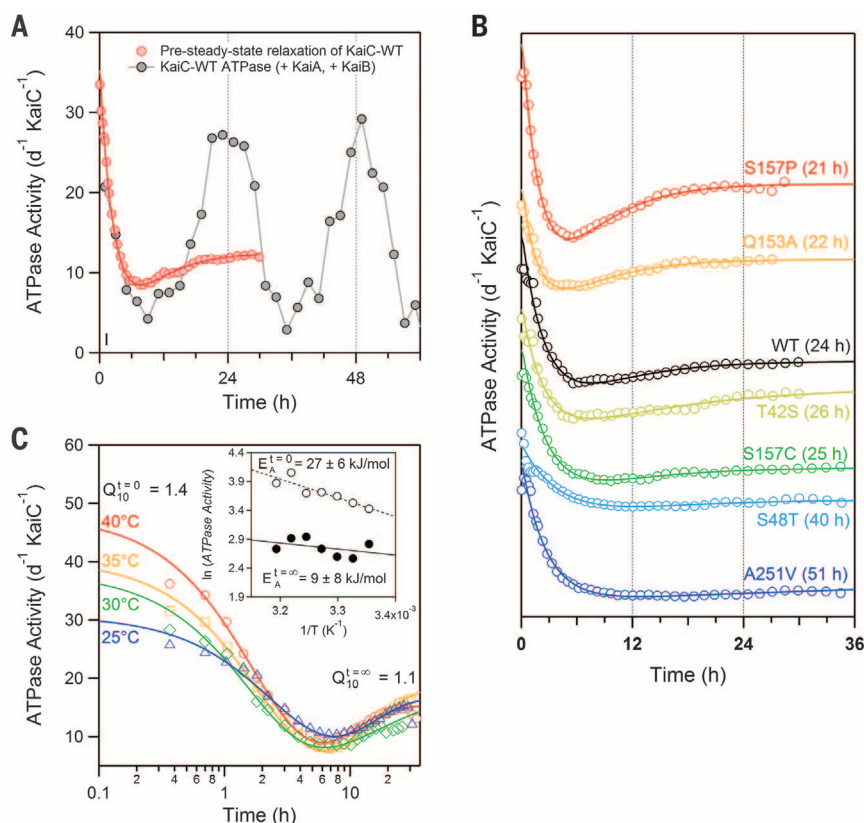


Fig. 4. Circadian periodicity determined by KaiC ATPase activity.

(A) Damped oscillation observed for pre-steady-state relaxation of KaiC ATPase as the source of the circadian time scale at 30°C. In the presence of both KaiA and KaiB, the ATPase activity of KaiC-WT exhibits a stable oscillation around its steady-state ATPase activity. The bar at lower left indicates the maximal contribution from ATP synthesis during the first hour (27). **(B)** Dynamic responses of the ATPase activity observed for period mutants of full-length KaiC. Each curve is offset longitudinally for clarity of presentation. The values in parentheses represent the periods of in vitro phosphorylation rhythms. **(C)** Temperature dependency of pre-steady-state dynamics of ATPase activity in KaiC-WT. The insets represent Arrhenius plots. Errors in values are derived from a linear regression analysis. E_A represents the apparent activation energy.

pre-steady-state relaxation of C1/C2-ATPase (fig. S6C). Thus, C1-C2-ATPase status is coupled to intermolecular interactions with KaiA and KaiB through a phosphorylation-dependent conformational change of KaiC. This idea is also supported by the fact that the C2-ATPase is essential for the complete autodephosphorylation (27). Thus, absolute slowness encoded in the C1-ATPase is transferred to and correlated with the TTO cycle (Fig. 1) through a pathway in which non-steady-state relaxation of C1-C2-ATPase leads to cycles of phosphorylation that alter intermolecular interactions with KaiA or KaiB or other clock-related proteins (28, 29).

In cyanobacteria, slow homeostatic regulation of the coupled C1-C2-ATPase is critical for circadian periodicity (Figs. 1 and 4A) and also has an important role in entrainment of individual KaiC molecules by external stimuli such as temperature changes (30). Our results suggest how ancient cyanobacteria have incorporated Earth's rotation period into their molecular systems.

REFERENCES AND NOTES

- C. S. Pittendrigh, *Annu. Rev. Physiol.* **55**, 17–54 (1993).
- D. Bell-Pedersen et al., *Nat. Rev. Genet.* **6**, 544–556 (2005).
- J. S. Takahashi, H. K. Hong, C. H. Ko, E. L. McDermott, *Nat. Rev. Genet.* **9**, 764–775 (2008).
- B. Schwanhäusser et al., *Nature* **473**, 337–342 (2011).
- J. Tomita, M. Nakajima, T. Kondo, H. Iwasaki, *Science* **307**, 251–254 (2005).
- M. Nakajima et al., *Science* **308**, 414–415 (2005).
- J. S. O'Neill, A. B. Reddy, *Nature* **469**, 498–503 (2011).
- M. Karplus, *J. Phys. Chem. B* **104**, 11–27 (2000).
- S. Akiyama, *Cell. Mol. Life Sci.* **69**, 2147–2160 (2012).
- M. Ishiura et al., *Science* **281**, 1519–1523 (1998).
- K. Terauchi et al., *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16377–16381 (2007).
- Y. Murayama et al., *EMBO J.* **30**, 68–78 (2011).
- T. Nishiwaki et al., *EMBO J.* **26**, 4029–4037 (2007).
- M. J. Rust, J. S. Markson, W. S. Lane, D. S. Fisher, E. K. O'Shea, *Science* **318**, 809–812 (2007).
- Y. G. Chang, N. W. Kuo, R. Tseng, A. LiWang, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 14431–14436 (2011).
- Y. G. Chang, R. Tseng, N. W. Kuo, A. LiWang, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16847–16851 (2012).
- S. Akiyama, A. Nohara, K. Ito, Y. Maeda, *Mol. Cell* **29**, 703–716 (2008).
- H. Kageyama et al., *Mol. Cell* **23**, 161–171 (2006).
- X. Qin et al., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14805–14810 (2010).
- R. Tseng et al., *J. Mol. Biol.* **426**, 389–402 (2014).
- Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
- T. Kondo, in *Clocks and Rhythms*, B. Stillman, D. Stewart, Eds. (Cold Spring Harbor Laboratory Press, New York, 2008), pp. 47–55.
- C. L. Parke, E. J. Wojcik, S. Kim, D. K. Worthylake, *J. Biol. Chem.* **285**, 5859–5867 (2010).
- C. A. Smith, I. Rayment, *Biochemistry* **35**, 5404–5417 (1996).
- M. W. Bowler, M. G. Montgomery, A. G. Leslie, J. E. Walker, *J. Biol. Chem.* **282**, 14238–14242 (2007).
- A. Mukaiyama, M. Osako, T. Hikima, T. Kondo, S. Akiyama, *BIOPHYSICS* **11**, 79–84 (2015).
- T. Nishiwaki, T. Kondo, *J. Biol. Chem.* **287**, 18030–18035 (2012).
- M. Hanaoka et al., *J. Biol. Chem.* **287**, 26321–26327 (2012).
- H. Iwasaki et al., *Cell* **101**, 223–233 (2000).
- T. Yoshida, Y. Murayama, H. Ito, H. Kageyama, T. Kondo, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1648–1653 (2009).

ACKNOWLEDGMENTS

Diffraction data were collected at BL44XU in the SPring-8 facility under the proposals 2009A6902, 2009B6902, 2010A6502, 2010B6502, 2011A6602, 2011B6602, 2012A6702, 2012B6702, and 2013B6700. MX225-HE was financially supported by Academia Sinica and National Synchrotron Radiation Research Center

(Taiwan, China). This work was supported by Grants-in-Aid for Scientific Research (25291039, 22687010, and 26102544 to S.A.; 24000016 to T.K.; 26888019 to T.M.; and 25288011 to S.Saito); the Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology of Japan; and partly by the Okazaki ORION project. The calculations were in part performed at the Research Center for Computational Science in Okazaki, Japan. S.A. designed the experiments. A.M., J.A., J.W., and M.O. conducted ATPase experiments. S.Son screened initial crystallization conditions, and J.A. and T.B.H. further refined them. S.Son, J.A., T.B.H., and S.A. collected diffraction data and analyzed the structures with inputs from E.Y. T.M. and S.Saito conducted molecular dynamics simulations. S.A., J.A., T.B.H., A.M., T.M., S.Saito, and T.K. drafted the manuscript with input from all authors. We declare no conflicts of interest. The atomic coordinates and structure factors are deposited in the Protein Data Bank with

accession codes 4TL8 (KaiC1-WT, P2;2;2), 4TL9 (ATPyS-bound KaiC1-S48T), 4TLA (ADP-bound KaiC1-S48T), 4TLB (KaiC1-S146P), 4TLC (KaiC1-S146G), 4TLD (KaiC1-S157P), 4TLE (KaiC1-S157C), 4TL7 (KaiC1-WT, P3;21), and 4TL6 (KaiC1-WT, C2).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/312/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S3
References (31–51)
Movies S1 to S3

9 September 2014; accepted 10 June 2015
Published online 25 June 2015;
10.1126/science.1261040

INFLAMMATION

Neutrophil extracellular traps license macrophages for cytokine production in atherosclerosis

Annika Warnatsch, Marianna Ioannou, Qian Wang, Venizelos Papayannopoulos*

Secretion of the cytokine interleukin-1 β (IL-1 β) by macrophages, a major driver of pathogenesis in atherosclerosis, requires two steps: Priming signals promote transcription of immature IL-1 β , and then endogenous “danger” signals activate innate immune signaling complexes called inflammasomes to process IL-1 β for secretion. Although cholesterol crystals are known to act as danger signals in atherosclerosis, what primes IL-1 β transcription remains elusive. Using a murine model of atherosclerosis, we found that cholesterol crystals acted both as priming and danger signals for IL-1 β production. Cholesterol crystals triggered neutrophils to release neutrophil extracellular traps (NETs). NETs primed macrophages for cytokine release, activating T helper 17 (T_H17) cells that amplify immune cell recruitment in atherosclerotic plaques. Therefore, danger signals may drive sterile inflammation, such as that seen in atherosclerosis, through their interactions with neutrophils.

Inflammation is critical against infection but must be regulated by multiple checkpoints to prevent inflammatory disease (1). During infection, cytokine transcription is triggered by microbial molecules that activate pattern recognition receptors (2). Release of mature active cytokines requires additional “danger” signals associated with host cell damage. Known as danger-associated molecular patterns (DAMPs), these secondary signals activate NLRP3 and other inflammasomes, promoting cleavage and activation of the protease caspase-1 that processes cytokines such as interleukin-1 β (IL-1 β) into their mature form (3).

IL-1 β plays a critical role in the development of atherosclerosis and other inflammatory diseases. Because of its low solubility, cholesterol crystallizes in circulation and is taken up by monocyte-derived macrophages (4–6), activating their inflammasomes to release IL-1 β and other proinflammatory cytokines (7). These molecules recruit myeloid cells to the vascular endothelium, where their cholesterol content generates obstructive lesions

(8). In atherosclerosis and other sterile inflammatory diseases, the endogenous priming signals that activate IL-1 β transcription prior to inflammasome activation remain unknown.

IL-1 β up-regulates chemokines that recruit neutrophils to atherosclerotic lesions (9–11). Neutrophils are implicated in disease (12, 13), but their role in pathogenesis remains poorly understood. To combat pathogens that evade phagocytosis (14), neutrophils release neutrophil extracellular traps (NETs) composed of decondensed chromatin and antimicrobials (15). NETs are implicated in several inflammatory diseases (16), but their pathogenic mechanism and role in atherosclerosis are unclear.

To examine how neutrophils respond during atherosclerosis, we investigated the effect of cholesterol crystals on human blood-derived neutrophils. Cholesterol crystals induced NET formation (NETosis) (Fig. 1, A and B) at concentrations required to activate the inflammasome (fig. S1, A and B) (7) as efficiently as microbes (14), triggering a reactive oxygen species (ROS) burst (fig. S1C) and neutrophil elastase (NE) translocation to the nucleus (fig. S1D), a critical step for NETosis (17). NETosis depended on ROS, as it was blocked by diphenylene iodonium [DPI]; an inhibitor of

Mill Hill Laboratory, Francis Crick Institute, London NW7 1AA, UK. *Corresponding author. E-mail: veni.p@crick.ac.uk

reduced nicotinamide adenine dinucleotide phosphate (NADPH) oxidase] or an inhibitor (NEi) of the neutrophil-specific proteases NE and proteinase 3 (PR3) (17) but not by Cl-amidine, which inhibits peptidylarginine deiminase (PAD) enzymes implicated in NETosis (18) (Fig. 1, A and B, and fig. S1E). Consistently, DPI or NEi blocked NE translocation to the nucleus driven by cholesterol (fig. S1D) (19).

Next, we examined whether NETs form during atherosclerosis. Previous studies reported the presence of NETs in lesions but showed intact neutrophils with condensed nuclei (20) or luminal rather than lesion-associated neutrophils in the absence of specific NET markers (21). We detected NETs as large amorphous extracellular structures in atherosclerotic lesions from apolipoprotein E (ApoE)-deficient mice that were placed on a high-fat diet (HFD) for 8 weeks to induce hypercholesterolemia (Fig. 1C). NETs formed in cholesterol-rich areas but were absent from adjacent adventitia (fig. S2A). To block NETosis in vivo, we crossed ApoE-deficient animals with mice deficient in PR3 and NE, because deleting both enzymes may abrogate NETosis more effec-

tively (22). NETs were completely absent in lesions of ApoE/PR3/NE-deficient mice after 8 weeks on HFD (Fig. 1C) and ApoE-deficient mice after 6 weeks on HFD receiving deoxyribonuclease (DNase), which degrades NETs (23) (fig. S2B).

Subsequently, we assessed the effect of NET deficiency on atherosclerosis. When placed on HFD, ApoE- and ApoE/PR3/NE-deficient mice exhibited similar weight gain (fig. S3A) and blood cholesterol, triglyceride, and low-density lipoprotein (LDL) concentrations (fig. S3B). Analysis of aortic root cross sections showed that the two groups were modestly different after 4 weeks on HFD (fig. S3, C and D), suggesting that NETs did not play a critical role early during atherogenesis. However, after 8 weeks on HFD, ApoE/PR3/NE-deficient mice exhibited a factor of 3 reduction in plaque size relative to ApoE-deficient controls (Fig. 2, A and B). These differences were also reflected by en face analysis of intact aortas (fig. S3, E and F). DNase injections into ApoE-deficient mice on HFD for 6 weeks resulted in a comparable factor of 3 reduction in lesion size, which excludes the possibility that the proteases played NET-independent roles (Fig. 2, C and D). Lesion growth

was unaffected by DNase in ApoE/PR3/NE-deficient mice that lack NETs.

NET-deficient mice exhibited a reduction in lesion growth that was comparable to mice lacking NLRP3 or the IL-1 receptor; this finding suggested that NETs may drive atherosclerosis by modulating cytokine production. Indeed, IL-1 α , IL-1 β , and IL-6 were elevated in the plasma of ApoE-deficient animals after 8 weeks on HFD but were largely absent in ApoE/PR3/NE-deficient mice (Fig. 3A and fig. S4A). After 16 weeks on HFD, IL-1 α but not IL-1 β concentrations were still elevated in ApoE-deficient controls relative to NET-deficient animals (fig. S4B). DNase administration abrogated plasma cytokine concentrations in ApoE-deficient controls but had no effect on ApoE/PR3/NE-deficient mice (fig. S4C). Furthermore, IL-1 β staining, which detects both immature and mature protein, was prominent in lesions from ApoE-deficient mice but was absent in ApoE/PR3/NE knockout animals and colocalized with NETs and macrophages (Fig. 3B and fig. S4D). By contrast, IL-1 β concentrations were similar in the spleen or in blood mononuclear cells (fig. S4E), although we cannot exclude that systemic cytokines were not produced differentially elsewhere. In addition, IL-1 β mRNA concentrations were significantly reduced in aortas from ApoE/PR3/NE-deficient mice (Fig. 3C). Together, these data indicate that NE and PR3 were not mediating IL-1 β maturation posttranslationally, and instead suggest that NETs are essential for the transcription of proinflammatory cytokines.

The requirement of NETs for cytokine production, and the proximity of NETs to macrophages (Fig. 3D) and IL-1 β in lesions (Fig. 3B), prompted us to examine whether NETs regulate cytokine production by macrophages. We prepared NETs from cholesterol crystal-stimulated neutrophils (fig. S5A) (24) and investigated their effects on CD14-purified, blood-derived human monocytes in vitro. Stimulation with NETs or cholesterol crystals separately yielded minor increases in IL-1 β and IL-6 concentrations in culture supernatants (Fig. 3E and fig. S5B). In contrast, monocytes released substantial cytokine concentrations when pretreated with NETs and subsequently stimulated with cholesterol crystals (Fig. 3E). By comparison, neutrophils were not a major source of cytokines, as they released negligible concentrations in response to cholesterol crystals (fig. S5C). Degradation of NETs by DNase treatment (fig. S4A) abrogated cytokine release (fig. S5B), indicating the requirement for a DNA moiety. Consistently, an oligonucleotide (ODN) antagonist of the pattern recognition DNA receptor Toll-like receptor 9 (TLR9) significantly reduced cytokine release in NET-treated monocytes, but not monocytes primed with bacterial lipopolysaccharide (LPS) (Fig. 3E). Because TLR9 is not expressed in monocytes, these data suggest that DNA is important for monocyte activation but that its detection is mediated via other DNA receptors blocked by ODN. Blocking with oligonucleotide did not fully inhibit IL-1 β induction, so we reasoned that additional non-DNA NET factors contribute to monocyte activation. Blocking TLR2 and TLR4, which bind endogenous proteins,

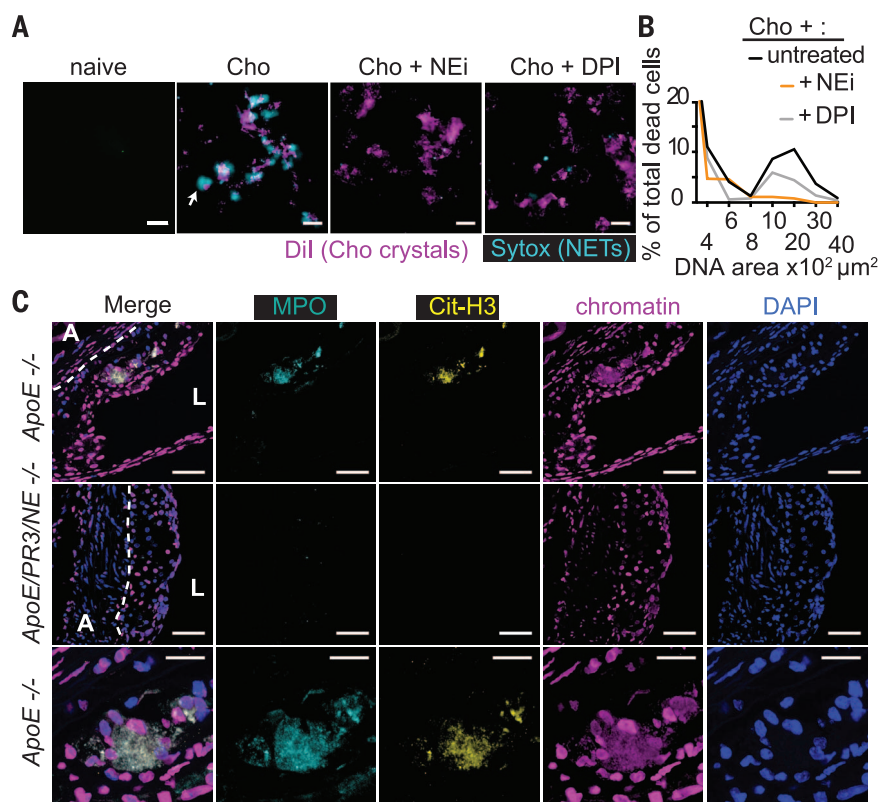


Fig. 1. Cholesterol crystals trigger NETosis. (A) Fluorescence micrograph of neutrophils incubated with cholesterol crystals and stained with the lipid dye Dil (magenta) and extracellular DNA (Sytox, cyan). Neutrophils were left untreated or treated with NE inhibitor (NEi) or the NADPH oxidase inhibitor DPI. Scale bars, 100 μ m. (B) Quantitation of NETosis in (A). Data are representative of three independent experiments. (C) Representative confocal immunofluorescence microscopy images of aortic root sections from ApoE $^{-/-}$ and ApoE/PR3/NE $^{-/-}$ mice on HFD for 8 weeks and stained for MPO (cyan), citrullinated histone 3 (Cit-H3, yellow), chromatin (magenta), and DNA (DAPI, blue). Borders between the adventitia (A) and the lesion (dotted line) and lumen (L) are shown; scale bars, 50 μ m. The third row shows detail from the first row (arrow); scale bars, 20 μ m. Data are representative of eight mice analyzed per strain from two independent experiments.

