Earth's heart of iron begins to yield its secrets p. 18

Microglia in chronic pain recovery and relapse pp. 33 & 86

Particle acceleration in a nova explosion p. 77

Science

$15
1 APRIL 2022
SPECIAL ISSUE
science.org
AAAS



FILLING THE GAPS

Closing in on a complete human genome p. 42

NOMIS & Science
YOUNG EXPLORER AWARD 2022

Research at the intersection of the social and life sciences

Unconventional. Interdisciplinary. Bold.

The **NOMIS & *Science* Young Explorer Award** recognizes and rewards early-career M.D., Ph.D., or M.D./Ph.D. scientists that perform research at the intersection of the social and life sciences. Essays written by these bold researchers on their recent work are judged for clarity, scientific quality, creativity, and demonstration of cross-disciplinary approaches to address fundamental questions.

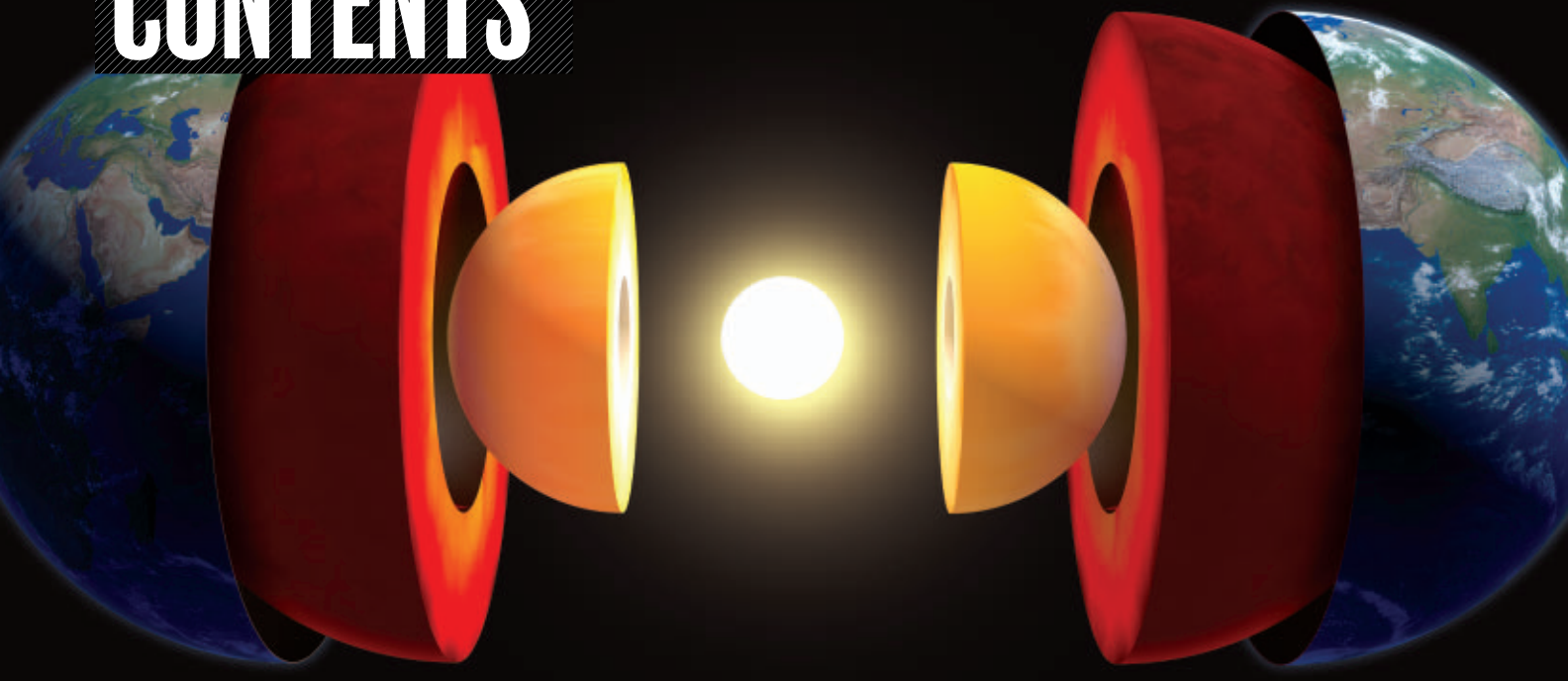
A cash prize of up to 15 000 USD will be awarded to essay winners, and their engaging essays will be published in ***Science***. Winners will also be invited to share their work and forward-looking perspective with leading scientists in their respective fields at an award ceremony as well as a meeting of the NOMIS Board of Directors to consider future funding.

Apply by May 15, 2022
at www.science.org/nomis

NOMIS FOUNDATION
Creating the Spark

Science
AAAS

CONTENTS



1 APRIL 2022 • VOLUME 376 • ISSUE 6588

18

NEWS

IN BRIEF

8 News at a glance

IN DEPTH

10 U.S. pares back neutrino experiment to beat rival
Decision to build troubled megaproject in two phases will add to ballooning cost *By A. Cho*

11 Biden's 2023 budget request for science aims high—again
New spending plan repeats many themes that Congress failed to support fully this year *By Science News Staff*

12 Dirty bomb ingredients go missing from Chornobyl lab
Insecure radioactive materials are the latest worry as Russia continues occupation of infamous nuclear reservation *By R. Stone*

14 Japan reboots HPV vaccination drive after 9-year gap
Modeling suggests 2013 move that undercut immunization will lead to thousands of preventable deaths *By D. Normile*

15 Most complete human genome yet is revealed
Unusual cell line helps sequencers read previously indecipherable stretches of DNA *By E. Pennisi*
COMPLETING THE HUMAN GENOME SECTION p. 42

16 Genital defects seen in sons of men taking major diabetes drug
Large study hints that taking metformin before conception could raise risk of birth defects by affecting sperm *By M. Wadman*

17 Software-designed miniproteins could create new class of drugs
Small versions of antibodies bind to virtually any target protein *By R. F. Service*

FEATURES

18 The planet inside
Scientists are probing the secrets of the inner core—and learning how it might have saved life on Earth *By P. Voosen*
PODCAST

INSIGHTS

LETTERS

24 NextGen Voices: Rules all PIs should follow

PERSPECTIVES

27 The road to a larger brain
Ecological opportunities in the early Cenozoic favored larger, not smarter, mammals *By F. A. Smith*
REPORT p. 80

28 Perovskite solar cells can take the heat
Layered surface structure adds durability to packaged perovskite cells *By J. M. Luther and L. T. Schelhas*
REPORT p. 73

30 Eco-evolutionary effects of keystone genes
The rapid evolution of specific genes within species can drive ecological changes *By P. Nosil and Z. Gompert*
REPORT p. 70

31 Selective cultural processes generate adaptive heuristics
Less intuitive, hard-to-learn cognitive heuristics can thrive *By J. Henrich*
REPORT p. 95

33 Pain-resolving microglia
An emergent subgroup of spinal cord microglia mediates recovery from persistent pain *By G. Sideris-Lampretsas and M. Malcangio*
REPORT p. 86

ILLUSTRATION: C. BICKEL/*SCIENCE*

NEWS FROM Science | AAAS

Subscribe to **News from Science** for unlimited access to authoritative, up-to-the-minute news on research and science policy.



bit.ly/NewsFromScience

SPECIAL SECTION

Completing the human genome

INTRODUCTION

42 Filling the gaps

RESEARCH ARTICLES

44 The complete sequence of a human genome *S. Nurk* et al.
PODCAST

54 A complete reference genome improves analysis of human genetic variation *S. Aganezov* et al.

55 Segmental duplications and their variation in a complete human genome *M. R. Vollger* et al.

56 Complete genomic and epigenetic maps of human centromeres *N. Altemose* et al.

57 From telomere to telomere: The transcriptional and epigenetic state of human repeat elements *S. J. Hoyt* et al.

58 Epigenetic patterns in a complete human genome *A. Gershman* et al.

SEE ALSO NEWS STORY p. 15
PERSPECTIVE p. 34

ON THE COVER

The Telomere-to-Telomere (T2T) Consortium has completed a challenging 8% of the human genome left unresolved by the initial Human Genome Project. In this data visualization, each chromosome begins at bottom right and wraps around, with chromosomes X and 1 through 22 arranged from the outside in (chromosome Y is not shown).



The newly completed regions are highlighted in red. See page 42. *Graphic: V. Altounian/* Science*; Data: T2T Consortium*

PHOTO: MATTHIAS FURLER

34 A next-generation human genome sequence
A near-complete sequence outlines a path for a more inclusive reference
By D. M. Church
COMPLETING THE HUMAN GENOME SECTION p. 42

36 Paul Farmer (1959–2022)
Humanitarian and public health pioneer
By B. Gates

POLICY FORUM

37 A push for inclusive data collection in STEM organizations
Professional societies could better survey, and thus better serve, underrepresented groups
By N. P. Burnett et al.

BOOKS *ET AL.*

40 The scientist-statesman
A new documentary capably navigates the life and legacy of Benjamin Franklin
By R. B. Bernstein

41 War in an automated, data-driven world
An anthropologist misses the mark on a topic of great concern
By G. Lucas

RESEARCH

IN BRIEF

59 From *Science* and other journals

RESEARCH ARTICLE

62 Human evolution
Bronze and Iron Age population movements underlie Xinjiang population history
V. Kumar et al.

REPORTS

70 Ecology
A keystone gene underlies the persistence of an experimental food web *M. A. Barbour* et al.
PERSPECTIVE p. 30

73 Solar cells
Damp heat–stable perovskite solar cells with tailored-dimensionality 2D/3D heterojunctions
R. Azmi et al.
PERSPECTIVE p. 28



A "keystone" gene in *Arabidopsis thaliana* regulates coexistence between a predatory wasp and its aphid prey.

77 Gamma-ray astronomy
Time-resolved hadronic particle acceleration in the recurrent nova RS Ophiuchi *H.E.S.S. Collaboration*

80 Paleontology
Brawn before brains in placental mammals after the end-Cretaceous extinction
O. C. Bertrand et al.
PERSPECTIVE p. 27

86 Chronic pain
A spinal microglia population involved in remitting and relapsing neuropathic pain
K. Kohno et al.
PERSPECTIVE p. 33

90 Membranes
An integrated materials approach to ultrapermeable and ultraselective CO_2 polymer membranes *M. Sandru* et al.

95 Cognitive science
Complex cognitive algorithms preserved by selective social learning in experimental populations *B. Thompson* et al.
PERSPECTIVE p. 31

DEPARTMENTS

7 Editorial
To solve climate, first achieve peace
By H. H. Thorp

102 Working Life
Bucking tradition *By K. H. Kim*

Science Staff 6
Science Careers 100

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2022 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $2212; Foreign postage extra: Air assist delivery: $98. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling available from backissues.science.org; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act can be obtained through the Copyright Clearance Center (CCC), www.copyright.com. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

Science AAAS

Editor-in-Chief Holden Thorp, hthorp@aaas.org

Executive Editor Monica M. Bradford

Editors, Research Valda Vinson, Jake S. Yeston **Editor, Insights** Lisa D. Chong

DEPUTY EDITORS Stella M. Hurtley (UK), Phillip D. Szuromi, Sacha Vignieri **SR. EDITORIAL FELLOW** Andrew M. Sugden (UK) **SR. EDITORS** Gemma Alderton (UK), Caroline Ash (UK), Brent Grocholski, Pamela J. Hines, Di Jiang, Priscilla N. Kelly, Marc S. Lavine (Canada), Yevgeniya Nusinovich, Ian S. Osborne (UK), Beverly A. Purnell, L. Bryan Ray, H. Jesse Smith, Keith T. Smith (UK), Jelena Stajic, Peter Stern (UK), Valerie B. Thompson, Brad Wible, Yuen Yiu **ASSOCIATE EDITORS** Michael A. Funk, Bianca Lopez, Seth Thomas Scanlon (UK), Yury V. Suleymanov, Ekeoma Uzogara **LETTERS EDITOR** Jennifer Sills **LEAD CONTENT PRODUCTION EDITORS** Harry Jach, Lauren Kmec **CONTENT PRODUCTION EDITORS** Amelia Beyna, Jeffrey E. Cook, Chris Filiatreau, Julia Haber-Katris, Nida Masiulis, Abigail Shashikanth, Suzanne M. White **SR. EDITORIAL COORDINATORS** Carolyn Kyle, Beverly Shields **EDITORIAL COORDINATORS** Aneera Dobbins, Joi S. Granger, Jeffrey Hearn, Lisa Johnson, Maryrose Madrid, Ope Martins, Shannon McMahon, Jerry Richardson, Hilary Stewart (UK), Alice Whaley (UK), Anita Wynn **PUBLICATIONS ASSISTANTS** Alexander Kief, Ronmel Navas, Isabel Schnaidt, Brian White **EXECUTIVE ASSISTANT** Jessica Slater **ASI DIRECTOR, OPERATIONS** Janet Clements (UK) **ASI SR. OFFICE ADMINISTRATOR** Jessica Waldock (UK)