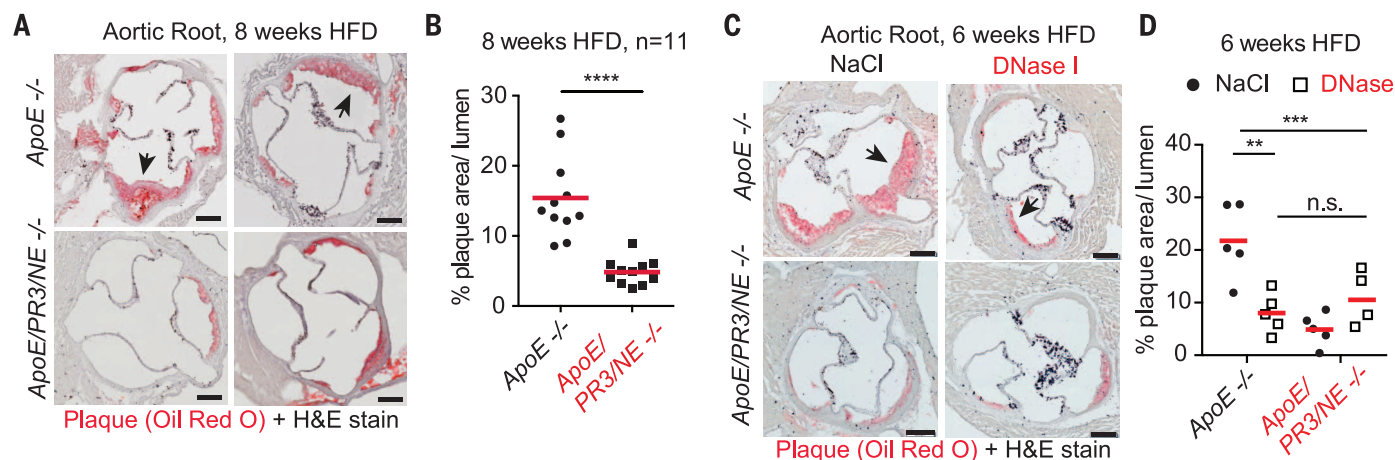


Fig. 2. NETs promote atherosclerosis. (A) Two representative microscopy images of aortic root sections from *ApoE*^{-/-} and *ApoE/PR3/NE*^{-/-} mice on HFD for 8 weeks and stained for lipid (Oil Red O, red) and hematoxylin. Scale bars, 200 μ m. (B) Quantitation of plaque area relative to the area of the aortic lumen from (A); data are representative of 11 mice per strain pooled from two independent experiments. Each point is the mean from multiple sections per animal. (C) Representative microscopy images of aortic root sections from *ApoE*^{-/-} and *ApoE/PR3/NE*^{-/-} mice on HFD for 6 weeks and regularly injected

intravenously with 120 U of DNase or vehicle control (0.9% NaCl). Stained for lipid (Oil Red O, red) and hematoxylin; scale bars, 200 μ m. (D) Quantitation of (C) as in (B). Data are representative of four or five mice analyzed per strain and condition. A power analysis revealed 92% power for the difference of means between NaCl- and DNase-treated *ApoE*^{-/-} mice. Statistics by Student's *t* test for single comparison and two-way analysis of variance (ANOVA) followed by Sidak's multiple comparison post test for multiple comparisons: ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001; n.s., not significant.

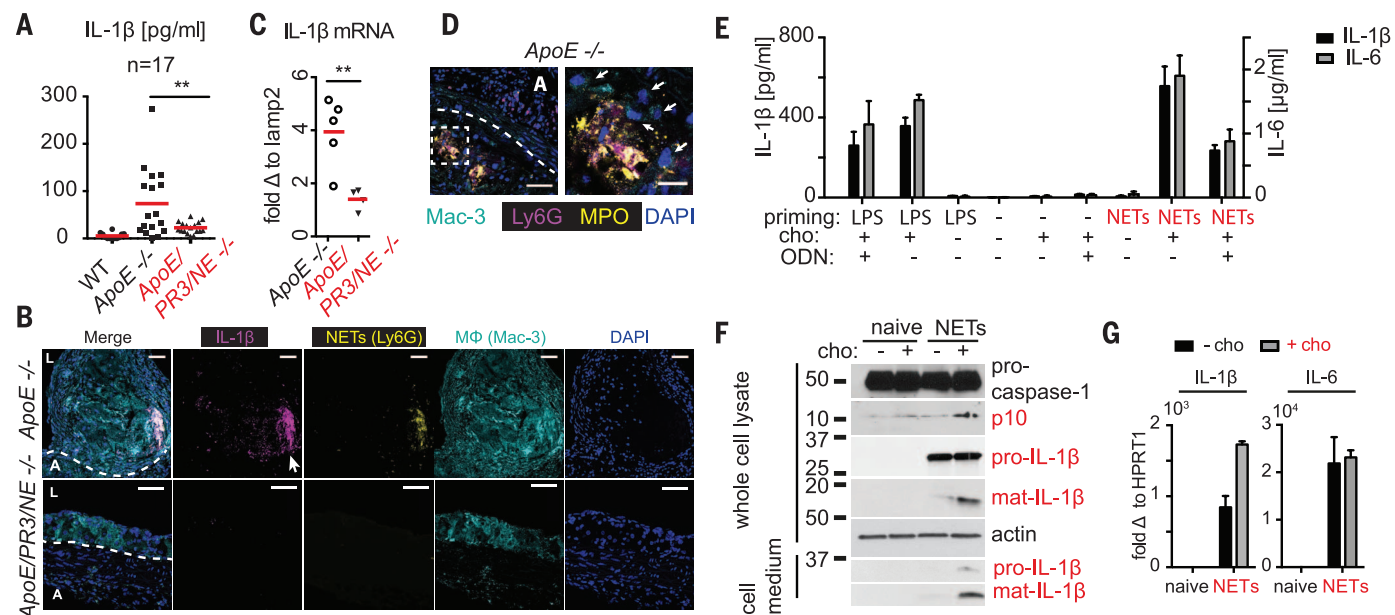


Fig. 3. NETs prime macrophages for cytokine release. (A) Plasma levels of IL-1 β from wild-type (WT), *ApoE*^{-/-}, and *ApoE/PR3/NE*^{-/-} mice on HFD for 8 weeks, measured by enzyme-linked immunosorbent assay (ELISA) in *n* = 17 mice per strain pooled from three independent experiments. (B) Representative confocal immunofluorescence microscopy images of aortic root sections from eight *ApoE*^{-/-} and five *ApoE/PR3/NE*^{-/-} mice on HFD for 8 weeks and stained with the macrophage marker Mac-3 (cyan), IL-1 β (magenta), the neutrophil marker Ly6G (yellow), and DNA (DAPI, blue) in two independent experiments. Dashed line denotes the adventitia (A)–lesion boundary; L, lumen. Scale bars, 50 μ m. (C) Representative IL-1 β mRNA levels in aorta of five *ApoE*^{-/-} and four *ApoE/PR3/NE*^{-/-} mice fed HFD for 8 weeks, repeated in two independent experiments and measured by quantitative polymerase chain reaction. mRNA levels were normalized to the monocyte/macrophage-specific gene *lamp2* and expressed relative to levels measured in wild-type mice. A power analysis measured 89% power for the difference of means between *ApoE*^{-/-} and *ApoE/PR3/NE*^{-/-} mice. (D) Representative confocal immunofluorescence microscopy images of aortic root sections from five *ApoE*^{-/-} mice on HFD

for 8 weeks and stained with the macrophage marker Mac-3 (cyan), Ly6G (magenta), MPO (yellow), and DNA (DAPI, blue). Scale bars, 50 μ m. Right panel is a close-up of the boxed area of the left panel; arrowheads point to macrophages. Scale bars, 20 μ m. (E) Mature IL-1 β (black bars, left y axis) or IL-6 (gray bars, right y axis) protein released by CD-14 blood-derived human monocytes untreated or treated with LPS or NETs alone or in the presence of cholesterol crystals. Where indicated, cells were treated with oligonucleotide inhibitor (ODN; 10 μ g/ml). (F) Whole-cell lysates or cell culture medium from naive CD-14 blood-derived human monocytes were treated with NETs or cholesterol crystals, analyzed by SDS–polyacrylamide gel electrophoresis, and immunoblotted for IL-1 β , caspase-1, and actin. (G) IL-1 β (left panel) or IL-6 (right panel) mRNA in naive CD-14 blood-derived human monocytes or treated with NETs alone (black bars) or in the presence of cholesterol crystals (gray bars). Statistics in (A) and (C) by two-tailed, unpaired Student's *t* test for single comparison and one-way ANOVA, followed by Tukey's multiple comparison posttest for multiple comparisons: ***P* < 0.01. In (E) and (G), data are representative of three independent experiments across three technical replicates; error bars denote SD.

decreased IL-1 β release and synergized with oligonucleotide inhibition, indicating that both protein and DNA moieties are important in NET-mediated priming (fig. S5D). The complete abrogation by DNase suggests that the association of these moieties is critical. In the absence of cholesterol, NETs did not induce substantial inflammasome activation, as reflected by the lack of caspase-1 and IL-1 β maturation (Fig. 3F), which were observed only upon costimulation with cholesterol crystals. Stimulation with NETs also up-regulated monocytic cytokine transcripts (Fig. 3G). These data are consistent with NETs providing priming signals in atherosclerosis.

IL-1 β up-regulates the T cell-derived cytokine IL-17, which drives the chemokines CXCL1 and CXCL2 to promote neutrophil recruitment during inflammation (9, 25). Both groups of mice exhibited comparable numbers of T cells (figs. S6 and S7), but aortas from ApoE/PR3/NE-deficient mice on 8-week HFD contained few IL-17 $^{+}$ T cells relative to ApoE-deficient controls (Fig. 4, A and B). IL-17-producing T cells could not be detected in aortas of ApoE $^{-/-}$ animals after 4 weeks on HFD (fig. S8, A and B), which suggests that strong T cell activation does not precede NET-driven

inflammation. The blood of NET-deficient mice did not exhibit alterations in immune cell populations, and circulating IL-17 $^{+}$ T cells were absent in both groups (figs. S9 and S10). The spleen and lymph nodes contained a small IL-17 $^{+}$ $\gamma\delta$ T cell population but no IL-17 $^{+}$ $\alpha\beta$ T cells (figs. S11 and S12, A and B). Furthermore, concentrations of IL-17A, CXCL1, CXCL2, and the monocyte chemokine CCL2 were also significantly reduced in aortas of ApoE/PR3/NE-deficient mice (Fig. 4C). Consistently, we counted fewer neutrophils in lesions and adventitia of ApoE/PR3/NE-deficient mice, by microscopy (Fig. 4D) and fluorescence-activated cell sorting (FACS) (fig. S7B). Although both groups contained a comparable makeup of immune cells (fig. S7A), ApoE/PR3/NE-deficient animals had significantly lower total immune cell counts per aorta (fig. S7B), consistent with reduced inflammation and smaller lesions. Adhesion molecule transcripts were similar in aortas from both groups, but differences may be difficult to detect because of the patchy lesion morphology (fig. S12C).

Neutrophil recruitment was comparable in the skin of wild-type and PR3/NE-deficient animals treated with Aldara imiquimod (fig. S13, A and B), which drives IL-1 exogenously to promote sterile

psoriatic inflammation (26). Hence, the reduction in neutrophil recruitment in ApoE/PR3/NE-deficient lesions was not due to intrinsic defects in neutrophil chemotaxis or extravasation. Therefore, NET-mediated priming of macrophages promotes a self-amplifying IL-1-IL-17 cascade and uncovers a mechanism for neutrophils to regulate T helper 17 (T_H17) cells that sustains chronic sterile inflammation (fig. S13C).

Our data reveal a requirement of NETs as priming cues in vivo and show that NETs are substantially more potent in priming cytokines than in activating the inflammasome. Although other endogenous molecules such as oxidized LDL can prime in vitro, their importance in vivo has not been demonstrated (7, 27). Interestingly, antibodies against oxidized LDL in lesions recognize oxidized phospholipids that suppress inflammation (28). NETosis may prime more efficiently than necrosis (29, 30), as it effectively exposes highly decondensed and proinflammatory DNA (31).

A recent study using Cl-amidine proposed that NETs drive a plasmacytoid dendritic cell (pDC)-derived interferon- α (IFN- α) autoimmune cascade in atherosclerosis via TLR9 ligation (20). However, TLR9 deficiency has little effect on atherogenesis (32). Moreover, PAD enzymes are expressed in many cell types (33) and were dispensable for NETosis triggered by cholesterol crystals (fig. S1E). Whereas IFN signaling down-regulates IL-1 β expression, genetic ablation of the IFN α/β receptor yields a modest 25% decrease in lesion size (34) because IFNs may contribute to pathogenesis via the up-regulation of caspases (3, 35, 36).

NET priming may drive inflammation in other NET-associated diseases such as cystic fibrosis and rheumatoid arthritis (24, 37). In contrast to the broad importance of IL-1 and IL-17, NETs play more specialized roles in immune defense (14) and NET-deficient individuals are primarily susceptible to localized fungal infection (38). Hence, blocking NETosis or degrading NETs may help to treat inflammatory diseases.

REFERENCES AND NOTES

1. M. Larnkanfi, V. M. Dixit, *Cell* **157**, 1013–1022 (2014).
2. K. Schroder, J. Tschopp, *Cell* **140**, 821–832 (2010).
3. E. Latz, T. S. Xiao, A. Stutz, *Nat. Rev. Immunol.* **13**, 397–411 (2013).
4. S. M. Lessner, H. L. Prado, E. K. Waller, Z. S. Galis, *Am. J. Pathol.* **160**, 2145–2155 (2002).
5. F. K. Swirski et al., *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10340–10345 (2006).
6. L. Landsman et al., *Blood* **113**, 963–972 (2009).
7. P. Duewell et al., *Nature* **464**, 1357–1361 (2010).
8. A. Grebe, E. Latz, *Curr. Rheumatol. Rep.* **15**, 313 (2013).
9. L. S. Miller et al., *Immunity* **24**, 79–91 (2006).
10. T. Naruko et al., *Circulation* **106**, 2894–2900 (2002).
11. P. Rotzsch et al., *Am. J. Pathol.* **177**, 493–500 (2010).
12. M. Drechsler, R. T. Megens, M. van Zandvoort, C. Weber, O. Soehnlein, *Circulation* **122**, 1837–1845 (2010).
13. Y. Döring et al., *Circ. Res.* **110**, 1052–1056 (2012).
14. N. Branzk et al., *Nat. Immunol.* **15**, 1017–1025 (2014).
15. V. Brinkmann et al., *Science* **303**, 1532–1535 (2004).
16. N. Branzk, V. Papayannopoulos, *Semin. Immunopathol.* **35**, 513–530 (2013).
17. V. Papayannopoulos, K. D. Metzler, A. Hakkim, A. Zychlinsky, *J. Cell Biol.* **191**, 677–691 (2010).
18. Y. Wang et al., *J. Cell Biol.* **184**, 205–213 (2009).

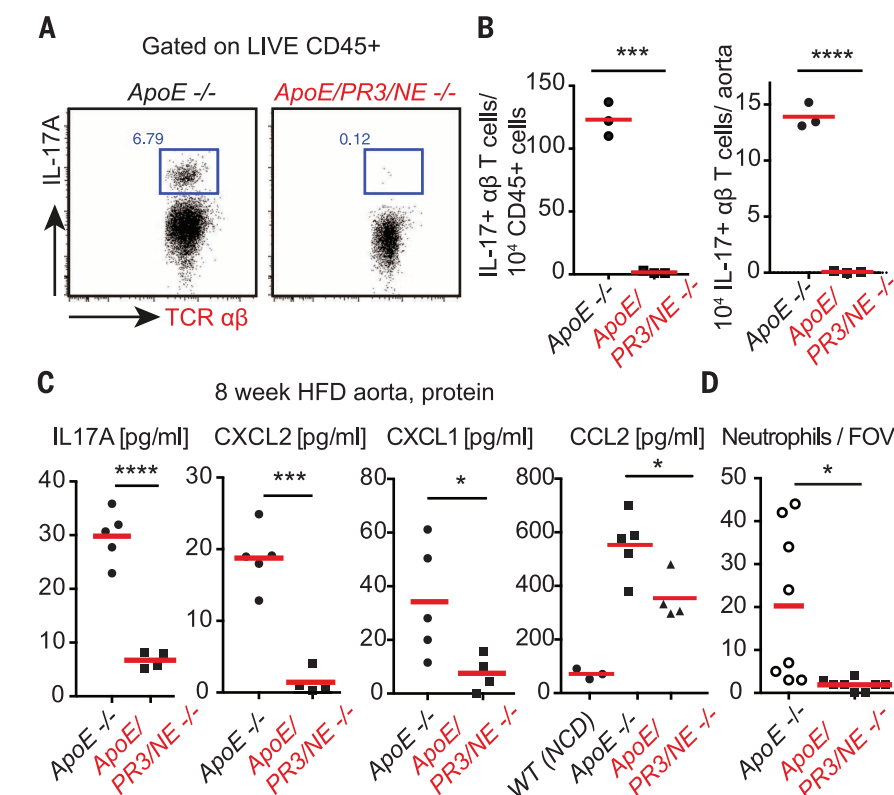


Fig. 4. NETs drive IL-17 and neutrophil chemokine production in atherosclerosis. (A) Representative FACS plot of IL-17 intracellular staining in phorbol myristate acetate (PMA)-restimulated $\alpha\beta$ and $\gamma\delta$ T cells from digested aortas of ApoE $^{-/-}$ or ApoE/PR3/NE $^{-/-}$ mice on HFD for 8 weeks. (B) Representative number of cells relative to CD45 $^{+}$ populations and whole aortas from (A) for three animals per group repeated in two independent experiments. (C) IL-17A, CXCL1, CXCL2, and CCL2 concentrations from whole aorta samples measured by ELISA; $n = 5$ ApoE $^{-/-}$ and $n = 4$ ApoE/PR3/NE $^{-/-}$ mice. (D) Number of total intact Ly6G-stained neutrophils per aortic root section field of view (FOV) measured in immunostained micrographs; eight mice per strain in two independent experiments. Statistics by two-tailed, unpaired Student's t test: * $P < 0.05$, *** $P < 0.001$, **** $P < 0.0001$.

19. K. D. Metzler, C. Goosmann, A. Lubojemska, A. Zychlinsky, V. Papayannopoulos, *Cell Rep.* **8**, 883–896 (2014).
20. J. S. Knight et al., *Circ. Res.* **114**, 947–956 (2014).
21. R. T. Megens et al., *Thromb. Haemost.* **107**, 597–598 (2012).
22. K. Kessenbrock et al., *J. Clin. Invest.* **118**, 2438–2447 (2008).
23. A. Hakikim et al., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9813–9818 (2010).
24. V. Papayannopoulos, D. Staab, A. Zychlinsky, *PLOS ONE* **6**, e28526 (2011).
25. H. Park et al., *Nat. Immunol.* **6**, 1133–1141 (2005).
26. A. Walter et al., *Nat. Commun.* **4**, 1560 (2013).
27. F. J. Sheedy et al., *Nat. Immunol.* **14**, 812–820 (2013).
28. P. Bretscher et al., *EMBO Mol. Med.* **7**, 593–607 (2015).
29. V. Hornung, E. Latz, *Nat. Rev. Immunol.* **10**, 123–130 (2010).
30. Y. Miyake, S. Yamasaki, *Adv. Exp. Med. Biol.* **738**, 144–152 (2012).
31. R. Lande et al., *Nature* **449**, 564–569 (2007).
32. C. Koulis et al., *Arterioscler. Thromb. Vasc. Biol.* **34**, 516–525 (2014).
33. S. Horibata, S. A. Coonrod, B. D. Cherrington, *J. Reprod. Dev.* **58**, 274–282 (2012).
34. P. Goossens et al., *Cell Metab.* **12**, 142–153 (2010).
35. V. A. Rathinam et al., *Cell* **150**, 606–619 (2012).
36. P. Broz et al., *Nature* **490**, 288–291 (2012).
37. R. Khandpur et al., *Sci. Transl. Med.* **5**, 178ra40 (2013).
38. K. D. Metzler et al., *Blood* **117**, 953–959 (2011).

ACKNOWLEDGMENTS

We thank Q. Xu for providing the ApoE^{-/-} mice and A. Zychlinsky for the PR3/NE^{-/-} mice; Z. Zhang for training; L. Mrowietz for help with NET preparations; and B. Stockinger, A. Zychlinsky, A. Schaefer, and M. Wilson for comments on the manuscript.

This work was supported by the UK Medical Research Council (grant MC_UP_1202/13) and was principally conducted at the MRC National Institute for Medical Research and completed at the Francis Crick Institute, which receives its core funding from the UK Medical Research Council, Cancer Research UK, and the Wellcome Trust. The data are contained in the manuscript and the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/316/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
References (39, 40)

30 January 2015; accepted 8 June 2015
10.1126/science.aaa8064

HIV-1 VACCINES

Protective efficacy of adenovirus/protein vaccines against SIV challenges in rhesus monkeys

Dan H. Barouch,^{1,2*} Galit Alter,² Thomas Broge,² Caitlyn Linde,² Margaret E. Ackerman,³ Eric P. Brown,³ Erica N. Borducchi,¹ Kaitlin M. Smith,¹ Joseph P. Nkolola,¹ Jinyan Liu,¹ Jennifer Shields,¹ Lily Parenteau,¹ James B. Whitney,¹ Peter Abbink,¹ David M. Ng'ang'a,¹ Michael S. Seaman,¹ Christy L. Lavine,¹ James R. Perry,¹ Wenjun Li,⁴ Arnaud D. Colantonio,⁵ Mark G. Lewis,⁶ Bing Chen,⁷ Holger Wenschuh,⁸ Ulf Reimer,⁸ Michael Piatak,⁹† Jeffrey D. Lifson,⁹ Scott A. Handley,¹⁰ Herbert W. Virgin,¹⁰ Marguerite Koutsoukos,¹¹ Clarisse Lorin,¹¹ Gerald Voss,¹¹ Mo Weijtens,¹² Maria G. Pau,¹² Hanneke Schuitemaker¹²

Preclinical studies of viral vector–based HIV-1 vaccine candidates have previously shown partial protection against neutralization-resistant virus challenges in rhesus monkeys. In this study, we evaluated the protective efficacy of adenovirus serotype 26 (Ad26) vector priming followed by purified envelope (Env) glycoprotein boosting. Rhesus monkeys primed with Ad26 vectors expressing SIVsmE543 Env, Gag, and Pol and boosted with AS01B-adjuvanted SIVmac32H Env gp140 demonstrated complete protection in 50% of vaccinated animals against a series of repeated, heterologous, intrarectal SIVmac251 challenges that infected all controls. Protective efficacy correlated with the functionality of Env-specific antibody responses. Comparable protection was also observed with a similar Ad/Env vaccine against repeated, heterologous, intrarectal SHIV-SF162P3 challenges. These data demonstrate robust protection by Ad/Env vaccines against acquisition of neutralization-resistant virus challenges in rhesus monkeys.

Despite the urgent need for a safe and effective global HIV-1 vaccine, only four vaccine concepts have been evaluated for protective efficacy in humans during more than 30 years (1, 2). In rhesus monkeys, vaccine protection has been reported against neutralization-sensitive viruses (3), but these data failed to predict protective efficacy in humans (4), which suggests the importance of using neutralization-resistant virus challenges for preclinical evaluation of HIV-1 and SIV vaccine candidates. We previously showed that priming with adenovirus vectors and boosting with poxvirus vectors expressing Env, Gag, and Pol resulted in a reduced per-exposure acquisition risk after challenges with neutralization-resistant SIVmac251, but the majority of these animals

were infected at the end of the challenge series (5, 6). To augment antibody responses, we evaluated the immunogenicity and protective efficacy of priming with adenovirus vectors and boosting with adjuvanted Env gp140 protein against SIVmac251 and SHIV-SF162P3 challenges in rhesus monkeys.

We immunized 32 adult rhesus monkeys (*Macaca mulatta*) that did not express the protective major histocompatibility complex class I alleles *Mamu-A*01*, *Mamu-B*08*, or *Mamu-B*17* with adenovirus serotype 26 (Ad26) vectors (7) expressing SIVsmE543 Env/Gag/Pol antigens (5) followed by either SIVmac32H Env gp140 protein (8) (Ad/Env; $n = 12$) or Ad35 vectors (9) expressing SIVsmE543 Env/Gag/Pol antigens (Ad Alone; $n = 12$), and a control group received

sham vaccines (Sham; $n = 8$). Animals in the Ad/Env group were primed with 3×10^{10} viral particles (vp) Ad26-Env/Gag/Pol vectors (10^{10} vp per vector) by the intramuscular route at weeks 0 and 24 and were boosted with 0.25 mg Env gp140 with AS01B Adjuvant System at weeks 52, 56, and 60. Animals in the Ad Alone group were primed with 3×10^{10} vp Ad26-Env/Gag/Pol vectors at weeks 0 and 24 and were boosted with 3×10^{10} vp Ad35-Env/Gag/Pol at week 52. One control animal died before challenge for reasons unrelated to the study protocol and was excluded from the analysis.

Binding antibody responses to heterologous SIVmac239 Env gp140 were detected by enzyme-linked immunosorbent assay (ELISA) (10) in all vaccinated animals after Ad26 priming at weeks 4 and 28 (Fig. 1A). In the Ad/Env group, ELISA end-point titers increased from 5.3 logs at week 28 to 6.4 logs after the SIV Env gp140 boosts at week 64 ($P < 0.0001$) (Fig. 1A), which confirmed that the Env boost effectively augmented Ad26-primed antibody responses. Neutralizing antibody (NAb) responses assessed in the TZM-bl cell line (11) against tier 1 heterologous SIVmac251_TCLA15 and homologous SIVsmE660 CP3C-P-A8 viruses also increased significantly after SIV Env gp140 boosting (fig. S1). NAb responses against tier 2 viruses were borderline (fig. S1).

In addition to neutralization, antibodies mediate a wide variety of additional antiviral functions through their ability to interact with Fc receptors, complement, and lectin-like proteins

¹Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02215, USA. ²Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Cambridge, MA 02139, USA. ³Thayer School of Engineering at Dartmouth, Hanover, NH 03755, USA. ⁴University of Massachusetts Medical School, Worcester, MA 01605, USA. ⁵New England Primate Research Center, Southborough, MA 01772, USA. ⁶Bioqual, Rockville, MD 20852, USA. ⁷Children's Hospital, Boston, MA 02115, USA. ⁸JPT Peptide Technologies GmbH, 12489 Berlin, Germany. ⁹AIDS and Cancer Virus Program, Leidos Biomedical Research, Frederick National Laboratory, Frederick, MD 21702, USA. ¹⁰Washington University School of Medicine, St. Louis, MO 63110, USA. ¹¹GSK Vaccines, 1330 Rixensart, Belgium. ¹²Janssen Infectious Diseases and Vaccines (formerly Crucell), 2301 Leiden, Netherlands
†Deceased. *Corresponding author. E-mail: dbarouch@bidmc.harvard.edu

(12, 13). Previous studies showed that antibody-dependent cellular phagocytosis (ADCP) (14) and antibody-dependent complement deposition (ADCD) responses correlated with protective efficacy in rhesus monkeys (6). To perform a comprehensive analysis of vaccine-elicited antibody responses, we evaluated 150 independent antibody Fc parameters by high-throughput antibody profiling, including multiple assessments of antibody Fc functionality [ADCP, ADCD, antibody-dependent cell-mediated cytotoxicity (ADCC), antibody-dependent NK cell expression of CD107a, interferon- γ (IFN- γ), and the chemokine CCL4], isotypes, glycosylation, complement binding, and Fc receptor binding (14–18). Integration of all 3600 data points in a principal-component analysis demonstrated that the Ad/Env vaccine and the Ad Alone vaccine elicited Env-specific antibodies that were phenotypically distinct ($P < 0.0001$) (Fig. 1B). A loadings plot (Fig. 1C) showed the distribution of all measured Fc features in the same multidimensional space, which demonstrated the specific features that drove the separation of antibody profiles (red arrows). Partial least-squares discriminant analysis (19) revealed that the six antibody Fc functions described above nearly completely sep-

arated these groups, with the majority of antibody Fc effector functions clustering with the Ad/Env-vaccinated animals (Fig. 1D). Univariate analyses showed that these antibody Fc functions were all significantly increased in Ad/Env group as compared with the Ad Alone group (Fig. 1E), and a combined analysis demonstrated that the number of antibody Fc functions was significantly greater in Ad/Env-vaccinated animals as compared with animals vaccinated with Ad Alone (Fig. 1, F and G). These data show that the protein boost resulted in a more polyfunctional antibody Fc effector profile.

Cellular immune responses measured by IFN- γ enzyme-linked immunospot (ELISPOT) assays in response to heterologous SIVmac239 and homologous SIVsmE543 Env/Gag/Pol peptide pools were also detected in all animals after vaccination (fig. S2). By multiparameter intracellular cytokine staining assays, SIV Env gp140 boosting primarily expanded Env-specific IFN- γ ⁺CD4⁺ T lymphocyte responses in the Ad/Env group, whereas Ad35-Env/Gag/Pol boosting substantially expanded IFN- γ ⁺CD8⁺ T lymphocyte responses in the Ad Alone group (fig. S2). Both CD28⁺CD95⁺ central and transitional memory and CD28⁺CD95⁺ effector

memory CD4⁺ and CD8⁺ T lymphocyte responses (20, 21) were elicited by both vaccines (fig. S3).

To evaluate the protective efficacy of these vaccine regimens, all animals were challenged with six repeated, intrarectal inoculations with 500 tissue culture infectious doses (TCID₅₀) of the heterologous, neutralization-resistant virus SIVmac251 (5, 22, 23) beginning week 96 (Fig. 2, A and B). All control animals were infected by this challenge protocol. The Ad Alone vaccine regimen resulted in a 75% reduction in the per-exposure acquisition risk as compared with controls [1 – (hazard ratio); $P = 0.039$, Cox proportional hazard model], which is consistent with our prior studies (5). In contrast, the Ad/Env vaccine regimen afforded a 90% reduction in the per-exposure acquisition risk as compared with controls ($P = 0.001$). Moreover, 50% (6 of 12) of animals in this group also appeared uninfected at the end of this challenge protocol ($P = 0.012$ compared with controls, chi-square test; $P = 0.044$, Fisher's exact test) (Fig. 2B). Protection in the Ad/Env group was greater than that in the Ad Alone group ($P = 0.042$, chi-square test; $P = 0.097$, Fisher's exact test). Binding antibody titers [$P < 0.0001$; correlation coefficient (R) = 0.75] and antibody Fc polyfunctionality

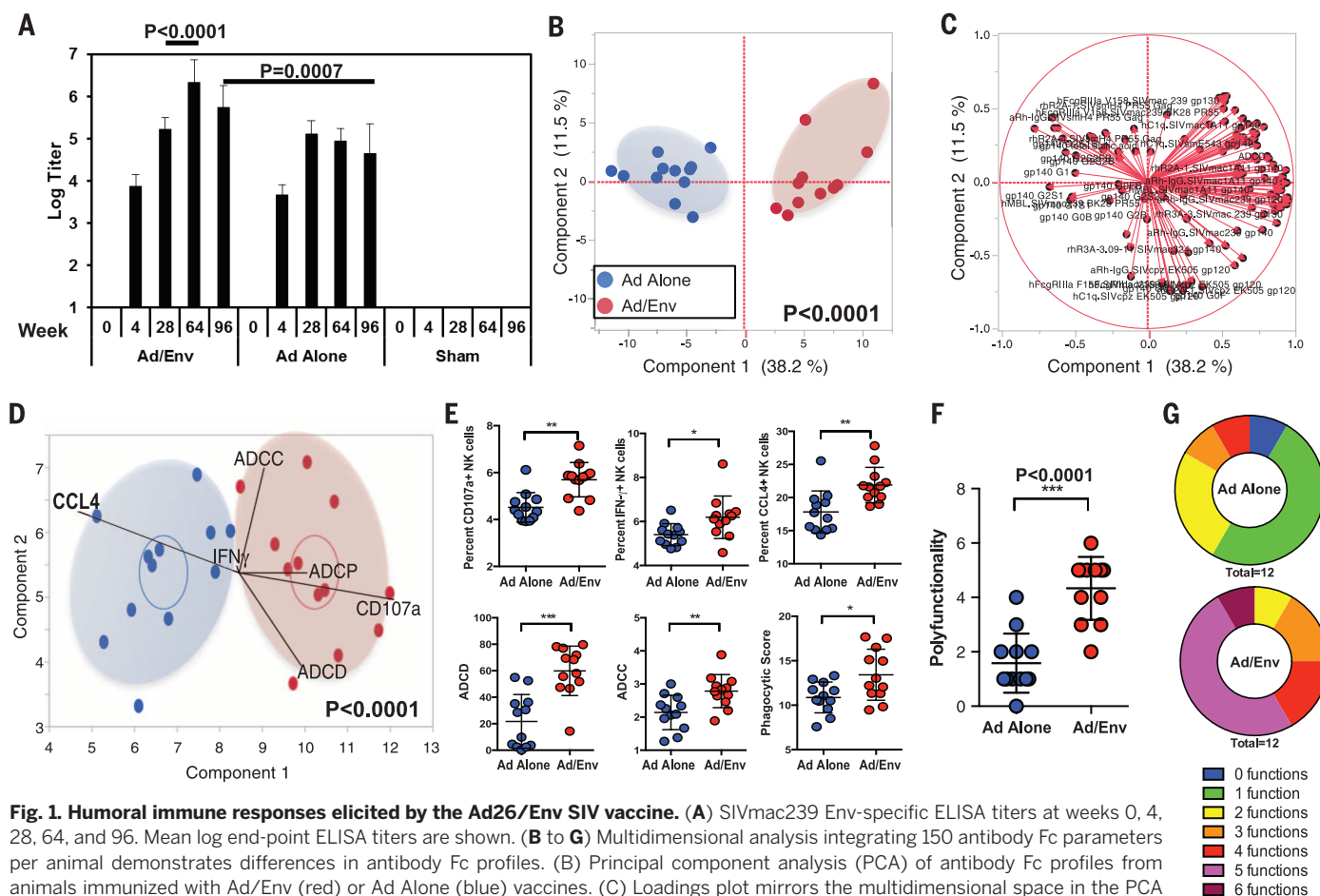


Fig. 1. Humoral immune responses elicited by the Ad26/Env SIV vaccine. (A) SIVmac239 Env-specific ELISA titers at weeks 0, 4, 28, 64, and 96. Mean log end-point ELISA titers are shown. (B to G) Multidimensional analysis integrating 150 antibody Fc parameters per animal demonstrates differences in antibody Fc profiles. (B) Principal component analysis (PCA) of antibody Fc profiles from animals immunized with Ad/Env (red) or Ad Alone (blue) vaccines. (C) Loadings plot mirrors the multidimensional space in the PCA but shows the distribution of all measured Fc features, which demonstrates the features that drove the separation of antibody profiles (red arrows). (D) Partial least-squares discriminant analysis (PLSDA) of antibody profiles from Ad/Env and Ad Alone vaccinees. (E) Univariate analyses of antibody functions identified in (D). A composite (F) dot plot and (G) pie chart show the overall functionality of antibody responses elicited by the Ad/Env and the Ad Alone vaccines. Error bars reflect SEM. P values reflect Mann-Whitney tests. * $P < 0.01$, ** $P < 0.001$, *** $P < 0.0001$.

($P = 0.004$; $R = 0.56$) best correlated with protection against acquisition of infection, as measured by the number of challenges required for infection (Fig. 2C). Individual antibody functions (ADCP, ADCC, CD107, CCL4) also correlated with protection ($P < 0.05$).

In the Ad/Env group, plasma viral loads were persistently negative in the six protected monkeys for 400 days after challenge (Fig. 2D). Of the six infected animals in this group, four animals developed measureable chronic set-point viremia, whereas two monkeys exhibited transient acute viremia and subsequently became elite controllers with undetectable plasma viral loads (Fig. 2D and fig. S4). In the Ad Alone group, plasma viral loads were persistently negative in 2 of 12 monkeys, and chronic viremia developed in 10 of 12 animals. In contrast, all sham controls developed high levels of chronic viremia with a median set-point viral load from days 100 to 400 after infection of 6.03 log copies/ml, which was at least 1.65 logs as high as the median set-point viral load in the animals in the Ad/Env group that became infected ($P = 0.035$) (fig. S5).

We previously reported that progressive SIV infection correlated with a marked expansion of the enteric virome in rhesus monkeys, particularly for picornavirus reads (24–27). Metagenomics sequencing of stool samples in the present study demonstrated that the enteric virome expanded by week 28 but not by week 10 in the sham controls ($P = 0.015$) (fig. S6). Both the Ad/Env and the Ad Alone vaccines reduced the expansion of total enteric reads including enteric picornaviruses ($P = 0.002$ and $P = 0.042$, respectively) (fig. S6). The Ad/Env vaccine also reduced AIDS-related mortality as compared with the sham controls ($P = 0.020$) (fig. S7).

We next investigated whether the vaccinated animals that exhibited persistently negative plasma viral loads were completely protected by comprehensive tissue analyses, adoptive transfer studies, and immunologic assays. We performed necropsies on the six protected animals in the Ad/Env group, the two protected animals in the Ad Alone group, and one of the elite controllers in the Ad/Env group at ~400 days after challenge (Fig. 2D). All of these animals had negative plasma

viral loads at the time of necropsy. We assessed 36 gastrointestinal, lymphoid, and reproductive tract tissues per animal (28 tissues in males) by ultrasensitive nested quantitative polymerase chain reaction or quantitative reverse transcription polymerase chain reaction assays for SIV DNA and SIV RNA as previously described (28). Viral DNA and RNA were readily detectable in all tissues in the elite controller (Fig. 3A, red circles) but not in the eight protected animals (Fig. 3A, black circles), except for one viral signal in a single animal, which is within the range of expected background false-positive signals in similar analyses of naïve animals (28).

We next performed adoptive transfer studies and infused 60 million peripheral blood and lymph node mononuclear cells by the intravenous route from the eight apparently protected animals and the two elite controllers into naïve rhesus monkey hosts. Cells from the elite controllers readily transferred infection and resulted in plasma viral loads of 6.87 to 7.12 log copies/ml in naïve recipients by day 14 after adoptive transfer (Fig. 3B, red lines). In contrast, cells from the

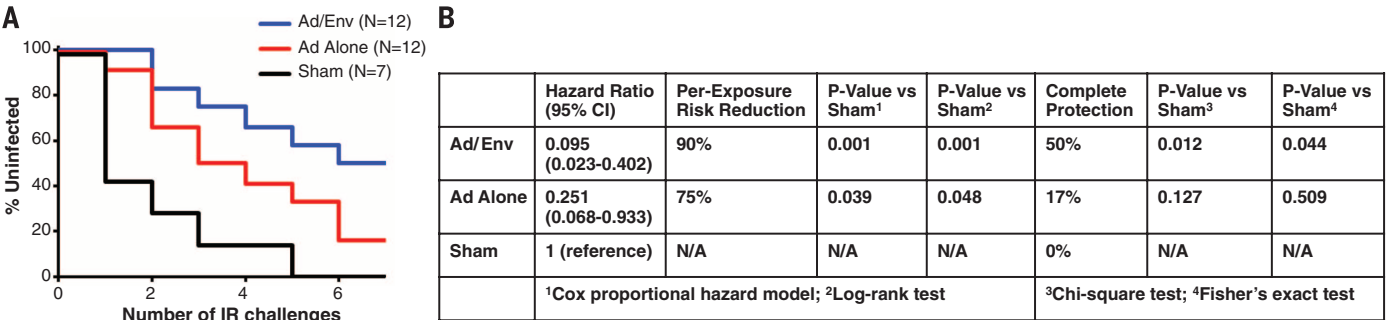
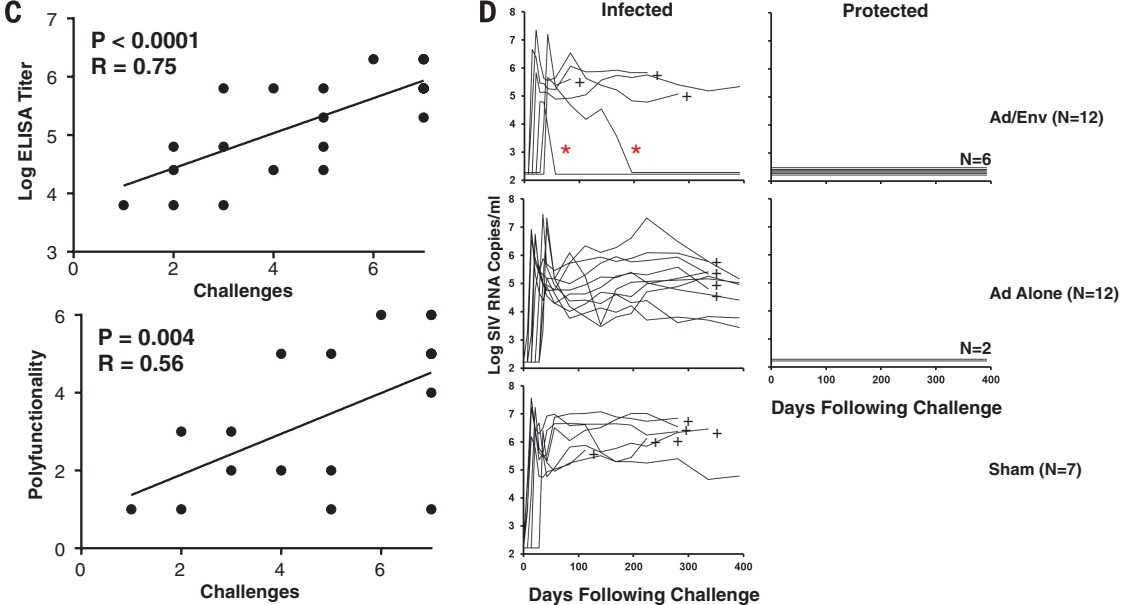


Fig. 2. Protective efficacy of the Ad26/Env SIV vaccine against repeated, intrarectal SIVmac251 challenges. (A) Number of challenges required for acquisition of infection in each vaccine group. (B) Statistical analyses include the hazard ratio with 95% confidence interval and the per-exposure reduction of acquisition risk in each group, with P values reflecting Cox proportional hazard models and log-rank tests. Additional statistical analyses include the percentage of completely protected animals at the end of the challenge series, with P values reflecting chi-square tests and Fisher's exact tests. (C) Correlation of log ELISA titers and antibody Fc polyfunctionality at week 64 with the number of challenges required to establish infection. The plotted data reflect only vaccinated animals and do not include the sham controls. Values plotted as >6 challenges reflect animals that remained uninfected. Overlapping data points are shown as a single symbol. P values reflect Spearman rank correlation tests. (D) Plasma SIV RNA copies/ml over time in infected and protected animals in each vaccine group. Red asterisks indicate elite controllers. + indicates mortality.



of log ELISA titers and antibody Fc polyfunctionality at week 64 with the number of challenges required to establish infection. The plotted data reflect only vaccinated animals and do not include the sham controls. Values plotted as >6 challenges reflect animals that remained uninfected. Overlapping data points are shown as a single symbol. P values reflect Spearman rank correlation tests. (D) Plasma SIV RNA copies/ml over time in infected and protected animals in each vaccine group. Red asterisks indicate elite controllers. + indicates mortality.

protected animals failed to transfer infection (Fig. 3B, black lines).