News Editor Tim Appenzeller

NEWS MANAGING EDITOR John Travis **INTERNATIONAL EDITOR** Martin Enserink **DEPUTY NEWS EDITORS** Elizabeth Culotta, Lila Guterman, David Grimm, Eric Hand (Europe), David Malakoff **SR. CORRESPONDENTS** Daniel Clery (UK), Jon Cohen, Jeffrey Mervis, Elizabeth Pennisi **ASSOCIATE EDITORS** Jeffrey Brainard, Kelly Servick, Catherine Matacic **NEWS REPORTERS** Adrian Cho, Jennifer Couzin-Frankel, Jocelyn Kaiser, Rodrigo Pérez Ortega (Mexico City), Robert F. Service, Erik Stokstad, Paul Voosen, Meredith Wadman **INTERN** Tess Joose **CONTRIBUTING CORRESPONDENTS** Warren Cornwall, Andrew Curry (Berlin), Ann Gibbons, Sam Kean, Eli Kintisch, Kai Kupferschmidt (Berlin), Andrew Lawler, Mitch Leslie, Eliot Marshall, Virginia Morell, Dennis Normile (Tokyo), Elisabeth Pain (Careers), Charles Piller, Gabriel Popkin, Michael Price, Joshua Sokol, Richard Stone, Emily Underwood, Gretchen Vogel (Berlin), Lizzie Wade (Mexico City) **CAREERS** Rachel Bernstein (Editor), Katie Langin (Associate Editor) **COPY EDITORS** Julia Cole (Senior Copy Editor), Morgan Everett, Cyra Master (Copy Chief) **ADMINISTRATIVE SUPPORT** Meagan Weiland

Creative Director Beth Rakouskas

DESIGN MANAGING EDITOR Marcy Atarod **GRAPHICS MANAGING EDITOR** Chris Bickel **INTERIM MULTIMEDIA MANAGING PRODUCER** Sarah Crespi **PHOTOGRAPHY MANAGING EDITOR** William Douthitt **WEB STRATEGY MANAGER** Kara Estelle-Powers **DESIGN EDITOR** Chrystal Smith **DESIGNER** Christina Aycock **GRAPHICS EDITOR** Nirja Desai **INTERACTIVE GRAPHICS EDITOR** Kelly Franklin **SENIOR GRAPHICS SPECIALISTS** Holly Bishop, Nathalie Cary **SENIOR SCIENTIFIC ILLUSTRATOR** Valerie Altounian **SCIENTIFIC ILLUSTRATOR** Ashley Mastin **VIDEO PRODUCER** Meagan Cantwell **SENIOR PHOTO EDITOR** Emily Petersen **PHOTO EDITOR** Kaitlyn Dolan **SOCIAL MEDIA STRATEGIST** Jessica Hubbard **SOCIAL MEDIA PRODUCER** Sabrina Jenkins **WEB DESIGNER** Jennie Pajerowski

Chief Executive Officer and Executive Publisher Sudip Parikh

Publisher, Science Family of Journals Bill Moran

DIRECTOR, BUSINESS SYSTEMS AND FINANCIAL ANALYSIS Randy Yi **DIRECTOR, BUSINESS OPERATIONS & ANALYSIS** Eric Knott **DIRECTOR OF ANALYTICS** Enrique Gonzales **MANAGER, BUSINESS OPERATIONS** Jessica Tierney **MANAGER, BUSINESS ANALYSIS** Cory Lipman **BUSINESS ANALYSTS** Kurt Ennis, Maggie Clark **FINANCIAL ANALYST** Isacco Fusi **DIGITAL/PRINT STRATEGY MANAGER** Jason Hillman **SENIOR MANAGER, PUBLISHING AND CONTENT SYSTEMS** Marcus Spiegler **ASSISTANT MANAGER DIGITAL/PRINT** Rebecca Doshi **SENIOR CONTENT & PUBLISHING SYSTEMS SPECIALIST** Jacob Hedrick **SENIOR CONTENT SPECIALISTS** Steve Forrester, Lori Murphy **PRODUCTION SPECIALIST** Kristin Wowk **DIGITAL PRODUCTION MANAGER** Lisa Stanford **CONTENT SPECIALIST** Kimberley Oster **ADVERTISING PRODUCTION OPERATIONS MANAGER** Deborah Tompkins **DESIGNER, CUSTOM PUBLISHING** Jeremy Huntsinger **SR. TRAFFIC ASSOCIATE** Christine Hall **SPECIAL PROJECTS ASSOCIATE** Sarah Dhere

ASSOCIATE DIRECTOR, BUSINESS DEVELOPMENT Justin Sawyers **GLOBAL MARKETING MANAGER** Allison Pritchard **DIGITAL MARKETING MANAGER** Aimee Aponte **JOURNALS MARKETING MANAGER** Shawana Arnold **MARKETING ASSOCIATES** Aaron Helmbrecht, Ashley Hylton, Mike Romano, Tori Velasquez, Jenna Voris **SENIOR DESIGNER** Kim Huynh

DIRECTOR AND SENIOR EDITOR, CUSTOM PUBLISHING Sean Sanders **ASSISTANT EDITOR, CUSTOM PUBLISHING** Jackie Oberst **PROJECT MANAGER** Melissa Collins

DIRECTOR, PRODUCT & PUBLISHING DEVELOPMENT Chris Reid **DIRECTOR, BUSINESS STRATEGY AND PORTFOLIO MANAGEMENT** Sarah Whalen **ASSOCIATE DIRECTOR, PRODUCT MANAGMENT** Kris Bishop **PRODUCT DEVELOPMENT MANAGER** Scott Chernoff **PUBLISHING TECHNOLOGY MANAGER** Michael Di Natale **SR. PRODUCT ASSOCIATE** Robert Koepke **PRODUCT ASSOCIATE** Caroline Breul, Anne Mason **SPJ ASSOCIATE MANAGER** Samantha Bruno Fuller **SPJ ASSOCIATE** Casey Buchta

MARKETING MANAGER Kess Knight **BUSINESS DEVELOPMENT MANAGER** Rasmus Andersen **SENIOR INSTITUTIONAL LICENSING MANAGER** Ryan Rexroth **INSTITUTIONAL LICENSING MANAGER** Marco Castellan, Claudia Paulsen-Young **SENIOR MANAGER, INSTITUTIONAL LICENSING OPERATIONS** Judy Lillibridge **SENIOR OPERATIONS ANALYST** Lana Guz **SYSTEMS & OPERATIONS ANALYST** Ben Teincuff **FULFILLMENT ANALYST** Aminta Reyes

DIRECTOR, GLOBAL SALES Tracy Holmes **US EAST COAST AND MID WEST SALES** Stephanie O'Connor **US MID WEST, MID ATLANTIC AND SOUTH EAST SALES** Chris Hoag **US WEST COAST SALES** Lynne Stickrod **ASSOCIATE DIRECTOR, ROW** Roger Goncalves **SALES REP, ROW** Sarah Lelarge **SALES ADMIN ASSISTANT, ROW** Victoria Glasbey **DIRECTOR OF GLOBAL COLLABORATION AND ACADEMIC PUBLISHING RELATIONS, ASIA** Xiaoying Chu **ASSOCIATE DIRECTOR, INTERNATIONAL COLLABORATION** Grace Yao **SALES MANAGER** Danny Zhao **MARKETING MANAGER** Kilo Lan **ASCA CORPORATION, JAPAN** Yoshimi Toda (Tokyo), Miyuki Tani (Osaka)

DIRECTOR, COPYRIGHT, LICENSING AND SPECIAL PROJECTS Emilie David **RIGHTS AND PERMISSIONS ASSOCIATE** Elizabeth Sandler **LICENSING ASSOCIATE** Virginia Warren **CONTRACT SUPPORT SPECIALIST** Michael Wheeler

MAIN HEADQUARTERS
Science/AAAS
1200 New York Ave. NW
Washington, DC 20005

SCIENCE INTERNATIONAL
Clarendon House
Clarendon Road
Cambridge, CB2 8FH, UK

SCIENCE CHINA
Room 1004, Culture Square
No. 59 Zhongguancun St.
Haidian District, Beijing, 100872

SCIENCE JAPAN
ASCA Corporation
Sibaura TY Bldg. 4F, 1-14-5
Shibaura Minato-ku
Tokyo, 108-0073 Japan

EDITORIAL
science_editors@aaas.org

NEWS
science_news@aaas.org

INFORMATION FOR AUTHORS
science.org/authors/
science-information-authors

REPRINTS AND PERMISSIONS
science.org/help/
reprints-and-permissions

MEDIA CONTACTS
scipak@aaas.org

MULTIMEDIA CONTACTS
SciencePodcast@aaas.org
ScienceVideo@aaas.org

INSTITUTIONAL SALES AND SITE LICENSES
science.org/librarian

PRODUCT ADVERTISING & CUSTOM PUBLISHING
advertising.science.org/
products-services
science_advertising@aaas.org

CLASSIFIED ADVERTISING
advertising.science.org/
science-careers
advertise@sciencecareers.org

JOB POSTING CUSTOMER SERVICE
employers.sciencecareers.org
support@sciencecareers.org

MEMBERSHIP AND INDIVIDUAL SUBSCRIPTIONS
science.org/subscriptions

MEMBER BENEFITS
aaas.org/membership/benefits

AAAS BOARD OF DIRECTORS
CHAIR Claire M. Fraser
PRESIDENT Susan G. Amara
PRESIDENT-ELECT Gilda A. Barabino
TREASURER Carolyn N. Ainslie
CHIEF EXECUTIVE OFFICER Sudip Parikh
BOARD Cynthia M. Beall
Rosina M. Bierbaum
Ann Bostrom
Janine Austin Clayton
Laura H. Greene
Kaye Husbands Fealing
Maria M. Klawe
Robert B. Millard
William D. Provine

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 or conflicting points of view. Accordingly, all articles published in *Science*—including editorials, news and comment, and book 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.

BOARD OF REVIEWING EDITORS (Statistics board members indicated with **S**)