Furthermore, the protected animals exhibited no increase in Env/Gag/Pol-specific cellular immune responses after challenge and also no responses to Vif, which was not included in the vaccine, whereas vaccinated animals that became infected developed massive anamnestic Env/Gag/Pol-specific cellular immune responses and primary Vif-specific responses (fig. S8). The protected animals also exhibited no anamnestic Env-specific ELISA antibody responses after challenge. Taken together (Fig. 3 and fig. S8), these data strongly suggest that the Ad/Env vaccine afforded complete sterilizing protection in 50% of animals against the SIVmac251 challenge protocol and that the mechanism of protection involved primary blocking of acquisition of infection.

To confirm these findings with analogous vaccines expressing HIV-1 immunogens, we used a

group of 20 rhesus monkeys that had been immunized previously at weeks 0 and 40 with Ad26 and Ad5HVR48 vectors expressing mosaic, consensus, or natural clade C HIV-1 Env/Gag/Pol immunogens (29). Two years after Ad priming, these animals were boosted six times with 0.25 mg HIV-1 clade C C97ZA012 Env gp140 (10, 30) with the AS01B Adjuvant System at weeks 156, 160, 164, 176, 180, and 184 (Ad/Env; $n = 20$). A second group of animals received only 0.25 mg Env gp140 with AS01B at the same six time points (Env Alone; $n = 8$), and a third control group received sham vaccines (Sham; $n = 12$). The Ad/Env vaccine elicited greater antibody responses than did the Env Alone vaccine by ELISA (fig. S9); functional nonneutralizing antibody assays (6, 14, 15) (fig. S9); tier 1 NAb TZM-bl assays (fig. S10); tier 2 NAb A3R5 assays (fig. S11); and linear peptide microarray assays (31, 32), including variable region 2-specific responses (33, 34) (fig. S12). Pro-

tection efficacy was assessed by six intrarectal challenges with 500 TCID₅₀ of the heterologous, neutralization-resistant virus SHIV-SF162P3 (6) beginning week 196 (Fig. 4, A and B). Although the Env Alone vaccine afforded only minimal protection, 40% (8 of 20) of Ad/Env-vaccinated animals were completely protected against this challenge series ($P = 0.006$ compared with controls, chi-square test; $P = 0.014$, Fisher's exact test) (Fig. 4B and figs. S13 to S15). Binding antibody titers ($P = 0.008$) and ADCP responses ($P = 0.001$) correlated with protection against acquisition of infection.

Our data demonstrate the protective efficacy of Ad/Env vaccine regimens against SIVmac251 and SHIV-SF162P3 challenges in rhesus monkeys and suggest that the Env protein boost improved protective efficacy by enhancing the functionality of vaccine-elicited, Env-specific antibody responses. In contrast, DNA/Ad5 vaccines afforded

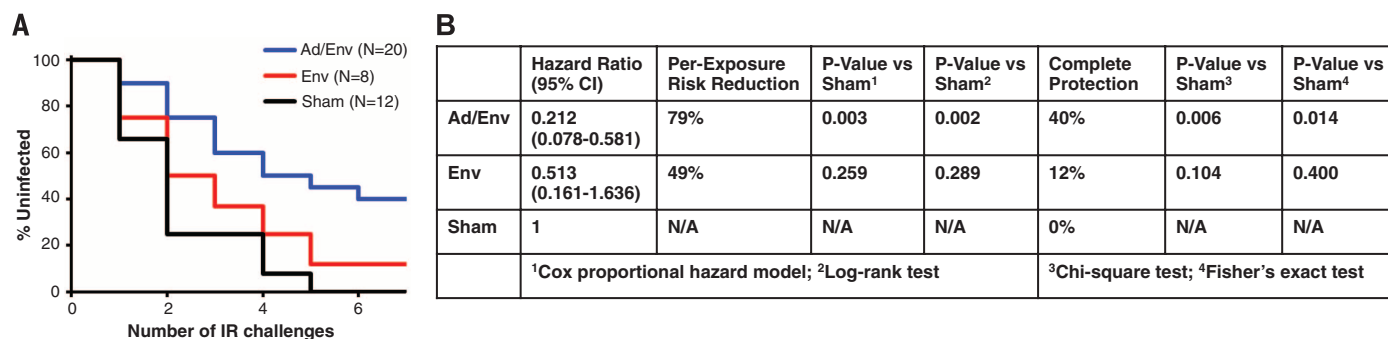
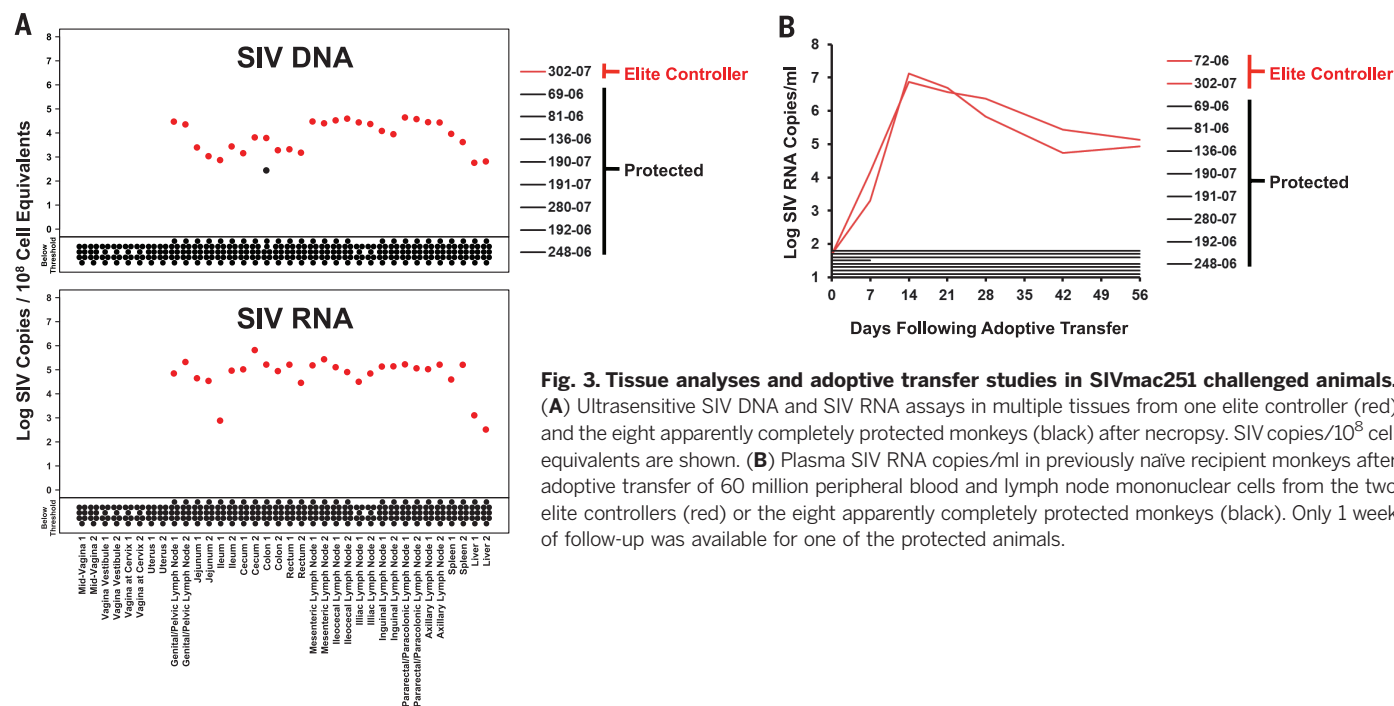


Fig. 4. Protective efficacy of the Ad/Env HIV-1 vaccine against repeated, intrarectal SHIV-SF162P3 challenges. (A) Number of challenges required for acquisition of infection in each vaccine group. (B) Statistical analyses include the hazard ratio with 95% confidence interval and the per exposure reduction of acquisition risk in each group, with P values reflecting Cox proportional hazard models and log-rank tests. Additional statistical analyses include the percentage of completely protected animals at the end of the challenge series, with P values reflecting chi-square tests and Fisher's exact tests.

no protection against SIVmac251 challenges (3), which reflects the ability of DNA/Ad5 vaccines to block only neutralization-sensitive virus clones (35). Alphavirus vector priming and Env protein boosting afforded partial protection against the neutralization-sensitive virus SHIV-SF162P4 but was not evaluated against neutralization-resistant viruses (36). Rhesus cytomegalovirus (CMV) vectors failed to block acquisition of infection but afforded post-infection virologic control and eventual viral clearance in about half of the animals after SIVmac239 challenges (28, 37).

The protective efficacy of Ad/Env vaccines against acquisition of neutralization-resistant virus challenges in rhesus monkeys in the present study has important implications for HIV-1 vaccine development and suggests the potential of Env protein boosting after Ad vector priming. Nevertheless, important differences exist between SIV/SHIV infection in rhesus monkeys and HIV-1 infection in humans. Clinical efficacy studies are therefore required to determine the protective efficacy of these HIV-1 vaccine candidates in humans.

REFERENCES AND NOTES

1. A. S. Fauci, H. D. Marston, *N. Engl. J. Med.* **370**, 495–498 (2014).
2. D. H. Barouch, *N. Engl. J. Med.* **369**, 2073–2076 (2013).
3. N. L. Letvin et al., *Sci. Transl. Med.* **3**, 81ra36 (2011).
4. S. M. Hammer et al., *N. Engl. J. Med.* **369**, 2083–2092 (2013).
5. D. H. Barouch et al., *Nature* **482**, 89–93 (2012).
6. D. H. Barouch et al., *Cell* **155**, 531–539 (2013).
7. P. Abbink et al., *J. Virol.* **81**, 4654–4663 (2007).
8. B. Chen et al., *J. Biol. Chem.* **275**, 34946–34953 (2000).
9. R. Vogels et al., *J. Virol.* **77**, 8263–8271 (2003).
10. J. P. Nkolola et al., *J. Virol.* **84**, 3270–3279 (2010).
11. D. Montefiori, *Curr. Protoc. Immunol. Chap. 12* (Unit 12.11), 1–15 (2005).
12. A. W. Chung, G. Alter, *Future Virol* **9**, 397–414 (2014).
13. M. E. Ackerman, G. Alter, *Curr. HIV Res.* **11**, 365–377 (2013).
14. M. E. Ackerman et al., *J. Immunol. Methods* **366**, 8–19 (2011).
15. V. R. Gómez-Román et al., *J. Immunol. Methods* **308**, 53–67 (2006).
16. A. W. Chung et al., *Sci. Transl. Med.* **6**, 228ra38 (2014).
17. E. P. Brown et al., *J. Immunol. Methods* **386**, 117–123 (2012).
18. A. W. Boesch et al., *MAbs* **6**, 915–927 (2014).
19. K. S. Lau et al., *Sci. Signal.* **4**, ra16 (2011).
20. L. J. Picker et al., *J. Clin. Invest.* **116**, 1514–1524 (2006).
21. H. Li et al., *J. Virol.* **85**, 11007–11015 (2011).
22. J. Liu et al., *Nature* **457**, 87–91 (2009).
23. J. Liu et al., *J. Virol.* **84**, 10406–10412 (2010).
24. S. A. Handley et al., *Cell* **151**, 253–266 (2012).
25. E. Aronesty, in *code.google.com/p/ea-utils*. (ea-utils, Durham, NC, 2011).
26. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, *Bioinformatics* **28**, 3150–3152 (2012).
27. D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, S. C. Schuster, *Genome Res.* **21**, 1552–1560 (2011).
28. S. G. Hansen et al., *Nature* **502**, 100–104 (2013).
29. D. H. Barouch et al., *Nat. Med.* **16**, 319–323 (2010).
30. J. M. Kovacs et al., *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12111–12116 (2012).
31. A. Masch, J. Zerweck, U. Reimer, H. Wenschuh, M. Schutkowski, *Methods Mol. Biol.* **669**, 161–172 (2010).
32. K. E. Stephenson et al., *J. Immunol. Methods* **416**, 105–123 (2015).
33. S. Rerks-Ngarm et al., *N. Engl. J. Med.* **361**, 2209–2220 (2009).
34. B. F. Haynes et al., *N. Engl. J. Med.* **366**, 1275–1286 (2012).
35. M. Roederer et al., *Nature* **505**, 502–508 (2014).
36. S. W. Barnett et al., *J. Virol.* **84**, 5975–5985 (2010).
37. S. G. Hansen et al., *Nature* **473**, 523–527 (2011).

ACKNOWLEDGMENTS

We thank M. Pensiero, M. Marovich, M. Beck, J. Kramer, S. Westmoreland, P. Johnson, W. Wagner, J. Yalley, C. Gittens, C. Cosgrove, M. Kumar, J. Schmitz, H. Peng, J. Hendriks, D. van Manen, W. Bosche, V. Cyril, Y. Li, F. Stephens, R. Hamel, K. Kelly, and L. Dunne for generous advice, assistance, and reagents. The SIVmac239 peptides were obtained from the NIH AIDS Research and Reference Reagent Program. The data presented in this paper are tabulated in the main paper and in the supplementary materials. The authors declare no competing financial interests. D.H.B. is a named co-inventor on vector, antigen, and protein patents (PCT/EP2007/052463, PCT/US2009/060494, PCT/US2009/064999). Correspondence and requests for materials should be addressed to D.H.B. (dbarouch@bidmc.harvard.edu). Vectors, antigens, proteins,

adjuvants, and viruses are subject to Material Transfer Agreements. We acknowledge support from the NIH (AI060354, AI078526, AI080289, AI084794, AI095985, AI096040, AI102660, AI102691, OD011170, and HHSN261200800001E), the Bill and Melinda Gates Foundation (OPP1032817), and the Ragon Institute of MGH, MIT, and Harvard.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/320/suppl/DC1
Materials and Methods

Figs. S1 to S15
References

20 April 2015; accepted 17 June 2015

Published online 2 July 2015;

10.1126/science.aab3886

CIRCADIAN RHYTHMS

A protein fold switch joins the circadian oscillator to clock output in cyanobacteria

Yong-Gang Chang,¹ Susan E. Cohen,² Connie Phong,³ William K. Myers,⁴ Yong-Ick Kim,² Roger Tseng,^{1,5} Jenny Lin,³ Li Zhang,¹ Joseph S. Boyd,² Yvonne Lee,⁶ Shannon Kang,⁶ David Lee,⁷ Sheng Li,⁷ R. David Britt,⁴ Michael J. Rust,³ Susan S. Golden,^{2,6} Andy LiWang^{1,2,5,8,9,*}

Organisms are adapted to the relentless cycles of day and night, because they evolved timekeeping systems called circadian clocks, which regulate biological activities with ~24-hour rhythms. The clock of cyanobacteria is driven by a three-protein oscillator composed of KaiA, KaiB, and KaiC, which together generate a circadian rhythm of KaiC phosphorylation. We show that KaiB flips between two distinct three-dimensional folds, and its rare transition to an active state provides a time delay that is required to match the timing of the oscillator to that of Earth's rotation. Once KaiB switches folds, it binds phosphorylated KaiC and captures KaiA, which initiates a phase transition of the circadian cycle, and it regulates components of the clock-output pathway, which provides the link that joins the timekeeping and signaling functions of the oscillator.

Endogenous circadian (~24-hour) rhythms are found in diverse organisms, arising as an adaptation to Earth's persistent cycles of night and day (1). To uncover the molecular mechanism of a circadian clock, we chose the cyanobacterial system because its oscillator can be reconstituted in vitro (2). The oscillator is composed of only three proteins KaiA, KaiB, and KaiC (3), which together generate a circadian rhythm of KaiC phosphorylation at residues serine 431 (S431) and threonine 432 (T432) in the CII domain (4, 5). KaiA promotes

KaiC (auto)phosphorylation during the subjective day (4, 6), whereas KaiB provides negative feedback to inhibit KaiA (7, 8) and promotes KaiC (auto)dephosphorylation during the subjective night. KaiB is also involved in regulating two antagonistic clock-output proteins—SasA (9) and CikA (10), which reciprocally control the master regulator of transcription, RpaA (11).

To determine the structure of KaiB in its KaiC-bound state, we used a monomeric variant of the KaiB-binding domain of KaiC, CI*, and a dimeric KaiB variant (12), KaiB*, with enhanced KaiC binding. Dimeric forms of free KaiB retain the same tertiary structure in crystals as tetrameric forms (13). Free KaiB has been shown by x-ray crystallography (14) to adopt a fold found in no other protein (15), despite clear sequence similarity with the thioredoxin-like fold at the N terminus of SasA, N-SasA (9). For structural studies, we used proteins from *Thermosynechococcus elongatus* (denoted by ^{te}), because they are more stable than those from *Synechococcus elongatus* (16). For functional studies, we used proteins from *S. elongatus* (denoted by ^{se}), the standard model for investigating in vivo

¹School of Natural Sciences, University of California, Merced, CA 95343, USA. ²Center for Circadian Biology, University of California, San Diego, La Jolla, CA 92093, USA. ³Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, IL 60637, USA. ⁴Department of Chemistry, University of California, Davis, CA 95616, USA. ⁵Quantitative and Systems Biology, University of California, Merced, CA 95343, USA. ⁶Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, USA. ⁷Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA. ⁸Chemistry and Chemical Biology, University of California, Merced, CA 95343, USA. ⁹Health Sciences Research Institute, University of California, Merced, CA 95343, USA.

*Corresponding author. E-mail: aliwang@ucmerced.edu

circadian rhythms (17). Analytical ultracentrifugation experiments indicated that KaiB^{te*} binds to Cl^{te*} as a monomer with a stoichiometric ratio of 1:1 (fig. S1A). Secondary chemical shifts of backbone resonances (18) of KaiB^{te*} in a complex with Cl^{te*} (fig. S1) revealed a thioredoxin-like secondary structure ($\beta\alpha\beta\alpha\beta\alpha$) (19), rather than the secondary structure of free KaiB ($\beta\alpha\beta\alpha\alpha\beta$) found in protein crystals (Fig. 1A). Hereafter, we refer to the $\beta\alpha\beta\alpha\alpha\beta$ form of KaiB as the ground state (gsKaiB), and the $\beta\alpha\beta\alpha\beta\alpha$ state as fold-switched (fsKaiB). Fewer than 10 proteins are known to switch reversibly between distinct folds under native conditions, and they are collectively known as metamorphic proteins (20). KaiB is the only metamorphic protein known to function in biological clocks.

Along a β strand, side chains typically alternate $\uparrow\downarrow\uparrow\downarrow\ldots$. In the β_4 strand of gsKaiB, the side chain pattern is $\uparrow\downarrow\uparrow\downarrow$, where the dash is G89; in fsKaiB, G89 lies in the α_3 helix. We reasoned that a G89A substitution would destabilize β_4 in gsKaiB but not α_3 in fsKaiB (Fig. 1B). A D91R substitution should also destabilize gsKaiB. Nuclear magnetic resonance secondary chemical-shift analysis revealed that, unlike KaiB^{te*}, the two single-point mutants had populations of both gsKaiB and fsKaiB states, but the double mutant was $\geq 98\%$ in the fsKaiB state (figs. S2 to S6). A structural model of G89A,D91R-KaiB^{te*} determined by CS-Rosetta (21), using chemical shifts and backbone amide $^1\text{H}^N$ - $^1\text{H}^N$ nuclear Overhauser effects as restraints, confirmed that G89A,D91R-KaiB^{te*} adopted a thioredoxin-like fold (fig. S7), similar to that of N-SasA (22).

The corresponding KaiB variants in *S. elongatus*, G88A-KaiB^{se}, D90R-KaiB^{se}, and G88A,D90R-KaiB^{se} also promoted the fsKaiB state relative to wild-type (WT) KaiB^{se} (figs. S8 and S9), although to a lesser extent than in the corresponding *T. elongatus* variants. G88A,D90R-KaiB^{se}

formed a complex with Cl^{se*}, with near-complete binding within 5 min (fig. S10). In contrast, WT KaiB^{se} bound Cl^{se*} marginally, even after 24 hours (fig. S11). In vitro oscillation assays showed that the KaiB^{se} variants disrupted KaiC^{se} phosphorylation rhythms (Fig. 2A and fig. S12). Amounts of KaiC^{se} phosphorylation were larger in the presence of D90R-KaiB^{se} or G88A,D90R-KaiB^{se} than they were with G88A-KaiB^{se}. G88A-KaiB^{se} formed a complex with KaiA^{se} (Fig. 2B), whereas the two KaiB^{se} mutants containing D90R did not (fig. S13), which indicated that although the D90R mutation promotes the fsKaiB state, it also disrupts binding. Larger amounts of unsequestered KaiA would be expected to lead to larger amounts of KaiC phosphorylation, which would account for the observed differences in the KaiC phosphorylation profiles when the D90R mutants were used. In vivo bioluminescence rhythms from cyanobacterial luciferase reporter strains harboring *kaiB^{se}* variants (Fig. 2C) were also disrupted, with phenotypes that closely matched the in vitro phosphorylation patterns (Fig. 2A). A chromosomal copy of *kaiB⁺* in addition to the mutant *kaiB* variants did not restore bioluminescence rhythms in vivo, which indicated a dominant-negative effect of the fold-switch mutations (Fig. 2D and fig. S14). KaiB fold switch-stabilizing mutants are less abundant in vivo than WT KaiB (fig. S15), so at equilibrium in vivo fsKaiB is probably rare.

Although KaiB^{se} variants disrupted rhythms in vitro and in vivo, each of them restored the cell-length phenotype in *kaiB⁺* strains (Fig. 2, E and F) in which the SasA-RpaA output pathway was hyperstimulated and cell division was inhibited (23). These functional variants indicate that fsKaiB regulates the clock-output enzymes SasA (9) and CikA (10, 23). SasA is activated when it binds KaiC (24), and SasA and KaiB compete for the CI domain of KaiC

(25). Preincubation with the D90R-KaiB^{se} and G88A,D90R-KaiB^{se} variants inhibited the ability of S431E-KaiC^{se} (a KaiC variant that mimics phosphorylation at the S431 residue) to trigger output signaling through SasA^{se} (Fig. 2G and fig. S16). Also, mixtures of KaiB and KaiC activate the phosphatase activity of CikA that dephosphorylates RpaA (26). Relative to WT KaiB^{se}, fsKaiB^{se} variants enhanced CikA^{se} phosphatase activity by about threefold in vitro (Fig. 2H and fig. S17) and suppressed RpaA^{se} phosphorylation in vivo (fig. S18). We propose that fsKaiB forms a complex with KaiC that both activates signaling through CikA and inhibits signaling through SasA by outcompeting SasA for binding to KaiC. It is likely that fsKaiB interacts with the CikA pseudo-receiver domain (PsR-CikA), because adding PsR-CikA to the in vitro oscillator shortened the period and reduced the amplitude (Fig. 2I and fig. S19). Overexpression of just the PsR-CikA domain in cyanobacteria similarly shortened the period of bioluminescence rhythms (27). Interaction with PsR-CikA was detected for G88A,D90R-KaiB^{se}, but not KaiB^{se} (fig. S20). Addition of the PsR domain of KaiA did not affect phosphorylation rhythms (Fig. 2J), even though it has the same tertiary structure as PsR-CikA (6, 28).

fsKaiB/gsKaiB equilibrium constants, $K = k_{B+}/k_{B-}$, were estimated by fitting the kinetics of binding of KaiB^{te} variants to Cl^{te*} (fig. S21). Equilibrium constants were larger for G89A-KaiB^{te} ($K = 0.13 \pm 0.02$), D91R-KaiB^{te} (1.2 ± 0.1), and G89A,D91R-KaiB^{te} (6.7 ± 0.4), relative to KaiB^{te} (0.08 ± 0.01). For KaiB^{te} variants binding S431E-KaiC^{te}, a distinct fast-binding phase was followed by slow binding (Fig. 3A). These multiphase binding kinetics were reproduced in a gsKaiB \leftrightarrow fsKaiB fold-switching model (Fig. 3B), in which KaiB was initially at equilibrium between the two folds. The pool of fsKaiB bound

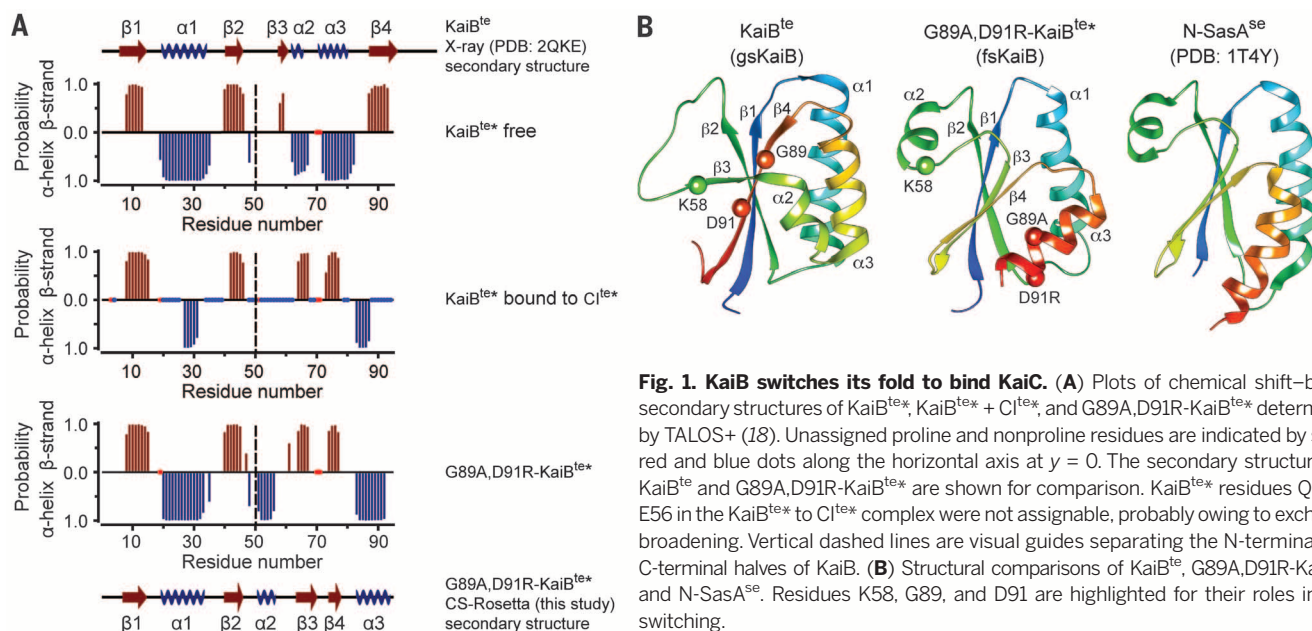


Fig. 1. KaiB switches its fold to bind KaiC. (A) Plots of chemical shift-based secondary structures of KaiB^{te*}, KaiB^{te*} + Cl^{te*}, and G89A,D91R-KaiB^{te*} determined by TALOS+ (18). Unassigned proline and nonproline residues are indicated by small red and blue dots along the horizontal axis at $y = 0$. The secondary structures of KaiB^{te} and G89A,D91R-KaiB^{te*} are shown for comparison. KaiB^{te*} residues Q52 to E56 in the KaiB^{te*} to Cl^{te*} complex were not assignable, probably owing to exchange broadening. Vertical dashed lines are visual guides separating the N-terminal and C-terminal halves of KaiB. (B) Structural comparisons of KaiB^{te}, G89A,D91R-KaiB^{te*}, and N-SasA^{se}. Residues K58, G89, and D91 are highlighted for their roles in fold switching.

rapidly upon adding S431E-KaiC (Fig. 3C), followed by a slow $\text{gsKaiB} \rightarrow \text{fsKaiB}$ population shift. As shown by our computational model (Fig. 3B and fig. S22), the slow formation of the KaiB-KaiC complex contributes to the delay that allows a population of KaiC proteins to become highly phosphorylated under continued stimulation by KaiA. Increasing the rate of KaiB-KaiC binding in our model causes the *in silico* phosphorylation rhythm to fail (Fig. 3D), similar to that observed *in vitro*. Comparing the kinetics of binding to CI (fig. S21) and full-length KaiC (Fig. 3A) by KaiB mutants shows that full-length KaiC contributes to the slow phase as well, and

this is likely due to the CI adenosine triphosphatase (29) exposing the KaiB-binding site upon its activation by CI-CII ring stacking (12).

N-SasA, which also adopts a thioredoxin-like fold (22), competes with KaiB for binding KaiC (25) (fig. S23). Intermolecular distances for the complexes $\text{CI}^{\text{te*}}\text{-G89A,D91R-KaiB}^{\text{te*}}$ and $\text{CI}^{\text{te*}}\text{-N-SasA}^{\text{te}}$ (figs. S24 and S25) were measured by using the pulsed electron paramagnetic resonance method of double electron-electron resonance (DEER), in combination with mutagenesis studies (figs. S26 to S29). The higher-quality DEER data for the $\text{CI}^{\text{te*}}\text{-N-SasA}^{\text{te}}$ complex allowed structural modeling (Fig. 4A and figs. S30 and S31).

In this model, the $\alpha 2$ helix of N-SasA binds to the B-loop region of the CI domain (residues A109 to D123), with additional interactions involving the CI α -helix that follows. Hydrogen-deuterium exchange mass spectrometry (HDX-MS) data on both complexes (Fig. 4, A and B, and figs. S32 to S35) also suggest that this region of KaiC is a common binding site for KaiB and SasA. This model is consistent with a report that truncation of the B loop abolished binding (25). B-loop truncation also restored the cell-length phenotype of a strain that lacked *kaiB* (fig. S36). Furthermore, a F121A mutation in the B loop of full-length KaiC (F121A-S431E-KaiC^{se}) abolished KaiC binding

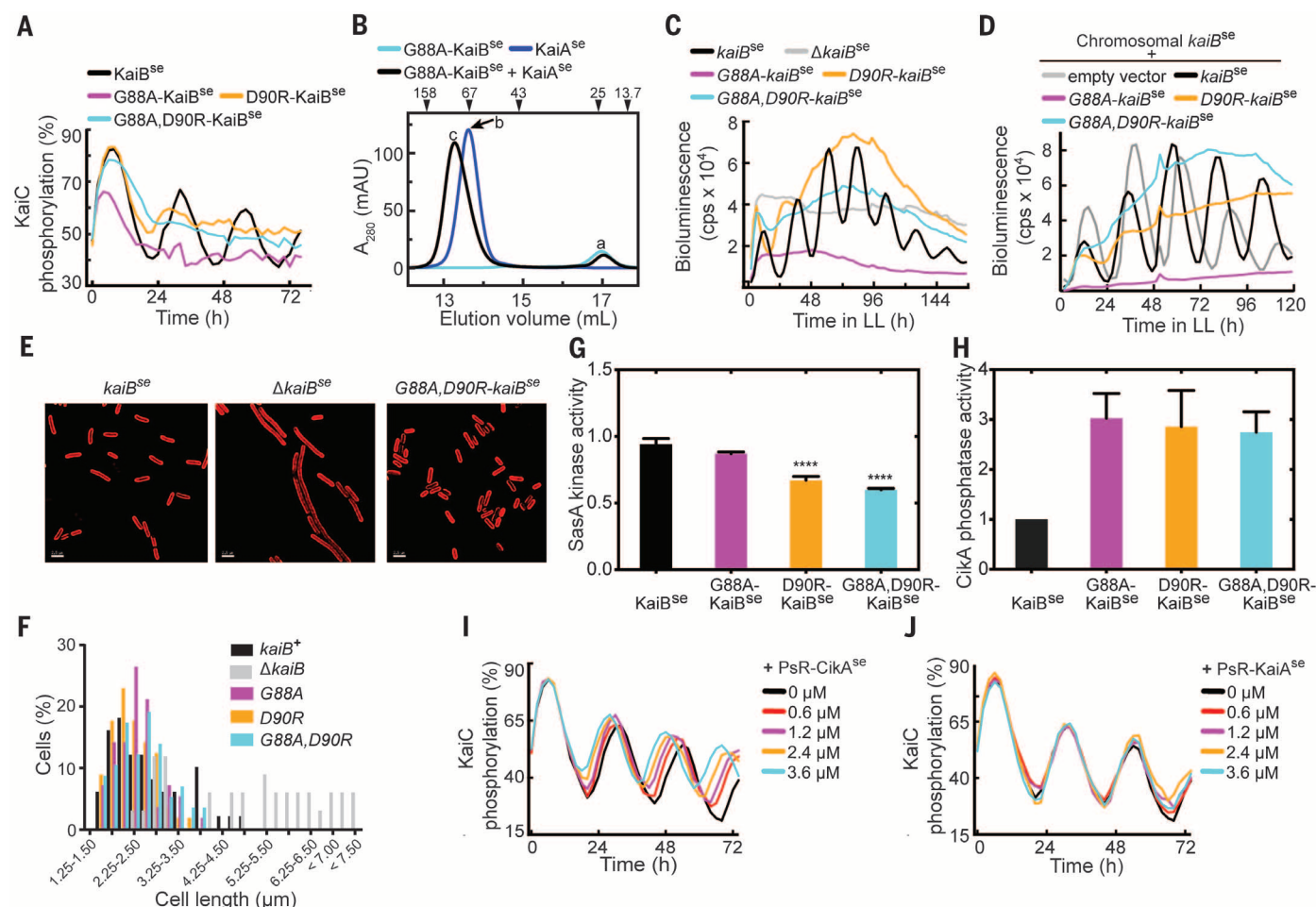


Fig. 2. KaiB fold switching regulates oscillator function and clock output.

(A) *In vitro* KaiC phosphorylation assays using KaiC^{se}, KaiA^{se}, and KaiB^{se}, G88A-KaiB^{se}, D90R-KaiB^{se}, or G88A,D90R-KaiB^{se}. (B) Gel-filtration profiles of G88A-KaiB^{se}, KaiA^{se}, and G88A-KaiB^{se} + KaiA^{se}. Peaks (a) to (c) were analyzed by SDS-PAGE (fig. S13). (C) Bioluminescence from strains that carry a P_{kaiB} *luc* reporter for circadian rhythmicity. Cells harbored *kaiB*^{se}, *G88A-kaiB*^{se}, *D90R-kaiB*^{se}, or *G88A,D90R-kaiB*^{se}, or cells with *kaiB*^{se} deletion. Time in LL, time under low-light conditions. (D) Bioluminescence from strains that carry a P_{kaiB} *luc* reporter expressing *kaiB*^{se}, *G88A-kaiB*^{se}, *D90R-kaiB*^{se}, *G88A,D90R-kaiB*^{se}, or empty vector, in addition to chromosomal *kaiB*^{se}. (E) Representative micrographs of cells expressing *kaiB*^{se}, lacking *kaiB*^{se}, or harboring *G88A,D90R-kaiB*^{se}. Cellular autofluorescence in red. Scale bars, 2.5 μm . (F) Histograms showing cell-length distributions of strains expressing *kaiB*^{se}, Δ *kaiB*^{se}, *G88A-kaiB*^{se}, *D90R-kaiB*^{se}, or *G88A,D90R-kaiB*^{se} as the only copy of *kaiB*. (G) SasA

kinase activities in the presence of S431E-KaiC^{se} and KaiB^{se}, G88A-KaiB^{se}, D90R-KaiB^{se}, or G88A,D90R-KaiB^{se}. The mixtures were incubated for 2 hours before SasA^{se}, RpaA^{se} and [γ -³²P]ATP (labeled adenosine triphosphate) were added. Relative kinase activities compare the mean steady-state amount of ³²P-labeled RpaA^{se} to that of a reaction of S431E-KaiC^{se} alone ($n = 4$, error bars denote SEM). One-way analysis of variance (ANOVA) gives $P < 0.001$, and **** denotes Bonferroni-corrected values ($P < 0.001$) for pairwise comparisons against kinase activity with KaiB^{se} ($\alpha = 0.05$). (H) CikA phosphatase activity toward phosphorylated RpaA in the presence of KaiC^{se} and KaiB^{se}, G88A-KaiB^{se}, D90R-KaiB^{se}, or G88A,D90R-KaiB^{se} ($n = 4$ –5, error bars denote SEM). KaiC^{se} alone or KaiB^{se} alone did not activate CikA phosphatase activity (fig. S17). (I) *In vitro* KaiC^{se} phosphorylation assays as a function of concentration of PsR-CikA^{se}. (J) Same as (I) except for using PsR-KaiA^{se} instead of PsR-CikA^{se}. The black curves in (I) and (J) are identical.

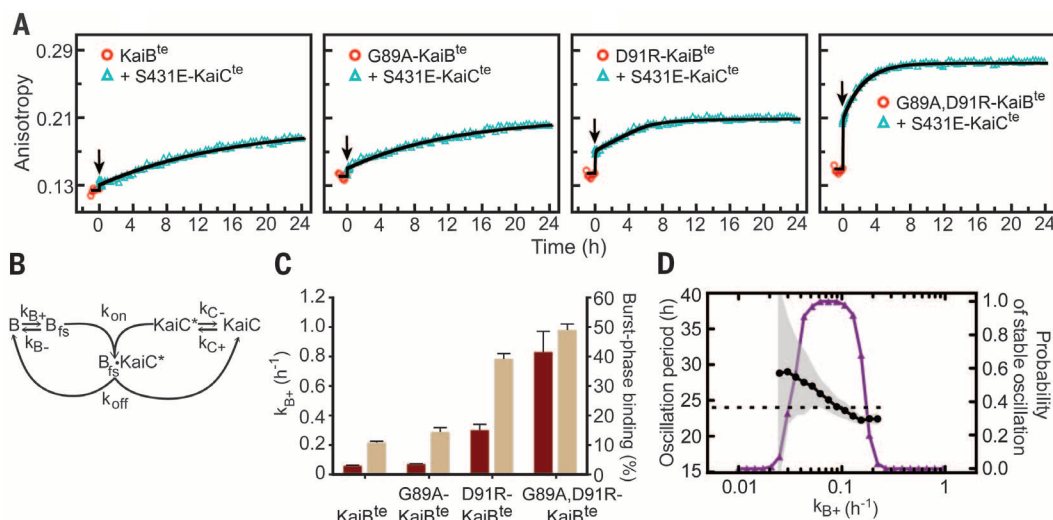
to full-length SasA^{sc}, KaiB^{sc}, and G88A,D90R-KaiB^{sc} (figs. S37 to S40). HDX-MS data (Fig. 4, A and B) indicated that N-SasA and KaiB induce long-range perturbations in isolated CI domains that, in the context of full-length KaiC, may affect

intersubunit interactions. Another HDX-MS-derived model (30) indicated that KaiB most likely binds to CII, on the basis of a calculation minimizing the docking energy between KaiC and gsKaiB, not fsKaiB.

Naturally occurring variations at residue positions 58, 89, and 91 of KaiB (fig. S41A) suggest that the ancestral KaiB protein had the fsKaiB thioredoxin-like fold, and the gsKaiB fold evolved later with the circadian clock. In support of

Fig. 3. KaiB fold-switching regulates slow formation of the KaiB-KaiC complex. (A)

Fluorescence anisotropies of 6-iodoacetamidofluorescein (6-IAF)-labeled KaiB^{te}, G89A-KaiB^{te}, D91R-KaiB^{te}, and G89A,D91R-KaiB^{te} in the presence of S431E-KaiC^{te}. KaiB samples were incubated for 1 hour (circles) before addition (arrow) of S431E-KaiC^{te}. A54C mutation was introduced to all KaiB for fluorescence labeling. (B) Scheme for modeling. (C) Forward fold-switching rate constants, k_{B+} (maroon), and burst-phase binding to S431E-KaiC^{te} (tan). Burst-phase binding—defined as the percentage of KaiB^{te}-S431E-KaiC^{te} complexes formed at $t = 0.1$ hours in the model relative to steady-state binding at $t = 24$ hours—were derived from fitting data after adding S431E-KaiC^{te} in (A) to the model shown in (B). Burst-phase error bars show the standard deviation from model calculations by bootstrap resampling the raw data ($n = 20$). k_{B+} values used in these fits were predetermined from analysis of the kinetics of binding of KaiB^{te} variants to the isolated CI^{te*} domain (fig. S21), a condition where we assumed the rate-limiting step in complex formation is due



only to KaiB fold switching. Error bars for k_{B+} were estimated by bootstrap resampling the original data set 500 times. (D) Mathematical modeling of KaiC phosphorylation period (black) and probability of stable oscillation (purple), as a function of k_{B+} . The black bars indicate the standard deviation in the model period from 100 oscillator calculations at each value of k_{B+} , with the other parameters randomly varied as described in supplementary materials.

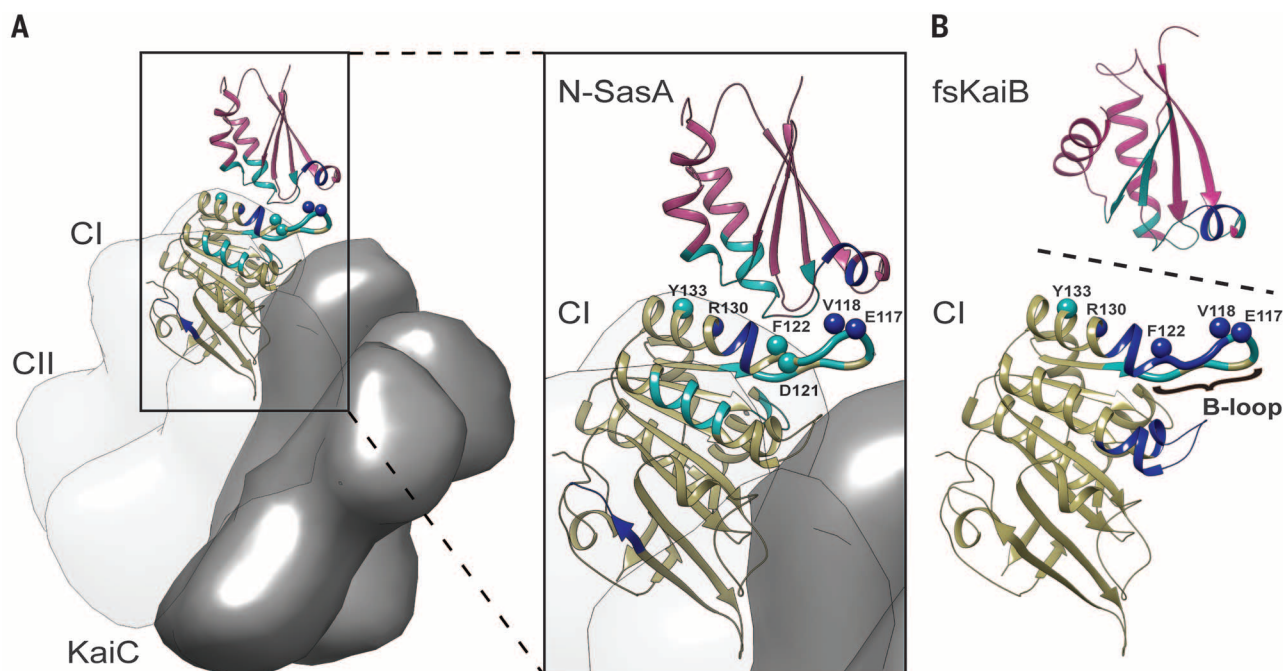
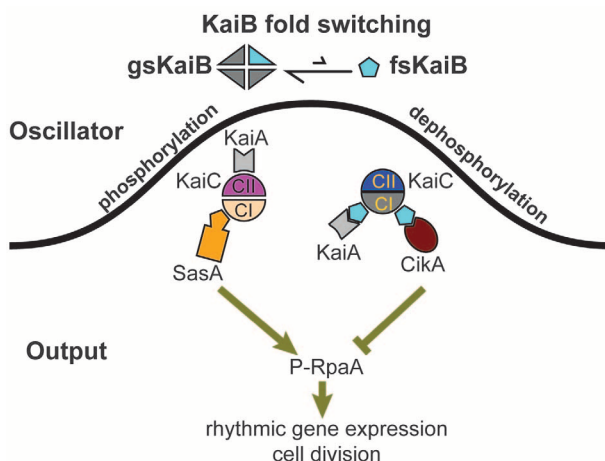


Fig. 4. KaiB and SasA bind to similar sites on CI. (A) An electron paramagnetic resonance)-restrained model of the CI^{te*}-N-SasA^{te} complex. The HADDOCK model of the complex with the best score is superimposed on the crystal structure of KaiC^{te} (PDB ID: 4o0m). (B) Qualitative structural model of the interaction of CI^{te*} and fsKaiB (G89A,D91R-KaiB^{te*}) based on HDX-MS data and mutagenesis. Dark blue and cyan spheres represent CI residues whose mutations strongly or moderately weaken binding, respectively. Dark blue and cyan ribbons represent protection against H/D exchange upon complex formation that are >1.5 and 0.5 to 1.5 standard deviations above the average, respectively, as determined by HDX-MS (figs. S32 to S35).