Erin Adams, *U. of Chicago*
Takuzo Aida, *U. of Tokyo*
Leslie Aiello, *Wenner-Gren Fdn.*
Deji Akinwande, *UT Austin*
Judith Allen, *U. of Manchester*
Marcella Alsan, *Harvard U.*
Sebastian Amigorena, *Inst. Curie*
James Analytis, *UC Berkeley*
Trevor Archer, *NIEHS, NIH*
Paola Arlotta, *Harvard U.*
David Awschalom, *U. of Chicago*
Delia Baldassarri, *NYU*
Nenad Ban, *ETH Zürich*
Nandita Basu, *U. of Waterloo*
Franz Bauer, *Pontificia U. Católica de Chile*
Ray H. Baughman, *UT Dallas*
Carlo Beenakker, *Leiden U.*
Yasmine Belkaid, *NIAID, NIH*
Philip Benfey, *Duke U.*
Kiros T. Berhane, *Columbia U.*
Joseph J. Berry, *NREL*
Alessandra Biffi, *Harvard Med.*
Chris Bowler, *École Normale Supérieure*
Ian Boyd, *U. of St. Andrews*
Malcolm Brenner, *Baylor Coll. of Med.*
Emily Brodsky, *UC Santa Cruz*
Ron Brookmeyer, *UCLA* (**S**)
Christian Büchel, *UKE Hamburg*
Dennis Burton, *Scripps Res.*
Carter Tribley Butts, *UC Irvine*
György Buzsáki, *NYU School of Med.*
Mariana Byndloss, *Vanderbilt U. Med. Ctr.*
Annmarie Carlton, *UC Irvine*
Simon Cauchemez, *Inst. Pasteur*
Ling-Ling Chen, *SIBCB, CAS*
M. Keith Chen, *UCLA*
Zhijian Chen, *UT Southwestern Med. Ctr.*
Ib Chorkendorff, *Denmark TU*
Amander Clark, *UCLA*
James J. Collins, *MIT*
Robert Cook-Deegan, *Arizona State U.*
Virginia Cornish *Columbia U.*
Carolyn Coyne, *Duke U.*
Roberta Croce, *VU Amsterdam*
Christina Curtis, *Stanford U.*
Ismaila Dabo, *Penn State U.*
Jeff L. Dangl, *UNC*
Chiara Daraio, *Caltech*
Nicolas Dauphas, *U. of Chicago*
Christian Davenport, *U. of Michigan*
Frans de Waal, *Emory U.*
Claude Desplan, *NYU*
Sandra Díaz, *U. Nacional de Córdoba*
Samuel Díaz-Muñoz, *UC Davis*
Ulrike Diebold, *TU Wien*
Stefanie Dimmeler, *Goethe-U. Frankfurt*
Hong Ding, *Inst. of Physics, CAS*
Dennis Discher, *UPenn*
Jennifer A. Doudna, *UC Berkeley*
Ruth Drdla-Schutting, *Med. U. Vienna*
Raissa M. D'Souza, *UC Davis*
Bruce Dunn, *UCLA*
William Dunphy, *Caltech*
Scott Edwards, *Harvard U.*
Todd Ehlers, *U. of Tübingen*
Nader Engheta, *UPenn*
Karen Ersche, *U. of Cambridge*
Beate Escher, *UFZ & U. of Tübingen*
Barry Everitt, *U. of Cambridge*
Vanessa Ezenwa, *U. of Georgia*
Toren Finkel, *U. of Pitt. Med. Ctr.*
Gwenn Flowers, *Simon Fraser U.*
Peter Fratzl, *Max Planck Inst. Potsdam*
Elaine Fuchs, *Rockefeller U.*
Jay Gallagher, *U. of Wisconsin*
Daniel Geschwind, *UCLA*
Ramon Gonzalez, *U. of South Florida*
Sandra González-Bailón, *UPenn*
Nicolas Gruber, *ETH Zürich*
Hua Guo, *U. of New Mexico*
Taekjip Ha, *Johns Hopkins U.*
Daniel Haber, *Mass. General Hos.*
Sharon Hammes-Schiffer, *Yale U.*
Wolf-Dietrich Hardt, *ETH Zürich*
Louise Harra, *U. Coll. London*
Carl-Philipp Heisenberg, *IST Austria*
Janet G. Hering, *Eawag*
Christoph Hess, *U. of Basel & U. of Cambridge*
Heather Hickman, *NIAID, NIH*
Hans Hilgenkamp, *U. of Twente*
Janneke Hille Ris Lambers, *ETH Zürich*
Kai-Uwe Hinrichs, *U. of Bremen*
Deirdre Hollingsworth, *U. of Oxford*
Randall Hulet, *Rice U.*
Auke Ijspeert, *EPFL*
Gwyneth Ingram, *ENS Lyon*
Darrell Irvine, *MIT*
Akiko Iwasaki, *Yale U.*
Stephen Jackson, *USGS & U. of Arizona*
Erich Jarvis, *Rockefeller U.*
Peter Jonas, *IST Austria*
Matt Kaeberlein, *U. of Wash.*
William Kaelin Jr., *Dana-Farber Cancer Inst.*
Daniel Kammen, *UC Berkeley*
Kisuk Kang, *Seoul Nat. U.*
Sabine Kastner, *Princeton U.*
V. Narry Kim, *Seoul Nat. U.*
Robert Kingston, *Harvard Med.*
Nancy Knowlton, *Smithsonian Institution*
Etienne Koechlin, *École Normale Supérieure*
Alex L. Kolodkin, *Johns Hopkins U.*
Julija Krupic, *U. of Cambridge*
Paul Kubes, *U. of Calgary*
Gabriel Lander, *Scripps Res.* (**S**)
Mitchell A. Lazar, *UPenn*
Luis Liz-Marzán, *CIC biomaGUNE*
Omar Lizardo, *UCLA*
Jonathan Losos, *Wash. U. in St. Louis*
Ke Lu, *Inst. of Metal Res., CAS*
Christian Lüscher, *U. of Geneva*
Jean Lynch-Stieglitz, *Georgia Inst. of Tech.*
David Lyons, *U. of Edinburgh*
Fabienne Mackay, *QIMR Berghofer*
Anne Magurran, *U. of St. Andrews*
Ari Pekka Mähönen, *U. of Helsinki*
Asifa Majid, *U. of York*
Oscar Marín, *King's Coll. London*
Charles Marshall, *UC Berkeley*
Christopher Marx, *U. of Idaho*
David Masopust, *U. of Minnesota*
Geraldine Masson, *CNRS*
C. Robertson McClung, *Dartmouth*
Rodrigo Medellín, *U. Nacional Autónoma de México*
C. Jessica Metcalf, *Princeton U.*
Baoxia Mi, *UC Berkeley*
Tom Misteli, *NCI, NIH*
Alison Motsinger-Reif, *NIEHS, NIH* (**S**)
Suresh Naidu, *Columbia U.*
Danielle Navarro, *U. of New South Wales*
Daniel Nettle, *Newcastle U.*
Daniel Neumark, *UC Berkeley*
Beatriz Noheda, *U. of Groningen*
Helga Nowotny, *Vienna Sci. & Tech. Fund*
Rachel O'Reilly, *U. of Birmingham*
Pilar Ossorio, *U. of Wisconsin*
Andrew Oswald, *U. of Warwick*
Isabella Pagano, *Istituto Nazionale di Astrofisica*
Elizabeth Levy Paluck, *Princeton U.*
Jane Parker, *Max Planck Inst. Cologne*
Giovanni Parmigiani, *Dana-Farber Cancer Inst.* (**S**)
Daniel Pauly, *U. of British Columbia*
Ana Pêgo, *U. do Porto*
Samuel Pfaff, *Salk Inst.*
Julie Pfeiffer, *UT Southwestern Med. Ctr.*
Philip Phillips, *UIUC*
Matthieu Piel, *Inst. Curie*
Kathrin Plath, *UCLA*
Martin Plenio, *Ulm U.*
Katherine Pollard, *UCSF*
Elvira Poloczanska, *Alfred-Wegener-Inst.*
Julia Pongratz, *Ludwig Maximilians U.*
Philippe Poulin, *CNRS*
Jonathan Pritchard, *Stanford U.*
Lei Stanley Qi, *Stanford U.*
Trevor Robbins, *U. of Cambridge*
Joeri Rogelj, *Imperial Coll. London*
Amy Rosenzweig, *Northwestern U.*
Mike Ryan, *UT Austin*
Miquel Salmeron, *Lawrence Berkeley Nat. Lab*
Nitin Samarth, *Penn State U.*
Erica Ollmann Saphire, *La Jolla Inst.*
Joachim Saur, *U. zu Köln*
Alexander Schier, *Harvard U.*
Wolfram Schlenker, *Columbia U.*
Susannah Scott, *UC Santa Barbara*
Anuj Shah, *U. of Chicago*
Vladimir Shalaev, *Purdue U.*
Jie Shan, *Cornell U.*
Beth Shapiro, *UC Santa Cruz*
Jay Shendure, *U. of Wash.*
Steve Sherwood, *U. of New South Wales*
Brian Shoichet, *UCSF*
Robert Siliciano, *JHU School of Med.*
Lucia Sivilotti, *U. Coll. London*
Richard Smith, *UNC* (**S**)
John Speakman, *U. of Aberdeen*
Tara Spires-Jones, *U. of Edinburgh*
Allan C. Spradling, *Carnegie Institution for Sci.*
V. S. Subrahmanian, *Northwestern U.*
Ira Tabas, *Columbia U.*
Eriko Takano, *U. of Manchester*
Patrick Tan, *Duke-NUS Med. School*
Sarah Teichmann, *Wellcome Sanger Inst.*
Rocio Titiunik, *Princeton U.*
Shubha Tole, *Tata Inst. of Fundamental Res.*
Maria-Elena Torres Padilla, *Helmholtz Zentrum München*
Kimani Toussaint, *Brown U.*
Barbara Treutlein, *ETH Zürich*
Jason Tylianakis, *U. of Canterbury*
Wim van der Putten, *Netherlands Inst. of Ecology*
Ivo Vankelecom, *KU Leuven*
Judith Varner, *UC San Diego*
Henrique Veiga-Fernandes, *Champalimaud Fdn.*
Reinhilde Veugelers, *KU Leuven*
Bert Vogelstein, *Johns Hopkins U.*
Julia Von Blume, *Yale School of Med.*
David Wallach, *Weizmann Inst.*
Jane-Ling Wang, *UC Davis* (**S**)
Jessica Ware, *Amer. Mus. of Natural Hist.*
David Waxman, *Fudan U.*
Chris Wikle, *U. of Missouri* (**S**)
Terrie Williams, *UC Santa Cruz*
Ian A. Wilson, *Scripps Res.* (**S**)
Hao Wu, *Harvard U.*
Li Wu, *Tsinghua U.*
Wei Xie, *Tsinghua U.*
Benjamin Youngblood, *St. Jude*
Yu Xie, *Princeton U.*
Jan Zaanen, *Leiden U.*
Kenneth Zaret, *UPenn School of Med.*
Bing Zhu, *Inst. of Biophysics, CAS*
Xiaowei Zhuang, *Harvard U.*
Maria Zuber, *MIT*

EDITORIAL

To solve climate, first achieve peace

The horrific invasion of Ukraine by Russia has many devastating effects. The most immediate are on the people of Ukraine, but the long-term implications for the entire planet are enormous. For science, the disruption to international collaboration must be addressed and we must give our strongest support to Ukrainian scientists, as outlined by Marcia McNutt and John Hildebrand in a recent *Science* editorial. But for climate change, the effects may be the greatest. If we want a positive energy future for a healthier climate, the West must start by recasting foreign policy with climate and energy issues at the forefront. That can only succeed if nations strengthen the commitment to settle differences with diplomacy, not war. The only truly life-sustaining climate will be one accompanied by international peace.

The impact of climate change on national and global security is not new. But this association was not always front and center. The push for renewable energy has deep roots in geopolitical and security concerns rather than in climate. But for too long, this geopolitical tension has been kept in a separate bucket from climate dangers. Last week, in addressing the Ukraine war, President Biden pledged that the United States would help Europe become less dependent on Russian oil. But nothing was said about climate. People should be made aware that the future of the planet is inextricably intertwined with international conflict. Many fossil-fuel–dependent countries, including the United States, have long been subjected to a volatile global oil market that is largely controlled by the Middle East. The war in Ukraine has served as a grim reminder that Europe has allowed itself to become dependent on fossil fuels from Russia—a country that is now an enemy. Without Russian oil, which nations can Europe turn to for fossil fuels without leaving itself the potential victim of future conflicts elsewhere on the planet? And that may be a short-term solution while it tries to transition to renewable energy.

Meanwhile, China keeps building coal-fired electricity plants. Why aren't countries working with China to slow this down? While the United States hurls invectives toward Beijing about the origins of COVID-19, economic espionage, and intellectual property protection, it is losing the opportunity to cooperate with China on energy sources cleaner than coal. Is a purely adversarial foreign policy the best strategy when China produces greenhouse gases (more than any other country in the world) at its discretion?

Unfortunately, solar and wind power won't provide enough energy in the foreseeable future. That leaves an opening for more nuclear power—but again, the war in Ukraine reminds us that it's not only a matter of building more nuclear plants but also of reducing the risks they present during times of tension. A nuclear power plant is a dangerous military target. The attacks by Russia on Ukrainian nuclear facilities are alarming—they have the potential to scatter nuclear waste over large areas and even provide the ingredients for a dirty bomb. Nuclear power cannot provide a path to a more equitable and healthy energy future in a world of geopolitical conflict. Also, if a country can build a nuclear power plant, it can build a nuclear weapon. The scientific case for replacing fossil fuels with nuclear power is strong, but the political case gets much weaker in a hostile world.

> "...the future of the planet is inextricably intertwined with international conflict."

Last year, President Biden said that climate change was the greatest threat to national security. That is correct and a reassuring statement from a president when so many climate deniers vie for power in the United States and around the world. But the reverse is also true. A hostile world is the greatest threat to a generative policy for dealing with climate change.

It's time for US and European foreign policy to be reframed in the context of addressing climate change. People around the world must better understand that a healthier world where all forms of energy are appropriately utilized requires a world that focuses first on peace. Without working toward peace as the first step, international efforts to tackle global climate change and promote renewable energy and sustainable development cannot progress. Global governance becomes more fragile, and as climate change worsens, it will threaten world stability, trigger humanitarian crises, and provoke more war, in a deadly cycle.

To win the climate war, we must win the climate peace.

–H. Holden Thorp



PHOTO: CAMERON DAVIDSON

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

10.1126/science.abq2761

NEWS

> “I’ve often heard folks say, ‘Let’s get back to normal.’ Well, normal was not equitable.”
>
> **Higher education researcher Christa Porter,** at a U.S. national science academies forum, on the pandemic’s long-term impact on the careers of women scientists.

IN BRIEF

Edited by
Jeffrey Brainard



PALEONTOLOGY

Stan the *T. rex* gets new home

The *Tyrannosaurus rex* fossil known as Stan, auctioned in October 2020 for a record $31.8 million, will be housed in a new natural history museum in Abu Dhabi, United Arab Emirates, to be completed by 2025, the city’s Department of Culture and Tourism announced last week. The 67-million-year-old, 11.7-meter fossil (above) was found in 1999 in South Dakota and has generated many research papers because it’s well preserved and almost complete. It was dug up and owned by the Black Hills Institute of Geological Research until its auction. The private buyer was anonymous, and scientists feared they wouldn’t be allowed to conduct further studies. But the museum will provide access, officials there told *National Geographic*, which first reported the new museum’s acquisition. Joining Stan will be a sample from the Murchison meteorite, which landed in Australia more than 40 years ago and has advanced scientists’ understanding of the early Solar System.

PHOTO: ANGELA WEISS/AFP/GETTY IMAGES

China seeks more ethics reviews

POLICY | China’s government last week released guidance for strengthening the ethical governance of research in life science, medicine, artificial intelligence, and other “sensitive” fields. Research institutions and companies working in these areas must establish committees to review and monitor research involving humans and animals and ensure compliance with national and international ethical standards. Many Chinese institutions already have such boards, but national oversight was lacking. China’s handling of such issues has been under scrutiny since biophysicist He Jiankui produced the world’s first gene-edited babies, for which he was sentenced in 2019 to prison. The guidelines are wide-ranging but rather vague, says bioethicist Jing-Bao Nie of the University of Otago, Dunedin. For example, they call for the public to participate in the reviews but don’t say how. Details are to

be worked out among the National Science and Technology Ethics Committee, national ministries, and local governments.