Fig. 5. Model of KaiB fold switching as linchpin for the cyanobacterial clock.

Excursion of KaiB to the rare fold-switch state causes fsKaiB to displace SasA for binding to KaiC. KaiC-stabilized fsKaiB captures KaiA, initiating the dephosphorylation phase of the cycle. These aspects control oscillator period. CikA and KaiA compete for binding to fsKaiB, which further links oscillator function related to KaiA and output activity via CikA-mediated dephosphorylation of RpaA. The competitive interactions of fsKaiB with SasA, and KaiA with CikA, implicate “output components” CikA and SasA as parts of an extended oscillator.



this notion, a homolog of KaiB from *Legionella pneumophila*, with no known circadian rhythms, has an alanyl residue at position 89 and crystallizes in the fsKaiB fold (31) (fig. S41B). Rare excursions of KaiB between two distinct folds are essential for a robust circadian period and reciprocally regulate mutually antagonistic clock output-signaling pathways (Fig. 5).

REFERENCES AND NOTES

- J. C. Dunlap, *Cell* **96**, 271–290 (1999).
- M. Nakajima et al., *Science* **308**, 414–415 (2005).
- M. Ishiura et al., *Science* **281**, 1519–1523 (1998).
- H. Iwasaki, T. Nishiwaki, Y. Kitayama, M. Nakajima, T. Kondo, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15788–15793 (2002).
- Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
- S. B. Williams, I. Vakonakis, S. S. Golden, A. C. LiWang, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15357–15362 (2002).
- Y. Kitayama, H. Iwasaki, T. Nishiwaki, T. Kondo, *EMBO J.* **22**, 2127–2134 (2003).
- Y. Xu, T. Mori, C. H. Johnson, *EMBO J.* **22**, 2117–2126 (2003).
- H. Iwasaki et al., *Cell* **101**, 223–233 (2000).
- O. Schmitz, M. Katayama, S. B. Williams, T. Kondo, S. S. Golden, *Science* **289**, 765–768 (2000).
- J. S. Markson, J. R. Piechura, A. M. Puszyńska, E. K. O'Shea, *Cell* **155**, 1396–1408 (2013).
- Y. G. Chang, R. Tseng, N. W. Kuo, A. LiWang, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16847–16851 (2012).
- R. G. Garces, N. Wu, W. Gillon, E. F. Pai, *EMBO J.* **23**, 1688–1698 (2004).
- S. A. Villarreal et al., *J. Mol. Biol.* **425**, 3311–3324 (2013).
- L. Holm, P. Rosenström, *Nucleic Acids Res.* **38** (Web Server), W545–W549 (2010).
- I. Vakonakis et al., *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1479–1484 (2004).
- G. Dong, Y. I. Kim, S. S. Golden, *Curr. Opin. Genet. Dev.* **20**, 619–625 (2010).
- Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *J. Biomol. NMR* **44**, 213–223 (2009).
- J. L. Martin, *Structure* **3**, 245–250 (1995).
- A. G. Murzin, *Science* **320**, 1725–1726 (2008).
- S. Raman et al., *Science* **327**, 1014–1018 (2010).
- I. Vakonakis, D. A. Klewer, S. B. Williams, S. S. Golden, A. C. LiWang, *J. Mol. Biol.* **342**, 9–17 (2004).
- G. Dong et al., *Cell* **140**, 529–539 (2010).
- R. M. Smith, S. B. Williams, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8564–8569 (2006).
- R. Tseng et al., *J. Mol. Biol.* **426**, 389–402 (2014).

- A. Gutu, E. K. O'Shea, *Mol. Cell* **50**, 288–294 (2013).
- X. Zhang, G. Dong, S. S. Golden, *Mol. Microbiol.* **60**, 658–668 (2006).
- T. Gao, X. Zhang, N. B. Ileva, S. S. Golden, A. LiWang, *Protein Sci.* **16**, 465–475 (2007).

SEX DETERMINATION

foxl3 is a germ cell-intrinsic factor involved in sperm-egg fate decision in medaka

Toshiya Nishimura,^{1,2} Tetsuya Sato,^{3,4} Yasuhiro Yamamoto,¹ Ikuko Watakabe,¹ Yasuyuki Ohkawa,⁵ Mikita Suyama,^{3,4} Satoru Kobayashi,^{2,6*} Minoru Tanaka^{1,2,†}

Sex determination is an essential step in the commitment of a germ cell to a sperm or egg. However, the intrinsic factors that determine the sexual fate of vertebrate germ cells are unknown. Here, we show that *foxl3*, which is expressed in germ cells but not somatic cells in the gonad, is involved in sperm-egg fate decision in medaka fish. Adult XX medaka with disrupted *foxl3* developed functional sperm in the expanded germinal epithelium of a histologically functional ovary. In chimeric medaka, mutant germ cells initiated spermatogenesis in female wild-type gonad. These results indicate that a germ cell-intrinsic cue for the sperm-egg fate decision is present in medaka and that spermatogenesis can proceed in a female gonadal environment.

In vertebrates, gonadal somatic cells instruct germ cells to adopt their sexual fates. In medaka (*Oryzias latipes*), the expression of *DMY/dmrt1bY* in the supporting somatic cells is critical for the fate decision of germ cells to enter spermatogenesis (1, 2). However, the molecular mechanism underlying the sexual fate decision in germ cells remains unknown. In this study, we show that *foxl3* serves as a germ cell-intrinsic cue for sperm-egg fate decision.

foxl3 is an ancient duplicated copy of *foxl2* that is expressed in gonads of teleost fish (3–5). Although *foxl2* is known to be essential for ovarian development and maintenance (6–9), the function of *foxl3* is not known. *foxl3* transcripts and FOXL3 protein (*foxl3*/FOXL3) were first detectable in germ

- C. Phong, J. S. Markson, C. M. Wilhoite, M. J. Rust, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1124–1129 (2013).
- J. Snijder et al., *Proc. Natl. Acad. Sci. U.S.A.* **111**, 1379–1384 (2014).
- M. Loza-Correa et al., *Environ. Microbiol.* **16**, 359–381 (2014).

ACKNOWLEDGMENTS

We thank R. Greenspan, J.-P. Changeux, M. Paddock, Y. Shen, R. Peterson, S. Chou, and N.-W. Kuo for discussions and A. Chavan for figure preparation. Work was supported by Air Force Office of Scientific Research grant 13RSL012 and NIH grant GM107521 to A.L., a Burroughs-Wellcome Career Award at the Scientific Interface to M.J.R., NIH grants GM100116 and GM062419 to S.S.G. and AI081982 and AI101436 for support of the University of California–San Diego HDX-MS Laboratory, American Cancer Society Postdoctoral Fellowship PF-12-262-01-MPC to S.E.C., and NSF Graduate Research Fellowship to R.T. The data reported in this paper are tabulated in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/324/suppl/DC1
Materials and Methods
Figs. S1 to S41
Tables S1 to S13
References (32–58)

15 August 2014; accepted 8 May 2015
Published online 25 June 2015;
10.1126/science.1260031

cells of both XX and XY embryos at stage 35, the time of onset of gonadal sex differentiation (figs. S1, A to D, and S2, A to D). After this stage in XX

¹Laboratory of Molecular Genetics for Reproduction, National Institute for Basic Biology, Okazaki 444-8787, Japan.

²Graduate University for Advanced Studies (SOKENDAI), Okazaki 444-8585, Japan. ³Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-8582, Japan. ⁴Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency (JST), Fukuoka 812-8582, Japan. ⁵Department of Advanced Medical Initiatives, JST-CREST, Faculty of Medicine, Kyushu University, Fukuoka 812-8582, Japan. ⁶Okazaki Institute for Integrative Bioscience, National Institute for Basic Biology, Okazaki 444-8787, Japan.

*Present address: Life Science Center of Tsukuba Advanced Research Alliance, University of Tsukuba, Tsukuba, Ibaraki 305-8577, Japan. †Corresponding author. E-mail: mtanaka@nibb.ac.jp

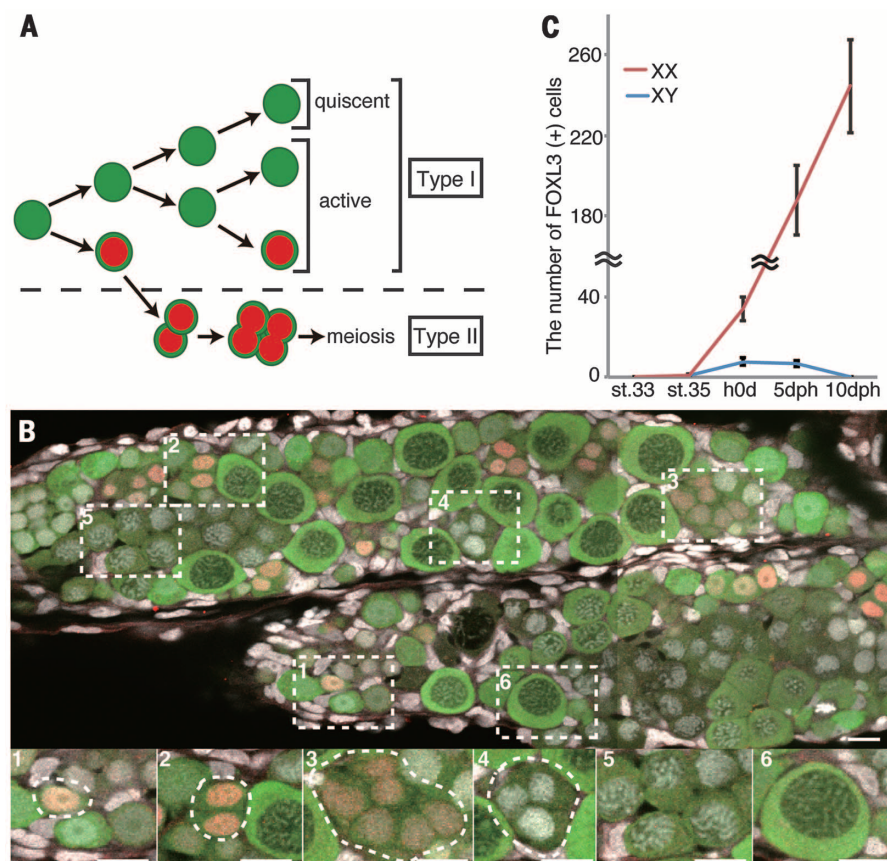


Fig. 1. Dimorphic expression of FOXL3 during gonadal development. (A) Schematic representation of type I and type II division and expression of FOXL3 protein (red) in germ cells (green). “Active” and “quiescent” indicate mitotic states. (B) Ventral view of the gonad, observed with confocal laser microscopy. FOXL3 expression (red) in XX gonad at 7 days after hatching. FOXL3 signals are detected in a subset of type I (1) and all mitotic type II (2, 2-cell cyst; 3, 8-cell cyst) germ cells, but not in meiotic germ cells (4, zygotene; 5, pachytene) or diplotene oocytes (6). (C) The number of FOXL3-positive cells during gonadal development in XX and XY gonads. Values are expressed as means \pm SEM. The number of samples examined at each stage is shown in fig. S2. st., stage; dph, days post-hatching. Scale bars, 10 μ m.

gonads, germ cells can be categorized into two types according to their division (Fig. 1A) (10, 11). Type I is stem-type self-renewal division, in which germ cells divide completely to generate two isolated daughter cells surrounded by supporting cells. Type I includes both mitotically active and quiescent germ cells (12). Type II is cystic division, in which germ cells divide synchronously with intercellular bridges. Type II is gametogenesis-committed division, followed by meiosis and oogenesis. In XY gonads, the transition from type I to type II is suppressed until 1 month after hatching.

foxl3/FOXL3 appeared in a subset of mitotically active type I germ cells, but not in quiescent type I germ cells, at stage 35 onward in both XX and XY gonads (Fig. 1, A and B, and fig. S3). In XX gonads, the signals continued to be detected in type II germ cells but disappeared in meiotic germ cells and oocytes (Fig. 1B and fig. S1, E and G). Whereas *foxl3*/FOXL3 were detected throughout gonadal development in XX fish, they disappeared in all germ cells of XY fish by 10 days after hatching (Fig. 1C and figs. S1 and S2).

To examine the function of *foxl3* in the gonadal sex differentiation, we generated transcription activator-like effector nuclease (TALEN)-induced mutants of *foxl3*. We designed two types of TALENs targeting *foxl3* and obtained two different mutant alleles (NΔ17 and FHΔ8), which contain frameshifts causing premature truncation upstream of and at the forkhead domain of FOXL3, respectively (fig. S4, A to D). Antiserum recognizing the C terminus of FOXL3 protein detected the protein in both heterozygous (+/-) mutants but failed to detect it in both homozygous (-/-) mutants (fig. S4, E to H), indicating that *foxl3* was successfully disrupted by TALEN. One week after hatching, a stage at which oocytes are formed in wild-type XX gonads (Fig. 1B), the *foxl3*^{-/-} XX gonads had no oocytes but were instead filled with cystic and meiotic germ cells (Fig. 2, A to C), suggesting that commitment to type II division and meiosis is not affected by loss

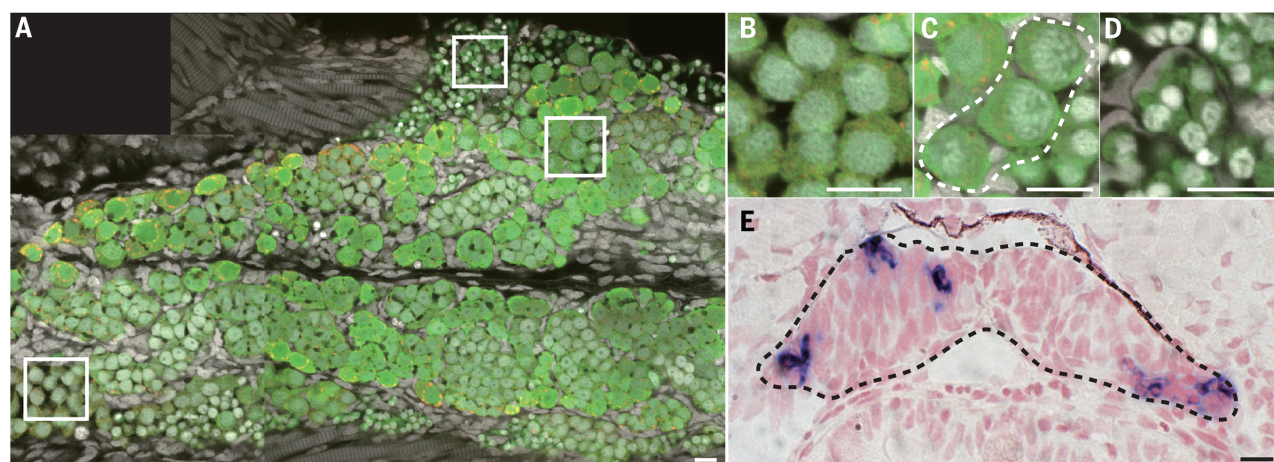


Fig. 2. *foxl3* is responsible for suppressing initiation of spermatogenesis in medaka germ cells. (A to D) Ventral view of a *foxl3*^{-/-} XX gonad at 7 days after hatching. The white boxes are magnified in (B) to (D). Green cells (*olvas*-enhanced green fluorescent protein) represent germ cells. Meiotic germ cells at zygotene (B) and pachytene [(C), dotted line] stages, and spermatid-like cells (D). Red dots in germ cells represent OLVAS protein. (E) Detection of protamine (purple signals) in *foxl3*^{-/-} XX gonad at 10 days after hatching. Scale bars, 10 μ m.

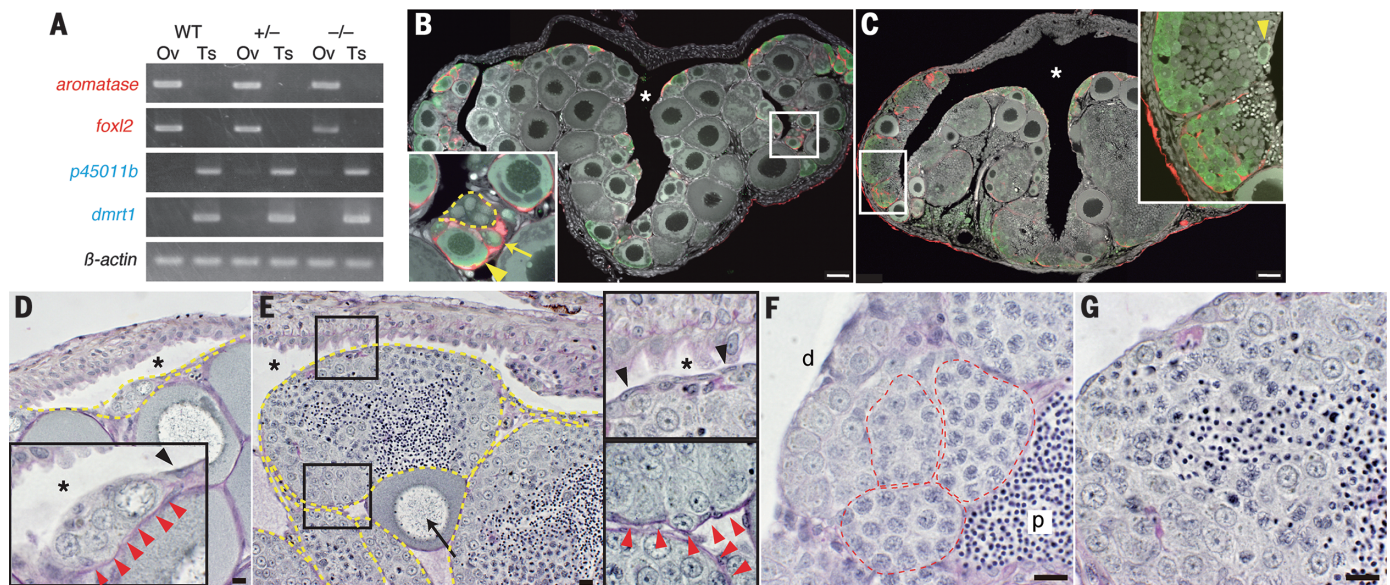
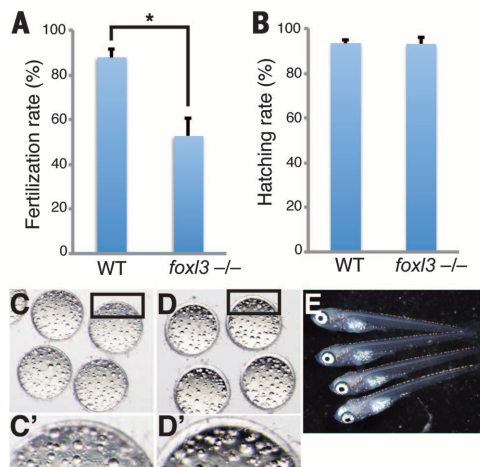


Fig. 3. Progression of spermatogenesis in a female gonadal environment. (A) RT-PCR analysis. Ov, ovary; Ts, testis. (B and C) Cross sections of adult gonads with immunohistochemistry. (B) Wild-type ovary. (Inset) A germinal cradle, in which a germline stem cell (arrow), cystic germ cells (yellow dotted line), and a diplotene oocyte (arrowhead) are surrounded by *sox9b*-expressing cells (red). (C) *foxl3*^{-/-} XX gonad. (Inset) Expanded germinal epithelium in which *sox9b*-expressing cells (red) surround germ cells (green). The yellow arrowhead indicates an oocyte. Scale bars, 50 μ m. (D to G) Cross sections of adult gonads subjected to periodic acid-Schiff staining. (D) Wild-type ovary. (Inset) The germinal cradle between basement mem-

brane (red arrowheads) and epithelial cells (black arrowhead) of the germinal epithelium (yellow dotted lines). (E) *foxl3*^{-/-} XX gonad. Spermatogenesis progresses within the expanded germinal epithelium (yellow dotted lines) that resides between the basement membrane (bottom inset, red arrowheads) and epithelial cells (top inset, black arrowheads). The arrow indicates an oocyte in the stromal compartment. Asterisks in (B to E) indicate ovarian cavities. [(F) and (G)] Spermatogenesis proceeds as units of cysts [(F), red dotted lines] from distal (d) to proximal (p) position in wild-type testis (F), but this arrangement is disorganized in the *foxl3*^{-/-} XX gonad (G). Scale bars, 10 μ m.