FDA OKs second booster

COVID-19 | The U.S. Food and Drug Administration (FDA) this week authorized a second booster dose of COVID-19 vaccines for adults ages 50 and older and for the immunocompromised ages 12 and older. The agency cited the waning effect of earlier doses and data it said revealed no new safety concerns. FDA will allow these groups, which are more vulnerable to poor outcomes from COVID-19 infections, to receive Pfizer's and Moderna's messenger RNA vaccines beginning 4 months after their first booster dose. Also this week, the U.S. Centers for Disease Control and Prevention issued guidance that members of these groups may receive an additional booster. FDA did not consult with a committee of outside advisers before making the decision, which came after Israeli scientists last week posted a preprint analyzing recent data from more than 563,000 people ages 60 and older. It found that a second booster shot substantially reduced COVID-19 mortality.

Microplastics in blood measured

POLLUTION | Researchers have measured tiny particles known as microplastics in human blood, a first that could aid research on the particles' health risks, which are unknown. The microscopic particles come from plastic degrading in the environment. People are exposed by breathing air and consuming food and drink laced with the microplastics, which have been found worldwide. Previously, they had been detected in human placentas and animal organs. But until now, measuring them in blood had proved technically difficult because of its complex mix of molecules. Heather Leslie, a chemist and ecotoxicologist at the Free University of Amsterdam, studied samples from 22 volunteers with a mass spectrometer. Blood from only one volunteer had no detectable microplastics, the team reported last week in *Environment International*.

WHO boosts traditional medicine

BIOMEDICINE | The World Health Organization inked a deal with the Indian government last week to launch a global center for the study of traditional medicine, a field that WHO Director-General Tedros Adhanom Ghebreyesus predicts will be "a game changer for health when founded on evidence, innovation, and sustainability." Indian Prime Minister Narendra Modi, too, has strongly promoted traditional medicine, but scientists have blasted his government's embrace of unproven therapies, including for COVID-19. The center, which opens on 25 April, will be based in Jamnagar, a city in Gujarat state that is a bastion of the ayurvedic school of traditional Indian medicine. India will invest about $250 million in the center to pay for land, a new building, and operating costs for 10 years.

Gene therapy gel shows promise

CLINICAL RESEARCH | A DNA-laden topical skin gel has helped heal wounds from a rare inherited disorder in what is proving to be the first clinical success for this form of gene therapy. In a small study published this week in *Nature Medicine*, a team used the gel, which contains a modified herpesvirus carrying a gene for the protein collagen VII, to treat nine people with recessive dystrophic epidermolysis bullosa. Due to by a mutated version of the collagen gene, the disease causes their skin to tear easily, resulting in painful open wounds, persistent infections, scars, and sometimes skin cancer. Repeated applications of the gel over 25 days healed most patients' wounds within 3 months after treatment; a wound treated with a placebo kept opening and closing. Trial sponsor Krystal Biotech Inc. reported positive results from a larger, 31-person trial at a meeting last week and plans to seek regulatory approval this year.



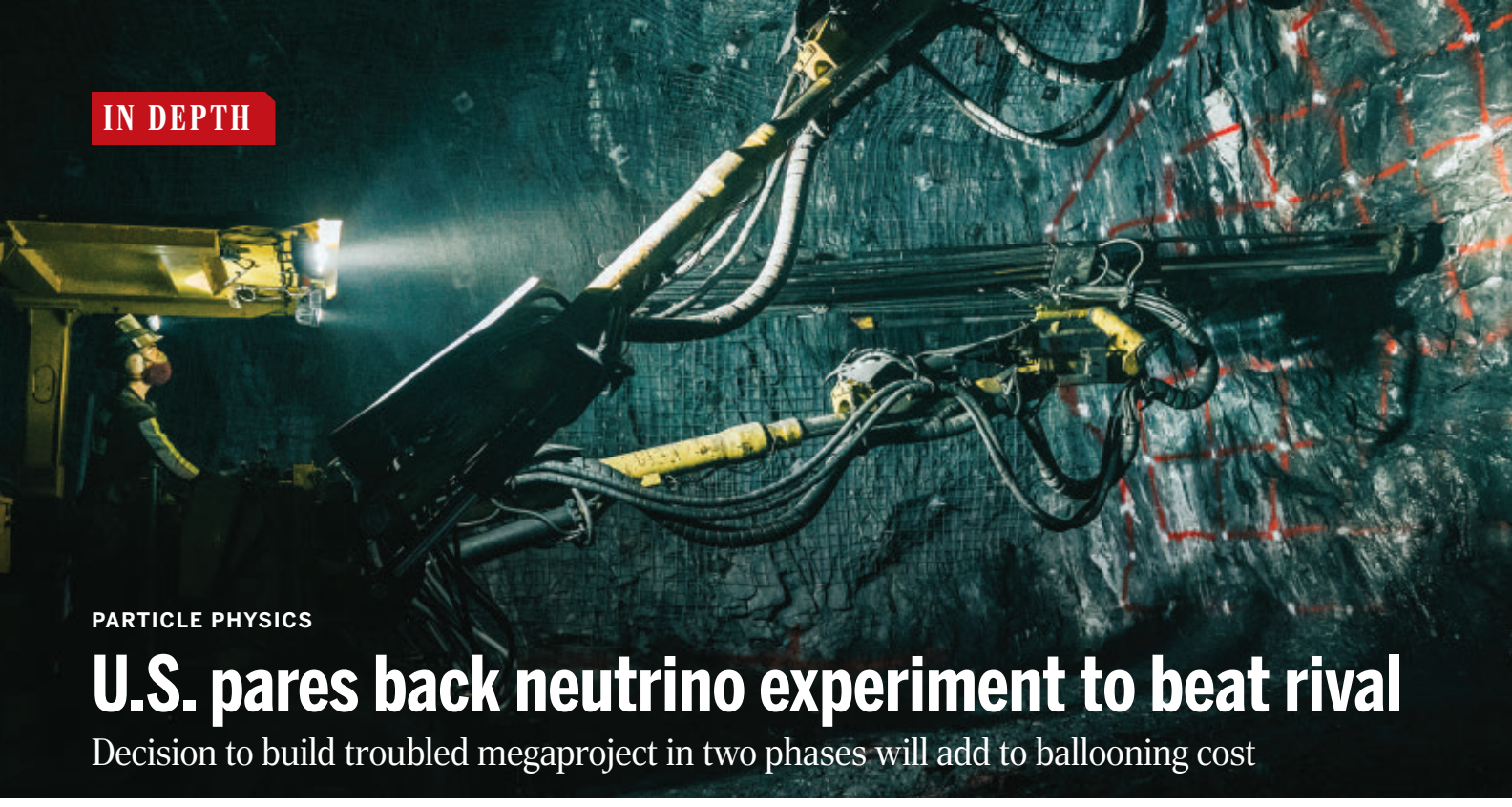
PHOTO: GLENN NICHOLLS/AFP/GETTY IMAGES

ECOLOGY

Bleaching hits Great Barrier Reef again

Warm ocean temperatures have again caused mass bleaching of coral that makes up the famed Great Barrier Reef, for the fourth time since 2016. The damage is occurring in many locations along its 2000-kilometer length, Australia's Great Barrier Reef Marine Park Authority said last week. The increasing frequency of bleaching "decreases the ability of the coral community to bounce back," says ecologist Kathy Townsend of the University of the Sunshine Coast, Fraser Coast. This year's damage was unexpected because it has occurred during a La Niña event, which typically brings cooler weather to the region. A UNESCO team has been visiting Australia to study whether to recommend that the Great Barrier Reef, which is a World Heritage Area, be formally classified as "in danger," a step that would raise pressure for Australia to take action to limit climate change. A year ago, the Australian government, concerned about the impact on tourism, successfully lobbied the World Heritage Committee to hold off making that designation.

A Great Barrier Reef coral turned white last month, after warm ocean temperatures caused it to eject symbiotic algae living in its tissues.



IN DEPTH

PARTICLE PHYSICS

U.S. pares back neutrino experiment to beat rival

Decision to build troubled megaproject in two phases will add to ballooning cost

By **Adrian Cho**

Struggling with cost overruns, the Department of Energy (DOE) has decided to build the United States's next great particle physics experiment in two phases, officials told physicists this month. The decision means the megaproject—actually two intertwined efforts called the Long-Baseline Neutrino Facility (LBNF) and the Deep Underground Neutrino Experiment (DUNE)—won't be completed to its original specs until the mid-2030s instead of late this decade, and that its already ballooning cost will increase even further.

But the adjusted plan should get a slimmed version of LBNF/DUNE running as early as 2029, researchers say. With that timeline, the experiment still has a chance to beat a rival in Japan called Hyper-Kamiokande (Hyper-K) to a key neutrino measurement, says Jim Siegrist, DOE's associate director for high energy physics, who laid the plan out by video on 16 March during a community planning exercise known as the Snowmass process. "DOE believes there is still a chance for LBNF/DUNE to come first," Siegrist wrote in an email.

Physicists involved in the project suggested the change in tack, says Regina Rameika, a neutrino physicist at Fermi National Accelerator Laboratory (Fermilab) and co-spokesperson for the 1400-member DUNE collaboration. "We are all on the same page about what comes first and what comes second," she says. But the stretched-out schedule presents a challenge, says Inés Gil-Botella, a DUNE physicist at Spain's Center for Energy, Environmental, and Technological Research. "We need to work harder to reduce as much as possible any delays between phase one and phase two," she says. The revised plan makes clear that the current $3 billion cost estimate—nearly twice the original—now applies to the first phase alone.

Nearly massless and rarely interacting with other matter, neutrinos come in three types—electron, muon, and tau—depending on how they're generated. One type can morph into another as the particles zip along at near–light-speed. To study those neutrino oscillations, physicists can fire a beam of muon neutrinos generated with a particle accelerator to a huge detector hundreds of kilometers away, which counts the arriving muon neutrinos—plus electron neutrinos that have appeared along the way.

LBNF/DUNE would shoot an intense beam of muon neutrinos from Fermilab to a detector 1300 kilometers away in the abandoned Homestake gold mine in Lead, South Dakota. Filled with frigid liquid argon, the novel DUNE detector would capture the collisions of neutrinos and atomic nuclei in unprecedented detail. Besides putting physicists' theory of neutrino oscillations to the acid test, it would search for charge-parity (CP) violation, an asymmetry between the oscillations of neutrinos and antineutrinos that could help explain how the universe created more matter than antimatter.

After squabbling for a decade, U.S. physicists rallied around the experiment in 2013 in the previous Snowmass. A year later, DOE's Particle Physics Project Prioritization Panel (P5) named it the new flagship domestic effort. In 2015, DOE embraced the P5 vision: an internationally funded experiment that would have a liquid argon detector with a target mass of at least 40,000 tons, a neutrino beam powered by a multimegawatt proton beam, and a smaller detector at Fermilab to monitor the outgoing beam. It would cost less than $1.9 billion and be completed by 2028, DOE estimated. But refurbishment and excavation at the mine have helped drive up the cost and delay progress (*Science*, 24 September 2021, p. 1424).

Now, Siegrist says the project will start smaller, with a detector in South Dakota containing just two of four planned argon modules, with a combined target mass of no more than 24,000 tons. The power of the proton beam at Fermilab will be limited to 1.2 megawatts, and the detector at the lab will be simpler. After completing the first phase, DOE could pursue the bigger detector and better beam in an upgrade likely costing hundreds of millions of dollars.

With just that first phase, LBNF/DUNE should be capable of measuring CP violation. But physicists with Japan's Hyper-K say they'll start to take data in 2028, a year before LBNF/DUNE. Hyper-K will consist of a subterranean cylindrical tank 70 me-

More than 1 kilometer down in a former gold mine in South Dakota, a worker helps excavate tunnels and chambers for a massive neutrino detector.

PHOTO: SANFORD UNDERGROUND RESEARCH FACILITY/CC BY-NC-ND 4.0

ters tall and filled with 260,000 tons of ultrapure water. Some 40,000 phototubes will capture the distinctive flashes of light generated when neutrinos fired from the Japan Proton Accelerator Research Complex in Tokai, 295 kilometers to the east, interact with the water.

Workers in Japan have already begun excavation, says Matthew Malek, a Hyper-K physicist at the University of Sheffield. Given Japan's relatively conservative approach to scheduling, that date is likely to hold, he says. "At this stage, there's really no chance that DUNE will turn on before Hyper-K," he says. But Marvin Marshak, a DUNE physicist at the University of Minnesota, Twin Cities, won't concede the race ahead of time. "One thing I learned over the last 50-some years that I've been doing this is not to overestimate your competition," he says.

Hyper-K and DUNE will also search for neutrinos from supernova explosions and signs that protons can decay. DUNE's ultimate goal is to probe physicists' model of three neutrino types with enough precision to tell whether it breaks down at some level, which could be proof of new particles and forces, Gil-Botella says. For technical reasons, Hyper-K will struggle to do so, she says.

DOE has asked physicists to weigh the scientific case for the second phase in the new Snowmass process and subsequent P5 prioritization, potentially giving skeptics a chance to question whether the whole project is worth the higher cost. Tao Han, a theorist at the University of Pittsburgh who has been involved in Snowmass organizing, predicts U.S. physicists will reiterate their support for it. "We will have the community support to reach our final goal," he says.

Malek wonders whether the new plan is DOE's way of saying the first stage is all it will build. "When you build something in multiple stages, the odds of getting beyond the first stage are more unlikely than likely," he says. Mary Bishai, a DUNE physicist at Brookhaven National Laboratory, reads things differently: "Jim Siegrist said 'Go to Snowmass and come back with a plan for the third and fourth detector modules.' He didn't say, 'We're done.'"

The issue comes at a time of flux for the U.S. program. Fermilab, the United States's only dedicated particle physics lab, is searching for a new director after Nigel Lockyer announced in September 2021 he would be stepping down—3 months before DOE failed the lab in its annual performance review, citing in part the troubles with LBNF/DUNE. And on 31 March, Siegrist, 69, will retire after 10 years leading DOE's high energy physics program. As U.S. particle physics enters rough waters, it's unclear who will take the helm. ■

DATA: WHITE HOUSE; AAAS

U.S. FUNDING

Biden's 2023 budget request for science aims high—again

New spending plan repeats many themes that Congress failed to support fully this year

By ***Science*** **News Staff**

President Joe Biden didn't forget research this week when he submitted to Congress a 2023 budget request that calls for a 9.5% increase in domestic discretionary spending. Biden is asking for a 19% increase at the National Science Foundation (NSF), a 9.3% boost for the National Institutes of Health (NIH), 4.5% more for the Department of Energy's (DOE's) Office of Science, and a 5% hike for NASA's science missions. Once again, fighting climate change and boosting sustainable energy technologies also rank high among Biden's research priorities.

But as always when a president submits an annual budget, the hard part will be getting Congress to go along. That process usually runs past the 1 October start of the fiscal year, leading to a temporary freeze on spending at current levels. This year, with midterm elections in November that could shift control of one or both chambers from Democrats to Republicans, a final agreement could easily be delayed until next year.

Even with his party now in control, Biden's first budget blueprint for science was seriously downsized when Congress passed a final 2022 spending bill last month. For example, legislators shrank Biden's proposed budget for a new Advanced Research Projects Agency for Health (ARPA-H) from $6.5 billion to $1 billion, instead giving NIH's existing institutes a boost of 5%. But ARPA-H remains a presidential favorite, with Biden seeking $5 billion for it in 2023.

Likewise, the final 2022 spending bill whittled his proposed increase for NSF from 20% to 4% and eliminated a $500 million request for a new technology directorate that would ramp up NSF's applied research efforts. But Biden has come back with a similar-size request for 2023, including a proposal to launch 10 megacenters to boost regional innovation.

The president's 2023 budget does have some new wrinkles. In a bid for bipartisan support, Biden wants to boost military spending by 4%. (His 2022 budget request only boosted defense by 1.8%, a figure Congress tripled.) And unlike last year, when climbing out of a pandemic-caused recession was his top priority, Biden says his 2023 budget will reduce the federal deficit by raising taxes on the superwealthy and curbing overall spending.

Here are some highlights from the budgets of four key agencies.

NIH

The biomedical research agency's budget would rise $4.3 billion in 2023, to $50.4 billion. But nearly all the new fund-

Wish list

Congress scaled back many of President Joe Biden's priorities in this year's budget, but he has revived his bid for major new initiatives in health, technology, climate, and energy.

AGENCY	2022 FINAL ($ BILLIONS)	2023 REQUEST	% INCREASE
National Institutes of Health	46.178	50.453	9.3%
ARPA-H	1	5	400%
National Science Foundation	8.838	10.492	19%
Technology directorate	0.37	0.88	140%
STEM education directorate	1.006	1.377	37%
NASA science	7.614	7.988	5%
DOE Office of Science	7.475	7.799	4%
ARPA-E	0.45	0.7	56%
NIST	1.23	1.468	38%
S&T labs	0.85	0.975	15%
U.S. Geological Survey	1.393	1.711	23%
NOAA	5.877	6.866	17%
EPA science and technology	0.729	0.863	18%
USDA Agriculture and Food Research Initiative	0.445	0.564	27%

ing would go to ARPA-H, which is intended to fund high-risk, cutting-edge research. The budgets for most of NIH's 27 institutes and centers would remain flat compared with this year.

"We're disappointed," says Eleanor Dehoney, vice president of policy and advocacy for Research!America. A $275 million increase to NIH's base budget, Dehoney says, "simply isn't enough to meet the agency's mission."

Although the president's budget places ARPA-H at NIH, many advocates would like it to be independent. Last month, Congress created it as a standalone agency within the Department of Health and Human Services but gave HHS Secretary Xavier Becerra until this week to decide whether to move it to NIH.

One of the few areas tagged for increases is NIH's Office of Nutrition Research. Its budget would roughly double, to $195 million. NIH's health disparities program would rise by 9%, to $4.4 billion. Biden has also requested $2 million for a Center for Sexual Orientation and Gender Identity Research. A recent National Academies of Sciences, Engineering, and Medicine report recommended federal agencies do a better job of collecting those data.

NIH would also receive $12.1 billion over 5 years as part of an HHS plan to improve pandemic preparedness using "mandatory" funds that do not require annual approval from Congress. The money would fund research on vaccines, treatments, and diagnostics to address high-priority viruses, as well as biosecurity and clinical trial infrastructure.

NSF

Biden's budget request for NSF sends a clear message to Congress: He believes in the agency. The $10.5 billion he has proposed for 2023 is consistent with aspirational spending levels in two pending bills intended to help the United States outcompete China, which the president has urged Congress to pass after reconciling their differences.

His 2023 budget also revives NSF's plan to spend $200 million on 10 regional innovation engines. Twice the size of NSF's current bevy of engineering and science centers, the new centers are designed not just to advance emerging technologies, but also address the workforce and economic needs of various regions of the country. This year's spending bill encourages NSF to start one center, from existing funds, but NSF Director Sethuraman Panchanathan says achieving the president's vision will require "tens of millions of dollars" invested in each of several centers.

Biden's new budget boosts spending for science, technology, engineering, and math education at all levels. Notably, it would grow by more than one-third the annual size of NSF's flagship graduate training program, increasing the number of new graduate research fellowships (GRFs) from 2000 to 2750. It would also raise annual stipends for the 3-year GRFs from $34,000 to $37,000.

DOE

The administration's focus on climate shaped its request for DOE's Office of Science, with the biological and environment research program receiving the biggest increase, a 10.9% boost, to $904 million.

The rest of the 4.5% overall increase, to $7.8 billion, would be spread more or less evenly across the Office of Science's other five major research programs. But Biden has also requested a 56% increase for DOE's Advanced Research Projects Agency-Energy (ARPA-E), to $700 million, in line with the administration's emphasis on developing clean energy technologies. The agency aims to rapidly turn basic research into prototype technologies that the private sector might take over.

The request says ARPA-E will also expand its remit "to invest in climate-related innovations necessary to achieve net zero climate-inducing emissions by 2050." The Biden administration had wanted to create a separate entity, called Advanced Research Projects Agency-Climate, to advance that goal. But last month Congress directed DOE to give the task to ARPA-E.

NASA

The biggest increase within the $8 billion that Biden has requested for the agency's science directorate is a 17% boost for the earth sciences program, to $2.4 billion. Within that, the budget for the Plankton, Aerosol, Cloud, Ocean Ecosystem (PACE) satellite to study marine plankton and clouds would grow by 38%, to $113 million, in preparation for a launch in May 2024. In contrast, planetary science, which jumped by 18% this year, would only increase by 1%, to $3.2 billion. One small casualty is planning for a Mars Ice Mapper mission. Even so, the Mars Sample Return mission, an agency priority, would see a 26% increase, to $822 million. NASA also hopes to nearly double spending next year on Dragonfly, a helicopterlike craft intended to explore Titan, Saturn's largest moon. The launch of the $850 million craft has been delayed a year, to 2027. ■

With reporting by Adrian Cho, Jocelyn Kaiser, Jeffrey Mervis, and Erik Stokstad.

EUROPE

Dirty bomb ingredients go missing from Chornobyl lab

Insecure radioactive materials are the latest worry as Russia continues occupation of infamous nuclear reservation

By **Richard Stone**

When the lights went out at Chornobyl's nuclear power plant on 9 March, the Russian soldiers holding Ukrainian workers at gunpoint became the least of Anatolii Nosovskyi's worries. More urgent was the possibility of a radiation accident at the defunct plant. If emergency generators ran out of fuel, the ventilators that keep explosive hydrogen gas from building up inside a spent nuclear fuel repository would quit working, says Nosovskyi, director of the Institute for Safety Problems of Nuclear Power Plants (ISPNPP) in Kyiv. So would sensors and automated systems to suppress radioactive dust inside a concrete "sarcophagus" that holds the unsettled remains of Chornobyl's Unit Four reactor, which melted down in the infamous 1986 accident.

Although power was restored to Chornobyl on 14 March, Nosovskyi's worries have multiplied. In the chaos of the Russian advance, he told *Science*, looters raided a radiation monitoring lab in Chornobyl village—apparently making off with radioactive isotopes used to calibrate instruments and pieces of radioactive waste that could be mixed with conventional explosives to form a "dirty bomb" that would spread contamination over a wide area. ISPNPP has a separate lab in Chornobyl with even more dangerous materials: "powerful sources of gamma and neutron radiation" used to test devices, Nosovskyi says, as well as intensely radioactive samples of material leftover from the Unit Four meltdown. Nosovskyi has lost contact with the lab, he says, so "the fate of these sources is unknown to us."

The drama at Chornobyl began on 24 February, the very first day of the invasion. At 5 a.m., as Russian troops poured across Ukraine's border with Belarus—just



An arching steel structure protects the smoldering remains of Chornobyl's Unit Four nuclear reactor, but it was not designed to withstand shelling.

15 kilometers from Chornobyl—ISPNPP managers were ordered to evacuate most staff, who monitor the safety of the plant, provide technical support for decommissioning, and develop protocols for managing radioactive waste in the off-limits "exclusion zone" surrounding Chornobyl. Within 2 hours, 67 had cleared out; two who live in Chornobyl village stayed behind to keep an eye on the institute's lab. "We've lost contact with these brave people," says ISPNPP senior scientist Maxim Saveliev.

By 5 p.m., Russian troops had taken control of all Chornobyl facilities. A shift supervisor, Valentin Geiko, negotiated a deal under which the plant's Ukrainian guards would disarm and the Russian soldiers would not interfere with civilian workers, Nosovskyi says. But for nearly a month, the soldiers forbade a shift change—essentially holding the workers hostage—and confiscated their cellphones. In a gesture of defiance, the workers played the Ukrainian national anthem every morning, cranking up the volume, Nosovskyi says. Last week, the occupiers finally allowed fresh staff to rotate in. But some captive workers chose to remain, he adds, "so as not to put at risk people who should come in their place."

Chornobyl is not the only Ukrainian nuclear installation at risk in the war. "There have already been several close calls," says Rafael Mariano Grossi, director general of the International Atomic Energy Agency, who traveled to Kyiv this week to discuss bolstering nuclear security. On 4 March, Russian forces shelled the Zaporizhzhya nuclear power plant—fortunately missing its reactor halls. Two days later, a rocket attack damaged a research reactor used to generate neutrons for experiments at the Kharkiv Institute of Physics and Technology. Nosovskyi labels the assaults as nothing short of state-sponsored "nuclear terrorism."

But Chornobyl has a unique set of radioactive hazards. On 11 March, wildfires ignited in the nearby forests, which harbor radioisotopes that were disgorged in the accident and taken up by plants and fungi. Russian military activities have prevented firefighters from entering the exclusion zone, Nosovskyi says. The fires continue to burn and could grow more intense as the weather warms, he says, releasing radiation that could lead to "significant deterioration of the radiation situation in Ukraine and throughout Europe."

So far, remote measurements suggest radioactive particle concentrations in the smoke do not pose a health hazard, Nosovskyi adds, but an automated radiation monitoring system that went down in the power outage has not yet been brought back online. That means "there is no information on the real situation in the exclusion zone," says Viktor Dolin, research director of the Institute for Environmental Geochemistry in Kyiv.

The restoration of electricity averted the nightmare of a hydrogen explosion in the spent fuel repository, where 8500 tons of uranium fuel rods continue to cool off in pools of water. The repository poses a major radioactive threat: Through radioactive decay, the assemblies have accumulated about 240 times more cesium-137 and 1500 times more strontium-90 than the destroyed reactor spewed in 1986, Dolin says. Staff intend to punch holes in the repository's walls to allow hydrogen gas to escape in the event of a future power outage, Nosovskyi says.