Fig. 4. *foxl3*^{-/-} XX gonads produce functional sperm. (A) Fertilization rate of artificial insemination by using sperm from wild-type testes ($n = 3$) and *foxl3*^{-/-} ovaries ($n = 8$). In each artificial insemination, one gonad was used. (B) Hatching rate of the fertilized eggs in (A). (C) Unfertilized eggs. (D) Eggs fertilized by sperm derived from *foxl3*^{-/-} ovaries. The activated egg membrane is shown in (D'). (E) Embryos hatched from the eggs in (D). Statistics by two-tailed Student's *t* test; **P* < 0.05.



of *foxl3*. Instead of oocytes, spermatid-like cells were present at the periphery of gonads in both mutant lines (NΔ17 in Fig. 2, A and D, and FHΔ8 in fig. S4H); these cells expressed *protamine*, a sperm-specific gene (Fig. 2E and fig. S5). In addition, other markers for spermatogenesis such as *kif17*, *tektin-t*, and *shippo1* were also expressed in *foxl3*^{-/-} XX gonads (fig. S5). At around the same stage in normal XY gonads, only type I germ cells are present (fig. S2J), and spermatogenesis does not begin until puberty, which occurs later than 1 month after hatching. This indicates that in *foxl3*^{-/-} XX mutants, spermatogenesis begins much earlier than puberty in wild-type males.

Chimeric analysis revealed that *foxl3*^{-/-} germ cells completed spermatogenesis in wild-type XX gonads, indicating that the phenotype is due to loss of *foxl3* function in germ cells but is not affected by gonadal somatic cells (fig. S6). Furthermore, introduction of a bacterial artificial chromosome (BAC) containing *foxl3* allele restored the formation of oocytes in *foxl3*^{-/-} XX larvae (fig. S7). Collectively, these data indicate that *foxl3* is a germline-intrinsic factor involved in sperm-egg fate decision in medaka.

To characterize the gonadal structures producing sperm in the XX mutants, we performed in situ hybridization and reverse transcription

polymerase chain reaction (RT-PCR) using sex-specific markers. Throughout gonadal development from 10 to 30 days after hatching and in adult gonads of *foxl3*^{-/-} XX fish, the female markers *foxl2* and *aromatase* were expressed in somatic cells surrounding spermatogenic cells, whereas the male markers *dmrt1* and *p45011b* were absent, reflecting the normal female expression pattern in somatic cells (Fig. 3A and figs. S8, A to F, and S9, A to F). In adults, the secondary sex characteristics, which are distinguished by the morphology of dorsal/anal fins and urogenital papilla (13), were of the female type in the XX mutants (fig. S10, A and B). Consistent with this, the *foxl3*^{-/-} XX gonads exhibited morphologically ovarian structures characterized by an ovarian cavity on the dorsal side and a stromal compartment on the ventral side (Fig. 3, B and C, and fig. S10, C and D). In wild-type ovaries, a thin multilayered tissue called the germinal epithelium separates the ovarian cavity from the stromal compartment (fig. S11A). In addition, the germinal cradle, a unit consisting of *sox9b*-expressing cells that harbor germline stem cells and support early oogenesis (Fig. 3B), lies between the epithelial cells of the germinal epithelium and the basement membrane bordering the stromal compartment (Fig. 3D and fig. S11B) (14). In *foxl3*^{-/-} XX gonads, spermatogenesis proceeded within the expanded germinal epithelium (Fig. 3E and fig. S11). Early stages of spermatogenic cells were inclined to localize on the basal side of the expanded germinal epithelium underlain by

the basement membrane and were surrounded by *sox9b*-expressing cells, forming a structure similar to the germinal cradles (Fig. 3C and figs. S11 and S12). A subset of these germ cells expressed *nanos2*, a marker of germline stem cells (14), suggesting that germline stem cells were established in the absence of *foxl3* function (fig. S9, G to I).

In wild-type testes, spermatogenesis proceeds synchronously in cysts composed of Sertoli cells surrounding spermatogenic cells at the same stage. The cysts are arranged along a distal-to-proximal axis, so that the spermatogonia are located most distally and matured spermatids are released into the efferent duct in the most proximal position (Fig. 3F). In the expanded germinal epithelium of *foxl3*^{-/-} XX gonads, however, the distal-to-proximal arrangement of the cysts was disorganized, so that spermatogenic cells at various stages were adjacent to each other (Fig. 3G). This observation suggests that feminized somatic cells are not able to arrange masculinized germ cell cysts in the proper distal-to-proximal manner, as Sertoli cells would otherwise do in wild-type testes. At the inner side of the expanded germinal epithelium, meiotic spermatogenic cells did not always contact *sox9b*-expressing cells (fig. S12). Therefore, it is possible that spermatocytes in *foxl3*^{-/-} XX gonads could differentiate autonomously into sperm within the expanded germinal epithelium, as previously reported occurring in culture (15). Occasionally, oocytes appeared in between spermatogenic cells in the expanded germinal epithelium (Fig. 3C), which subsequently exited as follicles into the stromal compartment (Fig. 3E), as seen in follicle formations of the wild-type ovary (fig. S11) (14). Collectively, the expanded germinal epithelium in *foxl3*^{-/-} XX gonads harbors germline stem cells and is able to support follicle formation but cannot organize the cysts in a distal-proximal direction. Any morphological trait of efferent ducts was not observed, possibly leading to the large accumulation of spermatogenic cells within the germinal epithelium of *foxl3*^{-/-} XX gonads. We conclude that *foxl3*^{-/-} XX mutants undergo spermatogenesis in histologically functional ovaries.

On the other hand, *foxl3*^{-/-} XY adult gonads exhibited morphologically normal testes with

functional sperm (fig. S10, E to H, L, and M), indicating that *foxl3* is dispensable for male development. In addition, chimeric analysis revealed that *foxl3*^{-/-} germ cells did not initiate precocious spermatogenesis in wild-type male gonads (fig. S6, B and D). These results imply that male, but not female, somatic cells are able to properly regulate spermatogenesis in the absence of *foxl3*. The absence of male somatic cell regulation may allow *foxl3*^{-/-} germ cells to complete spermatogenesis much earlier than puberty in wild-type males.

To investigate whether the sperm produced by *foxl3*^{-/-} XX mutants were functional, we performed artificial insemination. The gonads from 2- or 3-month-old XX mutants were minced with forceps to suspend sperm in medaka Ringer's solution, and ovulated eggs of wild-type females were subsequently inseminated by the resultant sperm. Approximately 50% of the eggs were successfully fertilized (Fig. 4A), as confirmed by the activation of the egg membrane (Fig. 4, C and D). The fertilized eggs exhibited normal development, and more than 95% of them hatched (Fig. 4, B and E). Therefore, the *foxl3*^{-/-} XX mutant produced functional sperm. Furthermore, *foxl3*^{-/-} XX mutants spontaneously spawned a few fertile eggs, although their fertility—as assessed by the number of spawned eggs, the fertilization rate, and the hatching rate—was lower than that of the *foxl3*^{+/-} mutants (fig. S10, I to K).

The presence of eggs in *foxl3*^{-/-} XX mutants suggests the existence of a *foxl3*-independent pathway controlling oocyte formation. The oocytes did not appear until around 20 to 30 days after hatching (fig. S8, G to I) in the mutants, a stage at which an ovarian cavity starts to form because of the action of estradiol (E2) (16). However, inhibition of *aromatase* activity and E2 signaling by the administration of fadrozole (FAD) and tamoxifen (TAM), respectively, did not block oocyte formation in the mutants (fig. S13). On the other hand, administration of E2 during the embryonic and larval stages of both XX and XY *foxl3*^{-/-} mutants did not induce the precocious oocyte formation observed in E2-treated *foxl3*^{+/-} XY larva (fig. S14). These findings suggest that after derepression of spermatogenesis upon loss of *foxl3* expression, oocytes are formed in an E2-independent manner.

Our findings suggest that suppression of spermatogenesis by *foxl3* is an important component of female fate decision in germ cells. In this respect, we note that another gene of forkhead box transcriptional factor, *foxl2*, is involved in the maintenance of female fate in the mammalian supporting cells by suppressing the male fate (8).

REFERENCES AND NOTES

1. M. Matsuda et al., *Nature* **417**, 559–563 (2002).
2. I. Nanda et al., *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11778–11783 (2002).
3. M. T. Geraldo, G. T. Valente, A. S. Braz, C. Martins, *Heredity* **111**, 57–65 (2013).
4. B. Crespo, O. Lan-Chow-Wing, A. Rocha, S. Zanuy, A. Gómez, *Gen. Comp. Endocrinol.* **194**, 81–93 (2013).
5. D. Baron et al., *J. Mol. Endocrinol.* **33**, 705–715 (2004).
6. C. Ottolenghi et al., *Hum. Mol. Genet.* **14**, 2053–2062 (2005).
7. D. Schmidt et al., *Development* **131**, 933–942 (2004).
8. N. H. Uhlenhaut et al., *Cell* **139**, 1130–1142 (2009).
9. M.-H. Li et al., *Endocrinology* **154**, 4814–4825 (2013).
10. D. Saito et al., *Dev. Biol.* **310**, 280–290 (2007).
11. T. Nishimura, M. Tanaka, *Sex Dev.* **8**, 252–261 (2014).
12. S. Nakamura et al., *Development* **139**, 2283–2287 (2012).
13. H. Kurokawa et al., *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16958–16963 (2007).
14. S. Nakamura, K. Kobayashi, T. Nishimura, S. Higashijima, M. Tanaka, *Science* **328**, 1561–1563 (2010).
15. A. Saiki et al., *Dev. Growth Differ.* **39**, 337–344 (1997).
16. A. Suzuki, M. Tanaka, N. Shibata, *J. Exp. Zool. A Comp. Exp. Biol.* **301A**, 266–273 (2004).

ACKNOWLEDGMENTS

We thank the National BioResource Project Medaka for supplying cDNAs, BAC, and fosmids. We are grateful to K. Suzuki and C. Kinoshita for fish maintenance. This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas (22132007), a Grant-in-Aid for Scientific Research (A) (25251034), and the Science and Technology Research Promotion Program for Agriculture, Forestry, Fisheries and Food Industry (26047A). T.N. and M.T. are authors on a patent applied for by the National Institute of Natural Sciences (provisional patent application number JP2014-241331) that relates to the function of *foxl3*.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6245/328/suppl/DC1
Materials and Methods
Figs. S1 to S14
References (17–27)

20 November 2014; accepted 4 June 2015
Published online 11 June 2015;
10.1126/science.aaa2657

Will you be meeting a Nobel Prize winner this December?

(If you have a recent PhD you could be.)

Stockholm in the second week of December is a special place. The city is alive with excitement as it welcomes and celebrates the new Nobel Laureates at the annual Nobel Prize ceremony.

If you are a PhD student, you could be here too – meeting a Nobel Laureate and receiving a rather special prize yourself.

The journal *Science* & SciLifeLab have established The *Science* & SciLifeLab Prize for Young Scientists, to recognize and reward excellence in PhD research and support young scientists at the start of their careers. It's about bright minds, bright ideas and bright futures.

Four winners will be selected for this international award. They will have their essays published in the journal *Science* and share a new total of 60,000 USD in prize money. The winners will be awarded in Stockholm, in December, and take part in a unique week of events including meeting leading scientists in their fields.

"The last couple of days have been exhilarating. It has been an experience of a lifetime. Stockholm is a wonderful city and the Award winning ceremony exceeds my wildest dreams."
–Dr. Dan Dominissini, 2014 Prize Winner

Who knows, The *Science* & SciLifeLab Prize for Young Scientists could be a major stepping stone in your career and hopefully one day, during Nobel week, you could be visiting Stockholm in December once again.

The 2015 Prize is now open. The deadline for submissions is August 1, 2015.

Enter today: www.sciencemag.org/scilifelabprize

The 2015 Prize categories are:

- Cell and Molecular Biology
- Ecology and Environment
- Genomics and Proteomics
- Translational Medicine



This prize is made possible with the kind support of the Knut and Alice Wallenberg Foundation. This Foundation grants funding in two main areas; research projects of high scientific potential and individual support of excellent scientists.

"What I do with my Octet HTX time? Climb."

Shave weeks off your lead selection programs.

Broader antibody cross-competition ups your odds of finding the best candidates, but larger epitope binning studies take time. The Octet HTX system lets you use any binning assay format, any size matrix, start a run and get analyzed results the same day or the next day for larger studies. You can also combine multiple experiments into one dataset to easily visualize and cluster antibodies in similar bins or binding groups.

Lucy gets out of the lab more often now to climb.
What will you do with your extra time?



fortéBIO[®]
A Division of **Pall Life Sciences**

fortebio.com | 888-OCTET-75

PALL Life Sciences

Fast. Accurate. EASY.

WILL YOUR RESEARCH LEAD TO BETTER LIVES FOR PATIENTS?



Gopinath Sutendra and Evangelos D. Michelakis, "Pulmonary Arterial Hypertension: Challenges in Translational Research and a Vision for Change", *Sci. Transl. Med.* 5, 208sr5 (2013) Credit: Science Source

Science Translational Medicine |  AAAS
INTEGRATING SCIENCE, ENGINEERING, AND MEDICINE

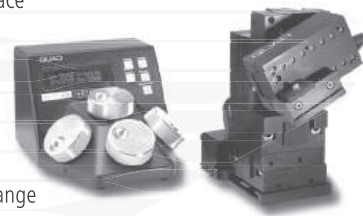
Find out more about the scope of the journal and submit your research today. ScienceTranslationalMedicine.org

QUAD™ Four-Axis Micromanipulator System

Sutter introduces a true fourth axis to move the electrode coaxially at exactly the desired angle of approach and significantly extend the reach of the system.

FEATURES

- Four independent axes – 30mm travel in diagonal for coaxial pipette movement, 25mm travel in X, Y and Z
- Sub-micron resolution
- True diagonal assures coaxial movement
- Suited for *in vivo* electrophysiological recording
- Quiet mode eliminates electrical noise
- User-friendly, fanless compact controller with ROE preserves bench space
- Push button control for multiple functions – Work, Home, Quiet, Pulse and Relative
- Robotic Home and Work position moves for easy automated pipette exchange



SUTTER INSTRUMENT

PHONE: 415.883.0128 | FAX: 415.883.0572
EMAIL: INFO@SUTTER.COM | WWW.SUTTER.COM



AAAS is here – Science Funding, Climate Regulation, Human Rights.

Around the world, governments turn to AAAS as an objective, multidisciplinary scientific authority to educate public officials and judicial figures on today's most pressing issues. And this is just one of the ways that AAAS is committed to advancing science to support a healthy and prosperous world. Join us. Together we can make a difference.

To learn more, visit
aaas.org/plusyou/policy

 AAAS + U = Δ

AAAS 2016
ANNUAL MEETING

WASHINGTON, DC
FEBRUARY 11–15

Global Science Engagement

Dear Colleagues:

On behalf of the AAAS Board of Directors, it is my honor to invite you to join us in Washington, DC for the 2016 AAAS Annual Meeting, February 11–15.

The AAAS Annual Meeting is the most widely recognized global science gathering with cutting-edge scientific sessions, renowned speakers, and valuable networking opportunities.



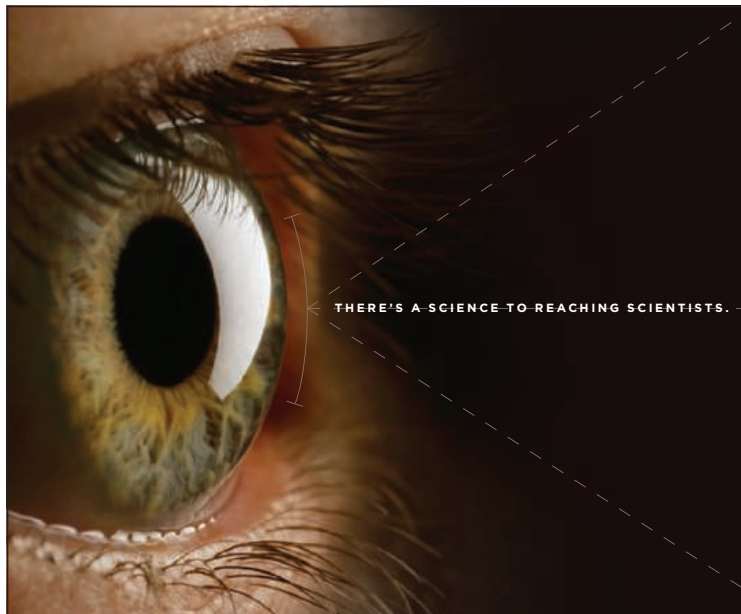
The 2016 meeting theme—Global Science Engagement—focuses on how the scientific enterprise can meet global challenges in need of innovation and international collaboration, such as food and water security, sustainable development, infectious disease and health, climate change, natural disasters, and energy.

We look forward to seeing you in Washington, DC. Registration and housing will open in August.

A handwritten signature in black ink, reading "Geraldine Richmond". The signature is fluid and cursive, with a large, stylized "G" and "R".

Geraldine Richmond
AAAS President
Presidential Chair and Professor of Chemistry
University of Oregon

www.aaas.org/AM16invite



Save these Dates!

Postdoc Careers

August 28, 2015

Reserve ads by August 11 to guarantee space

Faculty Careers

September 18, 2015

Reserve ads by September 1 to guarantee space

THERE'S A SCIENCE TO REACHING SCIENTISTS.

For recruitment in science, there's only one **Science**

Two Fantastic Recruiting Opportunities!

POSTDOC CAREERS | August 28, 2015

Be sure to promote your openings to the thousands of scientists who read *Science* to find out about the latest postdoc opportunities.

Reserve space by August 11, 2015.

FACULTY CAREERS | September 18, 2015

Gear up to recruit for the faculty positions at your university with this much anticipated issue that reaches thousands of Ph.D. scientists looking for positions in academia.

Reserve space by September 1, 2015.



Produced by the *Science*/AAAS Custom Publishing Office.

SCIENCECAREERS.ORG

Science Careers

FROM THE JOURNAL SCIENCE  AAAS

To book your ad: advertise@sciencecareers.org

The Americas
202-326-6582

Europe/RoW
+44(0)1223-326500

Japan
+81-3-3219-5777

China/Korea/Singapore/Taiwan
+86-186-0082-9345

"MOST OF THE PEOPLE THINK THEY'VE REACHED THE END OF EARTH WHEN THEY GET
TO THE REINDEER CAMP. BUT WE GO BEYOND THAT."

Paula T. DePriest

*Lichenologist and Mongolian cultural conservationist,
Paula DePriest, AAAS Member*



Every scientist
has a *story*

Read her story at membercentral.aaas.org, the website
that opens up new worlds. Connect with others who share
your passion.





Multimode Microplate Reader

The new Spark 10M multimode microplate reader is designed to offer greater flexibility and increased productivity for cell biology and genomics customers. From microbiology research and cell-based assays to rapid DNA quantification, the Spark 10M delivers a combination of advanced capabilities and exceptional ease-of-use to simplify your daily work. At the heart of the Spark 10M are Tecan's unique Fusion Optics which offer users the choice of filter- or monochromator-based measurements—or even a combination of both—at the touch of a button, meaning laboratories no longer have to make a trade-off between flexibility and sensitivity. This cutting-edge system is complemented by High Speed Monochromators, which provides a complete absorbance scan, from 200 nm to 1,000 nm, in less than 5 seconds. The Spark 10M reader has been developed from the outset with cell-based assays in mind, and includes a host of software and hardware features designed to simplify cell biology protocols.

Tecan

For info: +41-(0)-44-922-81-11
www.tecan.com

Automated Sample Storage

The Verso is a truly modular automated storage platform that is easily configured to meet the needs of the most demanding sample management applications. The system can be scaled to fit sample capacities ranging from 100,000 to over five million tubes at temperatures from ambient to -20°C. Verso is capable of storing and processing up to 1,500 tubes and plates per hour, and allows loading and unloading of up to 70 sample racks at a time. For a fully automated workflow, Verso can integrate with Hamilton automated liquid handling workstations. With end-users in mind, the intuitive system allows users to easily perform most jobs in three clicks or less. Features include one-touch loading and a job queue manager, which allows users to prioritize jobs and set jobs to process overnight. Every sample is tracked at all times to maintain a complete audit trail, including their temperature logs, who accessed the sample, and what function was completed.

Hamilton Storage Technologies

For info: 800-310-5866
www.hamilton-storage.com



Fluorescence Illumination System

The new Lumen 100-LED illuminator provides advanced, high-quality illumination for a wide range of fluorescence microscopy applications. With a 25,000-hour lifetime, the Lumen 100-LED offers a broad array of LED's covering wavelengths throughout the visible spectrum and is suitable for the majority of fluorophores used in fluorescence applications. A wide range of adapters is available to connect the Lumen 100-LED to most upright and inverted microscopes currently on the market. Directly coupled to the fluorescence port to maximize light efficiency, the unit can be further enhanced by the addition of excitation filters to LEDs to further optimize the bandwidth for your specific application. Engineered to completely eliminate power to the LED when set to zero light level, the Lumen 100 is ideally suited for electrophysiology and optogenetics as well as general fluorescence imaging applications. A combiner is available if two LEDs need to be used simultaneously.

Prior Scientific

For info: 800-877-2234
www.prior.com

Western Blot Validated Antibodies

PrecisionAb Antibodies are a catalog of rigorously tested primary antibodies validated for detection of endogenous proteins via Western blot. Scientists have struggled with antibody unreliability for decades because manufacturing standards and quality control vary widely among vendors, targets, and even among antibody lots. Each PrecisionAb antibody is screened on whole cell lysates from up to 12 different biologically relevant cell lines and only antibodies that detect endogenous proteins with high sensitivity and specificity were chosen for the catalog. Each antibody can be purchased as a trial size and comes with positive control lysates for easy in-lab validation to fit researchers' needs.