Another menace at Chornobyl are the fuel-containing masses (FCMs)—a mix of fuel rods, zirconium cladding, and other materials that melted in the accident and continue to smolder under Unit Four's sarcophagus, hastily erected in the wake of the disaster. For years Ukrainian scientists, with colleagues from Russia's Kurchatov institute, have kept a tense vigil. (The institute cut ties with Ukrainian partners earlier this month, issuing a statement supporting the war and the "denazification" of Ukraine.) Occasional spikes in the number of neutrons streaming from certain FCMs—a sign of fission—prompt sprinklers to spray gadolinium nitrate solution, which absorbs neutrons.

The odds of self-sustaining fission, or criticality, in an FCM are minuscule, and even if criticality triggered a small explosion, the burst would probably be contained within an arching steel structure, called the New Safe Confinement (NSC), that was erected over the sarcophagus in 2016 to shield it from the elements and create a safe space for cleanup work. But the NSC was not designed to withstand shelling, and a breach could disturb the FCMs. It could also release some of the hundreds of tons of highly radioactive dust that have accumulated in the sarcophagus over the years as the FCMs gradually disintegrate.

Thousands of other sites in Ukraine have radiological materials. Most are under the watchful eye of Ukraine's nuclear regulator. "There's a lot of ongoing effort to secure material," says Peter Martin, a nuclear physicist at the University of Bristol who collaborates with scientists at Chornobyl. That means, where possible, moving sources into vaults and repositories. But Vitaly Fedchenko, a nuclear security expert at the Stockholm International Peace Research Institute, notes that Ukraine, like other parts of the former Soviet Union, has not kept track of all the Soviet nuclear legacy. "There are a lot of radioactive sources that are not on anyone's radar," he says. "Even Ukraine's radar." ■

PHOTO: BRENDAN HOFFMAN/GETTY IMAGES



Japan's government again recommends vaccination against human papillomavirus for girls ages 12 to 16.

PHOTO: FERRANTRAITE/ISTOCKPHOTO.COM

CANCER

Japan reboots HPV vaccination drive after 9-year gap

Modeling suggests 2013 move that undercut immunization will lead to thousands of preventable deaths

By **Dennis Normile**

Nine years ago, Japan's health ministry made what many scientists regarded as a terrible mistake. Pressed by antivaccine activists who claimed debilitating side effects, it stopped recommending that Japanese girls get a vaccine that helps prevent cervical cancer. Now, in what public health officials say is a collateral benefit of the success of COVID-19 vaccines, the ministry has finally reversed its position. On 1 April, it will resume recommending that girls ages 12 to 16 get vaccinated against human papillomavirus (HPV)—"an important signal of confidence in the vaccine and its safety," says Paul Bloem, HPV vaccine strategy lead at the World Health Organization (WHO).

But part of the damage can't be undone. A modeling study published in *The Lancet* in 2020 estimated that the negligible vaccination rate between 2013 and 2019 would result in 25,000 preventable cervical cancer cases and up to 5700 deaths over time. A rapid catch-up campaign for the millions of women who missed their shots—which the government has now pledged to undertake—would only prevent 60% of that toll, the study said, because many of the women have already been infected with HPV. The disease causes nearly 3000 deaths annually in Japan, in part because cervical cancer screening rates are low.

Japan initially embraced HPV vaccines, approving GlaxoSmithKline's bivalent shot—which protects against the two HPV types carrying the greatest cancer risk—in 2009, and Merck & Co.'s quadrivalent vaccine in 2011. In April 2013, the health ministry added both to the national immunization program and started to recommend vaccination.

But just 10 weeks later, an advisory panel suggested suspending the recommendation after a number of girls reported chronic pain, headaches, motor impairment, and other symptoms after immunization. The ministry complied, and the vaccination rate plummeted from about 70% to less than 1% of those eligible.

Such safety problems had not emerged in clinical trials, and in 2017, WHO's Global Advisory Committee on Vaccine Safety said an extensive review of studies from around the world indicated the vaccines were "extremely safe." In Japan, a nationwide survey that same year found unvaccinated girls suffer the symptoms attributed to the vaccines at similar rates as vaccine recipients.

Evidence for effectiveness grew as well. The vaccines were approved because they prevent HPV infection, but by the late 2000s, studies showed they reduced the incidence of precancerous lesions as well. Large studies in Sweden and England, reported in 2020 and 2021, respectively, showed vaccination in the early teen years cut the risk of cervical cancer by age 30 by 87% to 88%.

In other countries roiled by reports of side effects—including Denmark, the United Kingdom, and Colombia—authorities kept recommending the vaccines while investigating the claims, says Heidi Larson, head of the Vaccine Confidence Project at the London School of Hygiene & Tropical Medicine. Vaccination rates dipped, but quickly recovered. But in Japan, the government was slow to review the evidence while antivaccine pressure "just became louder and louder," says Sharon Hanley, a cancer epidemiologist at Hokkaido University. Opponents held press conferences, seminars, and demonstrations, and more than 100 women and girls joined lawsuits against the health ministry and vaccine manufacturers.

Still, calls to reverse the policy increased. In 2017, 17 Japanese academic societies urged the ministry to resume support for vaccination. In 2020 and 2021, a group of parliamentarians led by cervical cancer survivor Junko Mihara asked the health ministry to reconsider its position. The COVID-19 pandemic demonstrated the power of vaccines to reduce severe illness and deaths, which eventually tipped the scales for the HPV vaccines as well, Hanley says. "Antivaccine rhetoric was also not given much space in the media" during the pandemic, she says. (About 80% of Japan's population is fully vaccinated against COVID-19.) In October 2021, the health ministry's advisory committee said there was no reason not to restart recommending HPV vaccination.

Many parents are still wary, and local governments and health care providers will have to convince them of the vaccines' benefits. The shots are in short supply globally, particularly Merck's latest, also approved in Japan, which protects against nine HPV types. And activists are not giving up. Resuming proactive recommendation "is without any scientific basis and is wrong as public health policy," says Masumi Mizuguchi, a lawyer representing plaintiffs suing the government. The lawsuits are working their way through Japan's legal system and will continue to generate publicity.

But vaccination supporters believe the spell may have been broken. "I am confident coverage will resume to previous levels as quickly as it fell simply due to peer power," Hanley says. To her, the 9-year interlude holds an important lesson: "When the government is not supporting [a vaccine], then the people won't support it." ■

GENOMICS

Most complete human genome yet is revealed

Unusual cell line helps sequencers read previously indecipherable stretches of DNA

By **Elizabeth Pennisi**

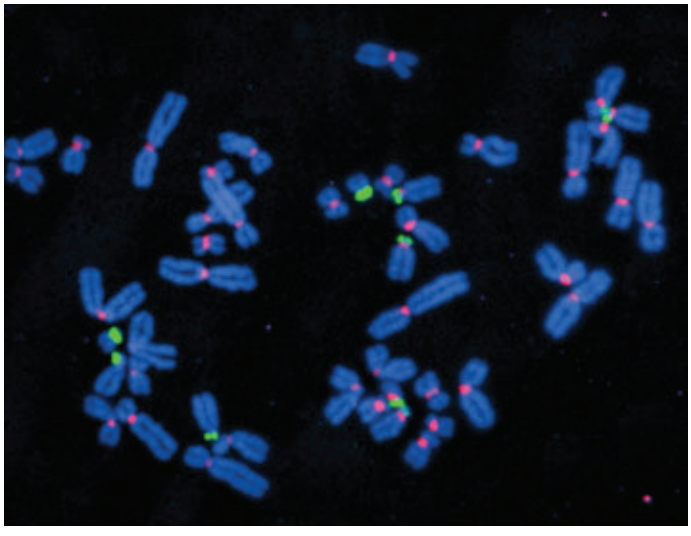
When it comes to sequencing the human genome, "complete" has always been a relative term. The first one, deciphered 20 years ago, included most of the regions that code for proteins but left about 200 million bases of DNA—8% of the human genome—untouched. Even as additional genomes were "finished," some stretches remained out of reach, because repetitive segments of DNA confounded the sequencing technologies of the time. Now, an international grassroots effort has sorted out those hard-to-read bases, producing the most complete human genome yet.

In six papers starting on p. 42, the Telomere-to-Telomere (T2T) Consortium—named for the chromosomes' end caps—fills in all but five of the hundreds of remaining problem spots, leaving just 10 million bases and the Y chromosome only roughly known. And on 31 March, the T2T consortium announced in a tweet it had deposited a correct sequence assembly of the missing Y.

"I don't think we could have imagined this even 5 years ago, certainly not 10 years ago," says bioinformaticist Ewan Birney, deputy director of the European Molecular Biology Laboratory and part of the original Human Genome Project "It's a tour de force." T2T researchers say the newly sequenced stretches reveal hotspots for gene evolution and underscore the chaotic history of the human genome. It "really gives us some insight into regions of the genome that have been invisible," says Deanna Church, a genomicist at Inscripta, a gene-editing company.

The previously indecipherable sequences of the genome that have now come into clear view include the protective telomeres and the dense knobs called centromeres, which typically reside in the middle of each chromosome and help orchestrate its replication. Also almost completely revealed are the short arms of the five chromosomes where centromeres are skewed toward one end. Those short arms were known to contain scores of genes coding for the backbone of ribosomes, the cell's protein factories.

When Birney, Church, and their colleagues introduced that first draft of a human genome in 2001, and even after they "completed" and published it in 2004, sequencer machines and genome assembly software could not wade through areas where the DNA sequence contained very repetitive stretches of bases: The repeats could too easily be skipped or their bases linked together incorrectly. As sequencing technology got better and costs dropped, scientists reduced the number of gaps or misassembled sequences, culminating in 2017 with the release of a human genome called GRCh38. With less than 1000 gaps, it became for many the "reference" against which other human genomes are compared.



Surprises awaited in the short arms (green) and centromeres (pink) of newly deciphered human chromosomes.

PHOTO: T. POTAPOVA AND J. GERTON/STOWERS INSTITUTE FOR MEDICAL RESEARCH

But Karen Miga and Adam Phillippy wanted to do better. Miga, a geneticist at the University of California, Santa Cruz, yearned to learn the exact sequences of the distinctive "satellite" DNA that helps form centromeres. Meanwhile, Phillippy, a bioinformatician at the National Human Genome Research Institute, was busy harnessing new sequencing technologies that could read very long stretches of DNA, reducing the need to piece together shorter sequences. After meeting at a conference, they joined forces. Then in 2019, Phillippy reported they had succeeded in sequencing the X chromosome from end to end, inspiring dozens of other researchers to join the cause. "It really took on a life of its own," Miga says.

To simplify the task, they decided to use an anonymized cell line that was derived more than 20 years ago from an unusual growth excised from the uterus of a woman—a failed pregnancy called a mole, produced when a sperm entered an egg that lacked its own set of chromosomes. With just the sperm's genetic material, such eggs can't develop into an embryo, but they can still replicate, especially if the sperm delivers an X instead of Y chromosome. In a boon for the project, both members of the resulting cell line's 23 pairs of chromosomes are identical. That "made a big difference" for eliminating gaps because sequencers didn't have to resolve differences between the parents' chromosomes, says Robert Waterston, a geneticist at the University of Washington, Seattle, who helped lead the Human Genome Project.

The T2T group combined sequencing technologies, including a so-called nanopore device that could read 100,000 bases at a time and another sequencer that was more accurate but only did about 10,000 bases at once. A final improvement to the latter method boosted accuracy, and together the three approaches were able to polish off all but five of the final trouble spots. "Just seeing the multiple ways they went after this [shows] these are really hard problems," Waterston says.

The approximately 200 million bases finally in the right order and in the right place include more than 1900 genes, most of them copies of known genes. The researchers cataloged duplicated regions and mobile elements—genetic material from viruses that became incorporated into the genome. In sequencing each centromere, they learned the duplicated regions vary greatly in size, unexpected because these knobs serve the same purpose in each chromosome.

The short chromosome arms held another surprise. As expected, they included

multiple copies, 400 in all, of the genes coding for the RNA that's used to make ribosomes. "This rDNA was the last domino to fall," as it was the hardest to sequence, Miga says.

The short arms are also "just chock-full of [other] repeats," says Jennifer Gerton, a chromosome biologist at the Stowers Institute for Medical Research. Those include mobile elements, duplicated segments and other types of repetitive DNA, as well as many copies of genes from other parts of the genome. "It's amazing how dynamic the human genome can be," Church says. In five spots along these chromosomes, the resulting jumble is so long that the researchers still can't clearly determine the order of the bases, although they have a rough idea of the sequence, Gerton says.

Short arms are likely hotspots for gene evolution, Phillippy notes, as gene copies parked there are free to mutate and take on new functions. The catalog of duplications could also shed light on neurological and developmental disorders, which have been linked to variations in the number of copies of specific sequences. Chemical modifications to the DNA in the complex repetitive areas likely play a role in disease as well, and those changes have been mapped.

Because the cell line used lacked a Y chromosome, the T2T group sequenced one from a well-studied genome belonging to Harvard University systems biologist Leonid Peshkin. "I'm excited to be part of this front-line, leading edge in science," says Peshkin, whose sequence will be described in a future paper.

Despite their latest milestone, human genome sequencers aren't packing their bags. "There's still some work to do," says Human Genome Project co-leader Richard Gibbs, a geneticist at Baylor College of Medicine. He and other researchers stress that the field now needs to get similarly complete genome sequences from a greater diversity of people to look for variation in the short arms and the other tough-to-read regions, which could play a role in diseases or traits.

The T2T team has made a start by deciphering 70 more genomes, with a goal of 350 from people of diverse ancestries. These genomes, sequenced as part of the Human Pangenome Reference Consortium, are more challenging to finish because they don't have identical pairs of chromosomes. So, for now, the team has settled for high-quality genomes that place as many of the bases as possible on their correct chromosomes. Next, the researchers plan to apply all their methods to Peshkin's whole genome. And, eventually, Phillippy says, "We want every genome to be telomere to telomere." ■

BIOMEDICINE

Genital defects seen in sons of men taking major diabetes drug

Large study hints that taking metformin before conception could raise risk of birth defects by affecting sperm

By **Meredith Wadman**

Metformin, a first-line diabetes drug used for decades, may boost the risk of birth defects in the offspring of men who took it during sperm development, according to a large Danish study. Sons born to those men were more than three times as likely to have a genital birth defect as unexposed babies, according to the paper, published in the *Annals of Internal Medicine* this week.

The genital defects, such as hypospadias, when the urethra does not exit the tip of the penis, were relatively rare, occurring in 0.9% of all babies whose biological fathers took metformin in the 3 months before conception. But epidemiologists say the findings are important because tens of millions of people worldwide take metformin, chiefly for type 2 diabetes.

"When I saw the paper … I thought: 'Yup, this is gonna go viral,'" says Germaine Buck Louis, a reproductive epidemiologist at George Mason University who wrote an editorial accompanying the report. "[Metformin] is widely used even by young men because of the obesity issue that we have. So that is potentially a huge source of exposure for the next generation."

However, Buck Louis and every other scientist interviewed stressed that the paper's findings are preliminary and observational and need to be corroborated. They cautioned men against abruptly stopping metformin before trying to conceive.

"Metformin is a safe drug, it's cheap, and it does what it needs to do" by controlling blood sugar levels, says the paper's first author, Maarten Wensink, an epidemiologist and biostatistician at the University of Southern Denmark. Any change in medication "is a complex decision that [a couple] should take together with their physicians."

Use of metformin, a synthetic compound that lowers blood sugar by boosting insulin sensitivity, has skyrocketed with the obesity epidemic and attendant diagnoses of type 2 diabetes. In the United States in 2004, 41 million prescriptions for the drug were written; by 2019 that number was 86 million.

The drug has been in use since the 1950s, but this is the first large study to rigorously analyze any paternally mediated impact on human birth defects. Metformin's use skews toward older people, but the rise in diabe-



A nurse measures blood sugar levels in a man with diabetes in Paris.

PHOTO: BSIP/UIG VIA GETTY IMAGES

tes means more younger men are taking it. In the United States, prescriptions to 18- to 49-year-olds with type 2 diabetes grew from fewer than 2200 in 2000 to 768,000 in 2015.

The researchers analyzed records from more than 1.1 million babies born in Denmark between 1997 and 2016, using the country's comprehensive medical registries to connect data on births, metformin prescriptions, and birth defects. In the 1451 offspring of men who filled metformin prescriptions during the 90 days before conception, the period when sperm are being made, the team found 5.2% had birth defects, compared with 3.3% among unexposed babies. That translated to a 1.4-fold increase of at least one major birth defect, including genital, digestive, urinary, and heart defects, after adjustments for paternal age and other factors.

For genital defects alone, the increased risk—only seen in male infants—was much larger. Among exposed babies, 0.9% had genital defects, compared with 0.24% in unexposed babies. The numbers were small—13 metformin-exposed boys were born with genital defects. But after the researchers adjusted for factors including parental ages and maternal smoking status, they found a 3.39-fold rise in the odds of a genital defect. "The rate per se was surprisingly high," Wensink says.

Reassuringly, the researchers saw no effect in offspring of men who took the drug earlier in life or in the year before or after the 90-day window of sperm production. "It really has to do with taking it in that window when the sperm … is being developed," says senior author Michael Eisenberg, a urologist at Stanford Medicine.

The team also found no additional risk in unexposed siblings of metformin-exposed babies, or in infants of diabetic fathers who took insulin or were not on metformin. All those findings suggest it's the drug's impact on sperm formation, rather than diabetes itself or another factor, that's responsible.

But the researchers acknowledge that men with diabetes who took metformin and those who didn't may have differed in attributes such as obesity or how well their disease was controlled—data that were not accessible to the researchers.

Nor are scientists sure how the drug may be impacting sperm. Studies in fish and mice suggest metformin can disrupt the development of male reproductive organs.

The caveats make scientists cautious about drawing conclusions from the paper. "This paper is the first word, not the last word," says Russell Kirby, a birth defects epidemiologist at the University of South Florida. "It's definitely going to require additional research." ■

BIOLOGY

Software-designed miniproteins could create new class of drugs

Small versions of antibodies bind to virtually any target protein

By **Robert F. Service**

Scientists have built their own "mini-antibodies" using innovative software that predicts how proteins fold. Designed to home in on other proteins, the tailor-made molecules could lead to a new class of drugs to fight everything from cancer to COVID-19.

"It's pretty amazing stuff," says Steven Mayo, a chemist at the California Institute of Technology who wasn't involved in the study. The strategy could also produce diagnostic probes that could detect diseases early, adds Tanja Kortemme, a bioengineer at the University of California, San Francisco, also not involved. "It opens up many possibilities."

Because antibodies excel at binding to proteins, such as those on invading microbes, they can serve as drugs to battle infections and cancer. But antibodies are large proteins, costly and often unstable. Artificial, miniature versions of these binders could be cheaper and more stable. But designing them to bind specific targets has been difficult.



A miniprotein (dark gold) binds influenza's hemagglutinin protein.

When antibodies bind to a protein, it's like a rock climber trying to scale a sheer cliff face, explains David Baker, a computational structural biologist at the University of Washington (UW), Seattle: The surface is mostly smooth with few reliable hand- and footholds. Antibodies are large, so they can snag many weak holds simultaneously, which collectively allow them to hold firm to a target.

Miniproteins have fewer options. Researchers have tried to overcome this deficiency by identifying hot spots on proteins—strong handholds—and then building miniproteins around those holds. That only works for a handful of targets studied enough that their hot spots have been mapped out, however.

To get around this problem, Baker and his colleagues turned to Rosetta, software they designed to predict protein structures based on their amino acid sequence (*Science*, 8 August 2008, p. 784). To find potential handholds, the team fed Rosetta experimental maps of target proteins and then instructed the program to first calculate how tightly specific amino acids would bind to different spots across the surface of a target protein. The software then looked for clusters of neighboring handholds and determined how to build a stable miniprotein that would grab onto as many holds in a cluster as possible. The researchers used the software to design tens of thousands of virtual miniproteins expected to bind well to their targets.

Baker and his colleagues tested their strategy with miniproteins designed to bind 12 targets, including proteins involved in cancer and the surface proteins of viruses such as influenza and SARS-CoV-2. Yeast turned these virtual binders into actual miniproteins that could be evaluated in the lab. After tweaking their designs, the researchers wound up with miniproteins that bound as strongly as, if not stronger than, most antibodies do, they reported last week in *Nature*.

Many of the targets play roles in disease, suggesting the minibinders could have potential as medicines. In July 2021, for example, Baker and colleagues reported an early example in a preprint. They produced miniproteins that bound to and neutralized the spike protein of SARS-CoV-2, the virus that causes COVID-19. The compounds were then shown to protect mice genetically engineered to be susceptible to SARS-CoV-2 infection. Clinical trials may begin this year, Baker says.

"We can design proteins to bind to any target," says co-author Longxing Cao of Westlake University. The resulting miniproteins could not only block disease proteins, he says, but also direct toxic payloads to cancerous tissues or ferry light-emitting or radioactive diagnostic molecules to cancer or other disease cells, to indicate the progress of treatment.

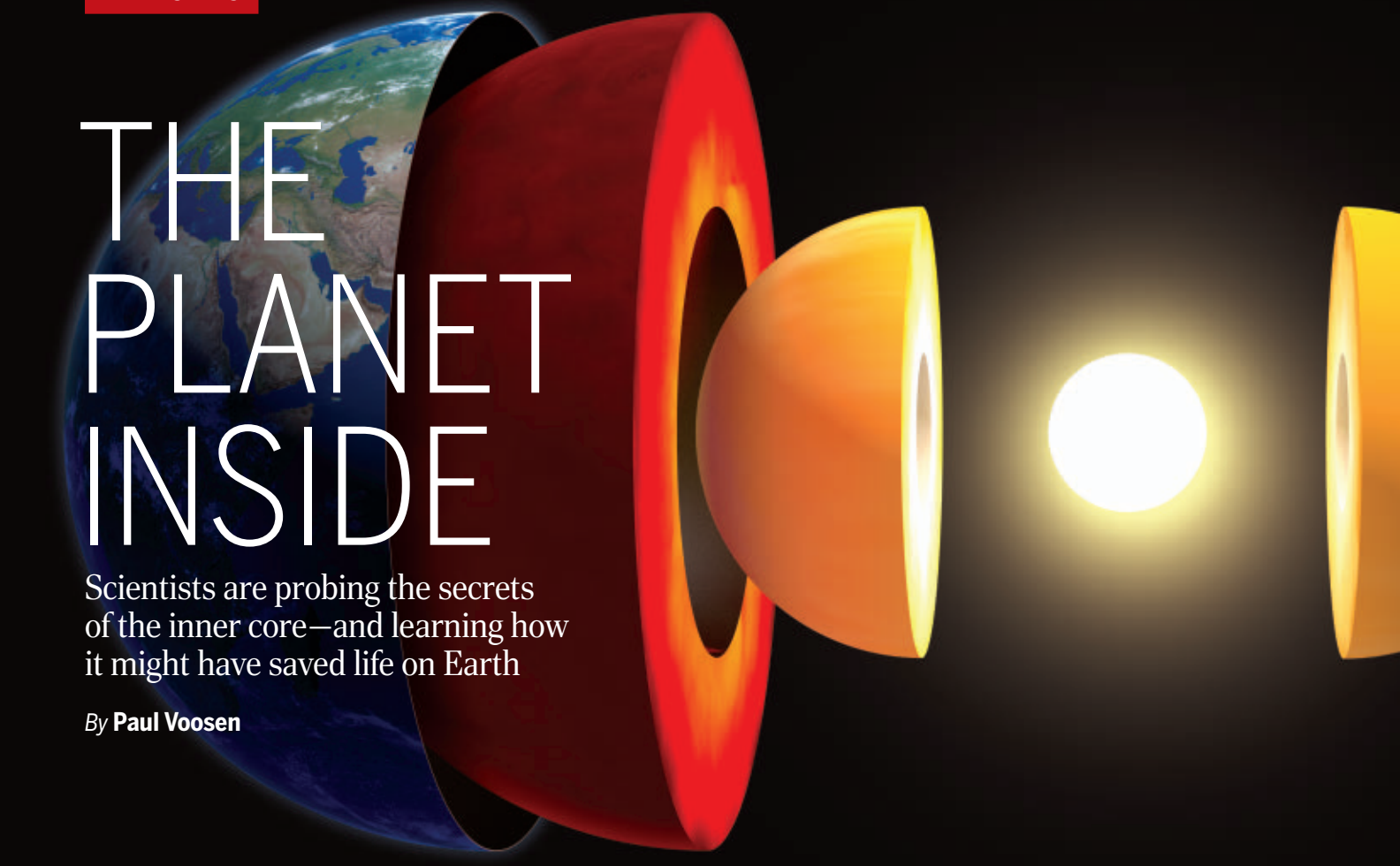
Still, Baker says, "There is room for improvement." Despite Rosetta's predictions, many miniproteins didn't stick to their target. But the software will continue to make adjustments, he says, so its designs are sure to improve over time. ■

IMAGE: IAN C. HAYDON/UW INSTITUTE FOR PROTEIN DESIGN

THE PLANET INSIDE

Scientists are probing the secrets of the inner core—and learning how it might have saved life on Earth

By **Paul Voosen**



Earth's magnetic field, nearly as old as the planet itself, protects life from damaging space radiation. But 565 million years ago, the field was sputtering, dropping to 10% of today's strength, according to a recent discovery. Then, almost miraculously, over the course of just a few tens of millions of years, it regained its strength—just in time for the sudden profusion of complex multicellular life known as the Cambrian explosion.

What could have caused the rapid revival? Increasingly, scientists believe it was the birth of Earth's inner core, a sphere of solid iron that sits within the molten outer core, where churning metal generates the planet's magnetic field. Once the inner core was born, possibly 4 billion years after the planet itself, its treelike growth—accreting a few millimeters per year at its surface—would have turbocharged motions in the outer core, reviving the faltering magnetic field and renewing the protective shield for life. "The inner core regenerated Earth's magnetic field at a really interesting time in evolution," says John Tarduno, a geophysicist at the University of Rochester. "What would have happened if it didn't form?"