Bio-Rad

For info: 800-424-6723
www.bio-rad.com/precisionabpr

Confocal Raman Imaging Module

The newest SWIFT v2 ultrafast confocal Raman imaging module is available for the XploRA PLUS and LabRAM Evolution. The new SWIFT v2 module greatly enhances the ability to obtain fast and detailed confocal Raman images with the click of a button. The module offers a 4x to 5x improvement in speed on even the fastest of Raman images, while maintaining the class leading confocal performance and sensitivity. It makes Raman imaging a realistic alternative to optical imaging techniques and can be used to survey 3-D sample structure and surface features. Its technology and speed will enable the user to focus on detail and the chemical information within an image, and is

not constrained by any compromise in sensitivity or confocal performance, which has previously challenged Raman imaging. The new module offers a powerful yet cost effective route to employing the full power of Raman imaging in the laboratory.

Horiba Scientific

For info: 732-494-8660
www.horiba.com/scientific

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/products/newproducts.dtl 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.

want new technologies?

antibodies

apoptosis

biomarkers

cancer

cytometry

data

diseases

DNA

epigenetics

genomics

immunotherapies

medicine

microbiomics

microfluidics

microscopy

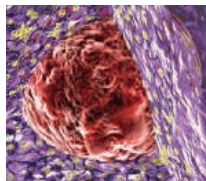
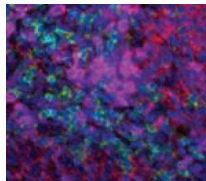
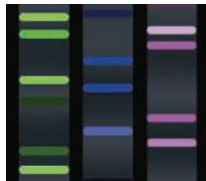
neuroscience

proteomics

sequencing

toxicology

transcriptomics



watch our **webinars**

Learn about the latest breakthroughs, new technologies, and ground-breaking research in a variety of fields. Our expert speakers explain their quality research to you and answer questions submitted by live viewers.

VIEW NOW!

webinar.sciencemag.org

Science
AAAS

Brought to you by the *Science*/AAAS
Custom Publishing Office



@SciMagWebinars

It's easy to find the antibody you need



Come find the antibody that's right for you. We're dedicated to making the experience of finding your antibody simple, fast and reliable. We have over 40,000 antibodies ready to ship to you, and a Web experience that makes it easy to find and order yours. We also offer custom antibody services.

Get your antibody now at
lifetechnologies.com/antibodies

ThermoFisher
SCIENTIFIC



There's only one **Science**

Science Careers Advertising

For full advertising details, go to ScienceCareers.org and click For Employers, or call one of our representatives.

Tracy Holmes

Worldwide Associate Director
Science Careers
Phone: +44 (0) 1223 326525

THE AMERICAS

E-mail: advertise@sciencecareers.org

Fax: +1 (202) 289 6742

Tina Burks

Phone: +1 (202) 326 6577

Nancy Toema

Phone: +1 (202) 326 6578

Online Job Posting Questions

Phone: +1 (202) 312 6375

EUROPE / INDIA / AUSTRALIA / NEW ZEALAND / REST OF WORLD

E-mail: ads@science-int.co.uk

Fax: +44 (0) 1223 326532

Sarah Lelarge

Phone: +44 (0) 1223 326527

Kelly Grace

Phone: +44 (0) 1223 326528

Online Job Posting Questions

Phone: +44 (0) 1223 326528

JAPAN

Katsuyoshi Fukamizu (Tokyo)

E-mail: kfukamizu@aaaas.org

Phone: +81 3 3219 5777

Hiroyuki Mashiki (Kyoto)

E-mail: hmashiki@aaaas.org

Phone: +81 75 823 1109

CHINA / KOREA / SINGAPORE / TAIWAN / THAILAND

Ruolei Wu

Phone: +86 186 0082 9345

E-mail: rwu@aaaas.org

All ads submitted for publication must comply with applicable U.S. and non-U.S. laws. Science reserves the right to refuse any advertisement at its sole discretion for any reason, including without limitation for offensive language or inappropriate content, and all advertising is subject to publisher approval. Science encourages our readers to alert us to any ads that they feel may be discriminatory or offensive.

ScienceCareers

FROM THE JOURNAL SCIENCE **AAAAA**

ScienceCareers.org



College of
Osteopathic
Medicine

FACULTY POSITION: BIOCHEMISTRY/GENETICS

New York Institute of Technology-College of Osteopathic Medicine (NYIT-COM) seeks an **Assistant Professor** with teaching expertise in biochemistry/genetics in the Department of Biomedical Sciences. The area of research focus is open. The Department currently has active research in cardiovascular disease/heart failure, renal physiology and development, neuroscience, and microbiology. Successful candidates are expected to contribute to the medical school teaching effort and participate in an active research program. Collaboration with existing faculty research activities is strongly encouraged. The medical school provides strong institutional support for both teaching and research activities, including maintenance of core facilities, support for student workers, and access to IT resources.

The successful candidate will possess a Ph.D., D.O., M.D., or D.V.M. and 2+ years of post-doctoral research experience and demonstrated successful medical education/teaching experience. We offer institutionally-supported faculty salaries, an extremely competitive benefits package, and a professional environment designed to enhance career development. To apply, please e-mail cover letter and resume to: **Dr. A. Martin Gerdes, Chair of Biomedical Sciences** at agerdes@nyit.edu. NYIT-COM, PO Box 8000 Northern Blvd, Old Westbury, NY 11568-8000.

EOE M/F/D/V.

Advance your career with expert advice from **Science Careers**.

Download Free Career Advice Booklets!
ScienceCareers.org/booklets

ScienceCareers
FROM THE JOURNAL SCIENCE **AAAAA**

POSITIONS OPEN

A **POSTDOCTORAL POSITION** is available at the University of Maryland School of Medicine to model inherited lipid storage diseases using iPSC technology. The candidate should have a Ph.D., experience in Cell/Molecular Biology, excellent oral and written communication skills, and be able to work independently. Experience in hESC/iPSC technology and hematopoietic/neuronal development is desirable. Salary is commensurate with experience. To apply, send curriculum vitae and contact information for three references to **Dr. Ricardo A. Feldman** at e-mail: rfeldman@som.umaryland.edu.

Your career is our cause.

Get help from the experts.

ScienceCareers.org

- Job Postings
- Job Alerts
- Resume/CV Database
- Career Advice
- Career Forum

ScienceCareers

FROM THE JOURNAL SCIENCE **AAAAA**

Post Your Jobs

1 million candidates*
151,000 job applications*



Reach Scientists.
Fill Positions.

*Jan-Dec 2014

ScienceCareers

employers.sciencecareers.org

PLAY A LEADING ROLE IN THE FIGHT AGAINST CANCER

Director of Population and Early Detection Research Competitive + excellent benefits | London

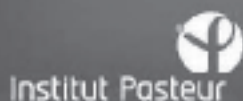
At Cancer Research UK, our goal is to bring forward the day when all cancers are cured. As Director of Population and Early Detection Research, you'll help us realise that vision by leading our research strategy for one of the key things in beating the disease: early detection & diagnosis.

Based in our London head office, you'll drive and implement strategy, generate excitement in the research community, and help build our capacity and expertise – enjoying autonomy and flexibility, as well the right resources, environment and freedom to flourish in your career.

To find out more and apply, visit the jobs section in cancerresearchuk.org

Cancer Research UK is a registered charity in England and Wales (1089464), Scotland (SC041666) and the Isle of Man (1103). Registered company in England and Wales (4325234) and the Isle of Man (5713F).

Let's beat cancer **sooner**.
cruk.org



FOR RESEARCH, FOR HEALTH,
FOR OUR FUTURE

The Institut Pasteur is hiring group leaders for its new Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI)

The new direction of the Institut Pasteur has defined Bioinformatics, Biostatistics and Integrative Biology as strategic priorities. A new center (C3BI) to foster research in these domains was set up in 2014. Substantial resources were allocated for the creation and development of the C3BI, with the recruitment of 40 research engineers in bioinformatics and biostatistics between 2014 and 2017. A building will be renovated on the Paris campus to house the C3BI.

In this context, the Institut Pasteur is looking for several new senior and/or junior group leaders in bioinformatics, biostatistics and integrative biology, with a strong methodological component. The main focus is on computational and statistical analysis of biological "big data", typically produced by new generation sequencing and -omics technologies, but all modeling and computational approaches of biological questions closely connected with Institut Pasteur research areas are eligible. Highly attractive packages to match the experience of the candidate will be provided, including institutional salaries (principal investigator, permanent scientists, secretary and postdoctoral fellows), a substantial contribution to running costs and equipment, as well as support for relocation expenses and administrative issues.

Candidate's profile

Successful candidates will possess the following qualifications:

- PhD with a minimum of 6 years of research experience
- Recognized scientific leadership in bioinformatics/biostatistics
- Broad experience in methodological development for the analysis of various types of data
- Consistent publishing record of cutting edge research as senior/first author
- Significant experience in managing and mentoring scientists, and in designing, financing and executing an innovative research program
- Demonstrated ability to collaborate with experimental and computational biologists

Application web site (forms and documents to be filled and uploaded):

<https://c3bi.pasteur.fr/research-teams-apply/>

Deadline: September 20, 2015 midnight CEST

Further information on the institute and C3BI can be found at <http://www.pasteur.fr> and <https://c3bi.pasteur.fr/>

Short-listed candidates will be invited for interview in October 2015, and results will be announced by mid-December. The new research groups will start by early 2016 (the precise date is negotiable).

Informal inquiries can be addressed to Olivier Gascuel (Head of C3BI – C3BICall2015@pasteur.fr)



Postdoc Careers

August 28, 2015

Reserve space by August 11

THERE'S A SCIENCE TO REACHING SCIENTISTS.



For recruitment in science, there's only one

Science

Post online and your job will be e-mailed to over 13,000 job seekers looking for postdoc positions.

This feature helps postdocs transition from one discipline or department to another. Topics include strategies for making such changes and discussions about the advantages/disadvantages of changing directions.

What makes *Science* the best choice for recruiting?

- Read and respected by 570,400 readers around the globe
- 78% of readers read *Science* more often than any other journal
- Your ad sits on specially labeled pages to draw attention to the ad
- Your ad dollars support AAAS and its programs, which strengthens the global scientific community.

Why choose this postdoc feature for your advertisement?

- Relevant ads in the career section with special postdoc banner
- 67% of our weekly readers are Ph.D.s.

Expand your exposure. Post your print ad online to benefit from:

- Link on the job board homepage directly to postdoc opportunities
- Dedicated landing page for postdoc opportunities.

SCIENCECAREERS.ORG

ScienceCareers

FROM THE JOURNAL SCIENCE AAAS

To book your ad: advertise@sciencecareers.org

The Americas
202-326-6582

Japan
+81-3-3219-5777

Europe/RoW
+44 (0) 1223-326500

China/Korea/Singapore/Taiwan
+86-186-0082-9345



SCHOOL OF MEDICINE
**Faculty Positions in Biochemistry and
Molecular Biophysics**

The Department of Biochemistry and Molecular Biophysics at Washington University School of Medicine invites applications for several tenured or tenure-track faculty positions at the level of Assistant, Associate or Full Professor. Successful candidates will have established a strong record of research. Applicants seeking tenured positions must have a strong record of external funding.

Outstanding individuals working in any area of biochemistry and molecular biophysics are encouraged to apply. The candidate's research should be aimed at addressing fundamental questions related to molecular mechanisms of biological or biomedical relevance. Current research in the department spans a wide range of topics including computational biology, membrane proteins, molecular motors, nucleic acid / protein interactions, protein structure, enzymology and signal transduction. Additional information about the department is available at <http://www.biochem.wustl.edu>. Washington University has a highly interactive research environment with vigorous interdisciplinary graduate and medical scientist training programs. Minority and women scientists are especially encouraged to apply.

Applicants should email their curriculum vitae and a brief description of their research interests to the Search Committee at bmbsearch@biochem.wustl.edu. Applicants should include contact information for three individuals who can write letters of recommendation. The committee will request letters as necessary.

Completed applications will be reviewed on a rolling basis, starting immediately. For full consideration, applications should be received by December 1, 2015.

Washington University is an Equal Opportunity Employer. We are committed to the recruitment of candidates traditionally underrepresented on university faculties. Individuals of any race, ethnicity, gender or sexual orientation are encouraged to apply, as are disabled individuals and veterans. The School of Medicine at Washington University is committed to finding solutions to global health problems, including ones that affect minority and disadvantaged populations.



**INDIANA UNIVERSITY
BLOOMINGTON**

Scientist at the Bloomington Drosophila Stock Center

The Bloomington Drosophila Stock Center at Indiana University Bloomington (<http://flystocks.bio.indiana.edu/>) is seeking a scientist to join its management team. The BDSC is a large, dynamic and heavily used repository for *Drosophila melanogaster* strains that serves researchers worldwide. IUB provides an outstanding intellectual environment for BDSC activities, with more than a dozen research groups using *Drosophila* for studies in molecular and cell biology, development, neurobiology, physiology, evolution and ecology. The work of BDSC scientists is wide ranging and includes management of collection contents (strain accessions, deaccessions, quality control), information management (biocuration, database construction and use, website development), scientific communication (user support, grant proposals and reporting, conference presentations, position papers, manuscripts), and business operations (personnel, user accounts, billing, financial recordkeeping and reporting). The new BDSC scientist will participate in all aspects of BDSC operations and eventually assume a primary leadership role in the organization. Candidates should have extensive knowledge of *Drosophila* genetics, excellent communication skills, proven grantsmanship, well-developed organizational abilities, interest in a long-term position, and a keen desire to support the *Drosophila* research community.

Applicants must hold a Ph.D. degree and have relevant postdoctoral experience with a solid record of research accomplishments. Experienced scientists are encouraged to apply. All applications received by **September 1, 2015** will be considered. Applicants should submit a cover letter, CV and the names of three references to <http://indiana.peopleadmin.com/postings/1603>. Please address inquiries to **Jeremy Bennett** at (812) 855-6283; jebennet@indiana.edu; 1001 E. Third St., Bloomington, IN 47405-7005.

Indiana University is an Equal Employment and Affirmative Action Employer and a provider of ADA services. All qualified applicants will receive consideration for employment without regard to age, ethnicity, color, race, religion, sex, sexual orientation or identity, national origin, disability status or protected veteran status.



**Center for Immunology and
Microbial Disease
Albany Medical College
Faculty Position**

The Center for Immunology & Microbial Disease at Albany Medical College invites applications for a tenure-track, junior or senior faculty position from individuals who have a doctoral degree, postdoctoral experience, and demonstrated research productivity. Those with an interest in host-pathogen interactions are particularly encouraged to apply. The successful candidate will be expected to establish an independent, extramurally-funded research program and participate in the teaching of medical and graduate students. The basic science departments at Albany Medical College are organized as interdisciplinary research centers and the Center for Immunology & Microbial Disease has a focus on microbial pathogenesis and immune defense, particularly as related to biothreat agents and emerging infections. The new faculty recruit will receive a competitive salary, an attractive start-up package, laboratory space in our newly constructed research building, and access to all departmental core services including the Center's fully-staffed Immunology and ABSL-3/BSL-3 Cores. Albany Medical College is located in a mid-sized city within the upstate New York Capital Region, and has easy access to Boston, New York City, and the Adirondack Mountains.

Applicants should send their curriculum vitae, a statement of research plans, and three letters of reference to:

**Faculty Search Committee
Center for Immunology & Microbial Disease
Albany Medical College
47 New Scotland Avenue, MC-151
Albany, NY 12208**

For further information about the Center, visit:
www.amc.edu/Academic/Research/imd.htm

*An Equal Opportunity/Affirmative Action Employer.
Women and minorities are encouraged to apply.*

**Scientific Director
For the Center for Predictive Medicine for
Biodefense and Emerging Infectious Diseases**

The University of Louisville seeks applications and nominations for the position of Scientific Director of its Center for Predictive Medicine for Biodefense and Emerging Infectious Diseases (CPM). The Scientific Director will work closely with the team responsible for managing the NIH-affiliated Regional Biocontainment Laboratory (BSL-2/ABSL-2, BSL-3/ABSL-3) at the University of Louisville and will be expected to lead research and educational activities that contribute towards meeting its operational and financial requirements. The Director will report to the Executive Vice President for Research and Innovation and will be appointed to the Faculty of the academic college and Department that is appropriate to their education and expertise.

The ideal applicant will have a PhD, MD, or equivalent degree, a proven publication record and an established research program with history of funding support. Experience and/or expertise in a regulated environment is preferable. Applicants should review information at the CPM website <http://louisville.edu/predictivemedicine>. Applicants should submit a CV, the names and contact information of four scientific references, and a letter describing the applicant's experience in leading a multidisciplinary research program that may encompass collaborations with individuals, academic institutions, government, industry, contract labs, and/or foundations. In addition, a statement that outlines plans for leading the CPM as its Scientific Director should be provided.

Applications should be received by **September 15, 2015** for full consideration, although applications will continue to be reviewed until the position is filled. Nominations and applications will remain confidential until invited for campus interview.

All applications and requested material must be submitted on-line at: https://higherdecisions.com/uofl/current_vacancies.asp **Job ID UL374**. Use the left margin option, **View Vacancy Announcement** to read the ad, and/or the **Applicant Guide** to apply.

The University of Louisville is an Affirmative Action, Equal Opportunity, Americans with Disabilities Employer.

**UNIVERSITY OF
LOUISVILLE**

The space roboticist

The first motorized vehicle that Vandi Verma ever operated was a tractor. “I must’ve been 11 years old at the time,” she told *Science*. During school vacations, she visited her grandparents, who lived in a village in central India. At their farm, her uncle let her take a few turns behind the tractor wheel. Later, when she was a teenager, her father, who was a pilot with the Indian Air Force, taught her how to drive a car. That was unusual in India at that time, where those who could afford a car hired a driver.

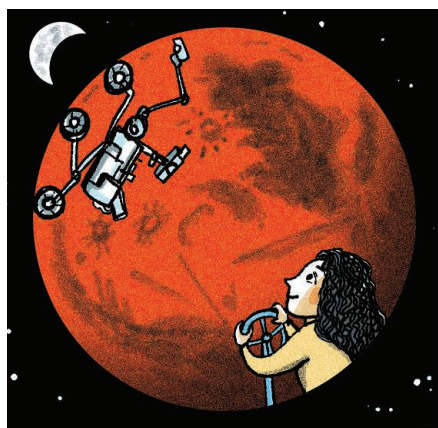
Today, Verma is one of the few people in the world who is qualified to drive a vehicle on Mars.

Verma majored in electrical engineering in India and came to the United States to study artificial intelligence. She was captivated by the landing of the Sojourner Mars rover in 1997 and decided to apply her engineering skill to space exploration. She pursued a Ph.D. in robotics at Carnegie Mellon University and did internships with NASA’s Ames Research Center. She also got her first taste of robotic exploration here on Earth by field testing a rover that surveyed South America’s Atacama Desert for signs of life.

After graduating, in 2005, Verma joined the intelligent systems division at Ames Research Center as a research scientist. Later, she moved to NASA’s Jet Propulsion Lab (JPL), the command center for the Mars rover missions. There, her robotics expertise and experience with field testing rovers won her a chance to drive the Opportunity rover. Verma drove Opportunity for 3 years before graduating to the nuclear-powered Curiosity rover, which is now prowling Mars, examining its rocks to see whether it is, or was ever, a suitable habitat for life.

Each day, before the rover shuts down for the frigid martian night, it calls home, Verma says. Besides relaying scientific data and images it gathered during the day, it sends its precise coordinates. They are downloaded into simulation software Verma helped write. The software helps drivers plan the rover’s route for the next day, simulating tricky maneuvers. Operators may even perform a dry run with a duplicate rover on a sandy replica of the planet’s surface in JPL’s Mars Yard. Then the full day’s itinerary is beamed to the rover so that it can set off purposefully each dawn.

For the first 3 months after landing, from 5 August to 5 November 2012, the team was working on Mars time. “I loved that we didn’t have to wait long after we uplinked our



“I am happy to be ... pushing the envelope on space exploration.”

commands to see the results from Curiosity, and every sol (a martian day), we were doing something we’d never done before,” Verma says.

Curiosity can scoop dirt, drill rock, and hand off samples to the onboard lab. While a sample is being analyzed, Curiosity is already on its way to the next site. Verma helped write the code that lets Curiosity juggle these tasks. “We have to drive on to find newer things for the slew of instruments to analyze without compromising the rover hardware or the sample,” she says.

She loves her day-to-day responsibility for the machine. “You definitely don’t want to be the one who drove the rover off a cliff! But I find it energizing rather than stressful. You’re completely focused.”

But she has not left research behind. One of Verma’s key research goals has been to give rovers greater autonomy to decide on a course of action. She is now working on a software upgrade that will let Curiosity be true to its name. It will allow the rover to autonomously select interesting rocks, stopping in the middle of a long drive to take high-resolution images or analyze a rock with its laser, without any prompting from Earth.

Originally designed for a 2-year mission, Curiosity is still going strong and has already made many scientifically significant finds. “With every drive, we get to explore new terrain that no human has seen in this kind of detail,” Verma says.

Although human spacefaring has stalled, Verma says the spirit of exploration is alive and well in space robots. “I am happy to be working in robotics, pushing the envelope on space exploration,” she says. “We have reached Mars, our neighboring planet. We have only just begun.” ■

Vijaysree Venkatraman is a Boston-based science journalist. For more on life and careers, visit sciencecareers.org. Send your story to SciCareerEditor@aaas.org.