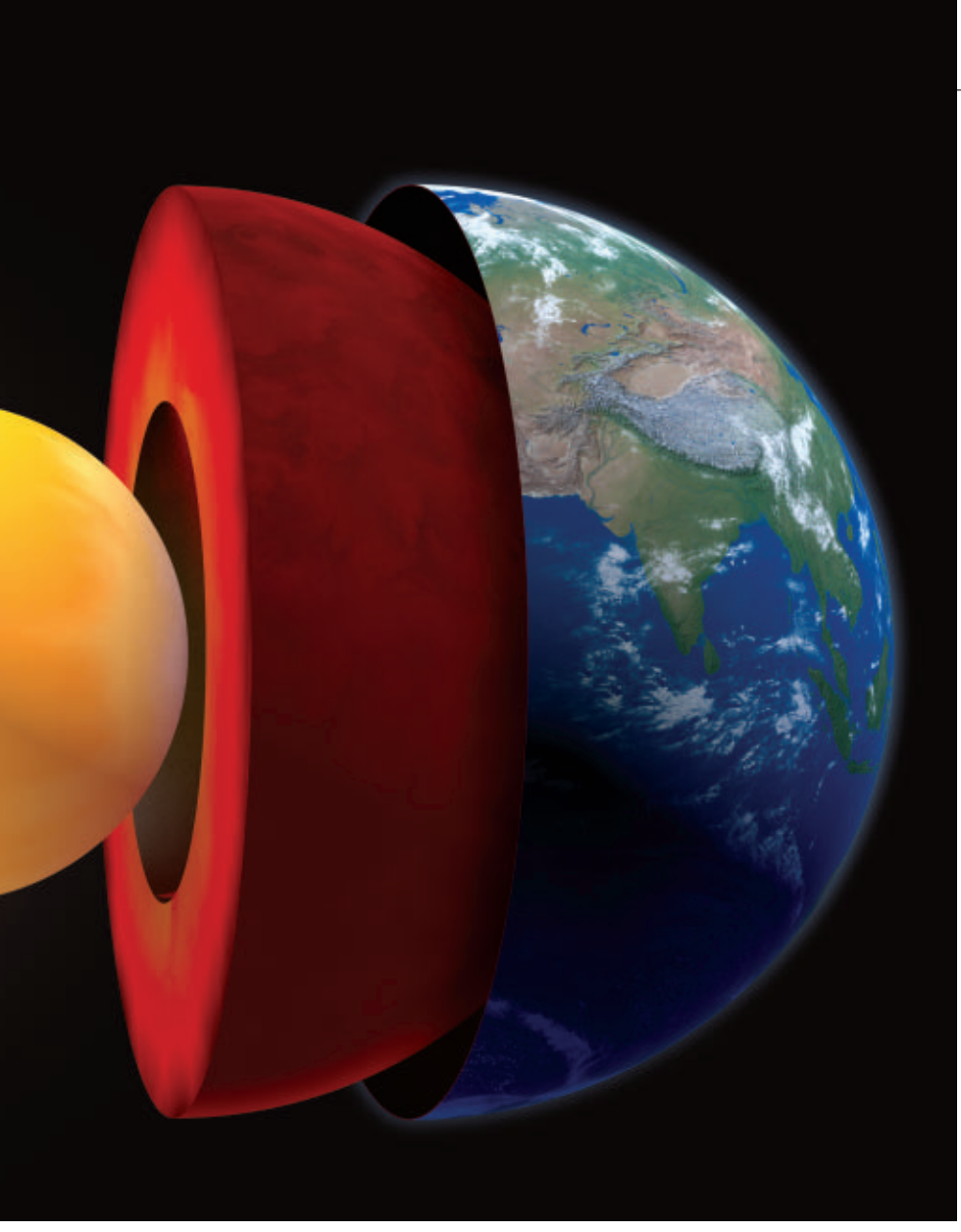
Just why and how the inner core was born at that moment is one of many lingering puzzles about the Pluto-size orb 5000 kilometers underfoot. "The inner core is a planet within a planet," says Hrvoje Tkalčić, a seismologist at Australian National University (ANU)—with its own topography, its own spin rate, its own structure. "It's beneath our feet and yet we still don't understand some big questions," Tkalčić says.

But researchers are beginning to chip away at those questions. Using the rare seismic waves from earthquakes or nuclear tests that penetrate or reflect off the inner core, seismologists have discovered it spins independently from the rest of the planet. Armed with complex computer models, theorists have predicted the structure and weird behavior of iron alloys crushed by the weight of the world. And experimentalists are close to confirming some of those predictions in the lab by re-creating the extreme temperatures and pressures of the inner core.

Arwen Deuss, a geophysicist at Utrecht University, feels a sense of anticipation that may resemble the mood in the 1960s, when researchers were observing seafloor spreading and on the cusp of discovering plate tectonics, the theory that makes sense of Earth's surface. "We have all these observations now," she says. It's simply a matter of putting them all together.

THE ANCIENTS thought Earth's center was hollow: the home of Hades or hellfire, or

ILLUSTRATION: C. BICKEL/*SCIENCE*



a realm of tunnels that heated ocean waters. Later, following erroneous density estimates of the Moon and Earth by Isaac Newton, Edmond Halley suggested in 1686 that Earth was a series of nested shells surrounding a spinning sphere that drove the magnetism witnessed at the surface.

Basic tenets of planet formation provided a more realistic picture. Some 4.5 billion years ago, Earth was likely born from the collisions of many asteroidlike "planetesimals." The dense iron in the planetesimals would have sunk to the core of the molten proto-Earth, while lighter silicate rocks rose like oil on water to form the mantle. At temperatures of thousands of degrees and millions of atmospheres of pressure, the core would have remained molten, even as Earth's mantle and crust cooled and hardened.

Early 20th century seismologists confirmed that view with a key bit of evidence: an earthquake shadow. When an earthquake strikes, the rupture emits primary, or pressure, waves (P waves) that ripple out in all directions. Secondary, or shear, waves (S waves) follow. For large earthquakes, seismologists were able to detect P waves on the other side of the planet, after they were bent and refracted by Earth's interior layers. But strangely, S waves were missing. That only made sense if the iron core was liquid, because liquids lack the rigidity that allows S waves to sashay through.

It wasn't until the early 1930s that Inge Lehmann, a pioneering Danish seismologist, noticed another breed of P waves that showed the core was not entirely liquid. These waves arrived at angles that were only possible if they had bounced off something dense. By 1936 she had deduced the existence of a solid inner core, ultimately measured to be about 2440 kilometers in diameter: the planet inside.

THE SOUTH SANDWICH ISLANDS are inhospitable volcanic crags in the far southern Atlantic Ocean. They are also earthquake factories, thanks to the nearby subduction of the South American tectonic plate. Seismologists like them for another, geometric reason: Earthquake waves that rocket from the islands to a lonely seismic station in Alaska shoot straight through the inner core.

Nearly 30 years ago, Xiaodong Song and Paul Richards—both seismologists then at Columbia University—thought they could use those waves to get a handle on the spin of the inner core, which, suspended in liquid, is under no obligation to rotate in sync with the rest of the planet. Combing through archival seismic records, they looked for subtle variations in the travel times of P waves for several dozen South Sandwich earthquakes over the course of decades. Their travel times through the outer core and mantle stayed constant, as expected. But with each passing year, P waves going through the inner core sped up a bit. "It was delicate, but you could see the changes," Song says.

There was only one way he and Richards could account for this puzzling trend: The inner core was rotating faster than the rest of the planet, by about 1° per year. This superrotation was gradually realigning the seismic wave paths with a north-south axis in the inner core known to boost P wave speeds. Every 400 years, they suggested in a 1996 *Nature* paper, the inner core made an extra revolution inside Earth.

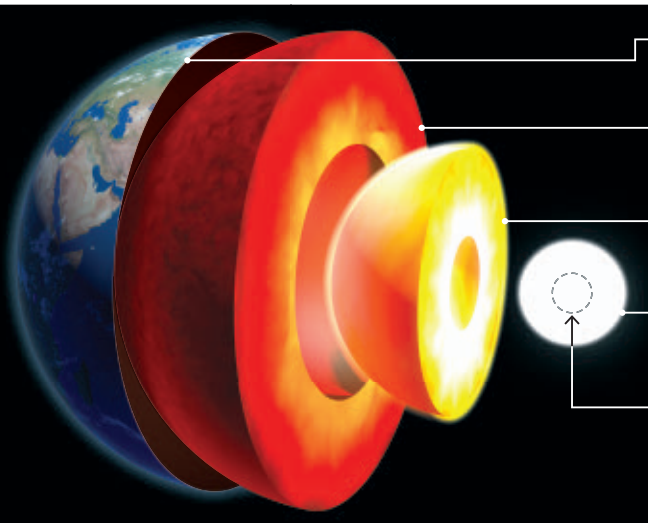
A few years later, John Vidale, a seismologist now at the University of Southern California, validated the result using a slightly different method. Vidale specializes in using records from the Large Aperture Seismic Array (LASA), a U.S. Air Force facility in Montana, closed in 1978, that operated more than 500 sensors in deep boreholes to detect atomic bomb tests. "It's still the best data, better than the best arrays today," he says. Seismic waves from nuclear tests were ideal because, unlike earthquakes, the source can be precisely located.

Vidale used the waves from two Soviet underground bomb tests detonated in 1971 and 1974 beneath Novaya Zemlya, a remote Arctic archipelago. Instead of looking for waves that passed through the inner core, as Song and Richards did, Vidale chose ones that ricocheted off it, registering its spin like the beam of a radar gun. "We could see one side of the inner core getting closer, and one side getting further away," he says.

He found that over the 3 years between the tests, the inner core rotated 0.15° per year faster than the rest of the planet—much less than Song's first estimate. But subsequent work by Song in 2005, using 18 pairs of South

Iron heart

Earth's solid inner core, buried 5000 kilometers below our feet, has remained enigmatic since its discovery nearly 100 years ago. About the size of Pluto and growing several millimeters every year, it helps power Earth's magnetic field. It also possesses a strange interior structure that is only now coming into view with advances in seismology.



Crust
Life sits on a layer of rock that is vanishingly thin compared with the rest of the planet.

Mantle
Earth's thickest layer is made of 3000 kilometers of sticky silicate rock.

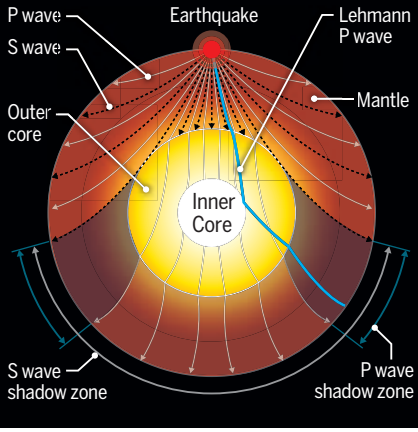
Outer core
The molten iron outer core was born along with Earth 4.5 billion years ago.

Inner core
At 6000°C and 3 million atmospheres of pressure, the inner core is solid iron but soft.

Innermost inner core
At the core's center is an off-center globe with odd seismic characteristics.

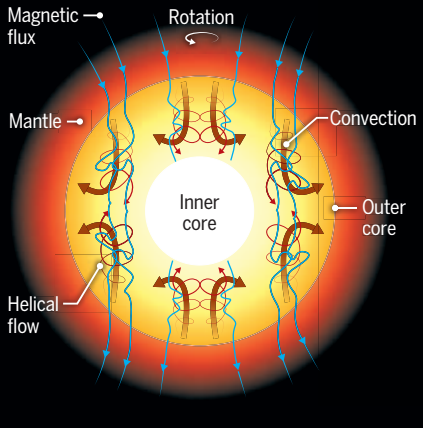
A mysterious reflection

Earthquake pressure (P) and shear (S) waves refract as they pass through Earth, but the liquid outer core stymies S waves. In 1936, Inge Lehmann discovered P waves in a shadow zone associated with an entirely molten core—only possible if the waves were reflections off a solid sphere rather than refractions.



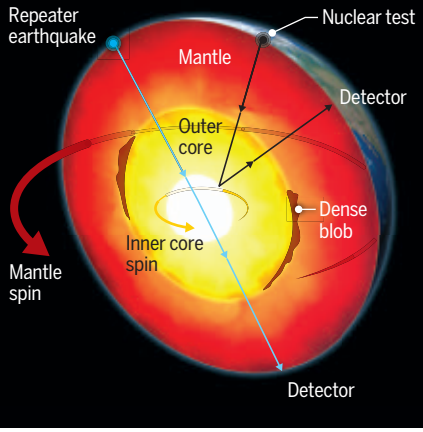
Magnetic driver

Earth's magnetic field, which protects life from radiation, is driven by convective motions in the molten outer core. Growth of the inner core turbocharges those motions. As iron crystallizes, it spits out light elements like oxygen or silicon, which rise toward the mantle, dragging iron with them.



Erratic spinner

Waves from repeating earthquakes and nuclear tests have shown that the inner core does not rotate in sync with the rest of the planet. Some researchers believe gravitational tugs from dense blobs at the bottom of the mantle could be responsible for the erratic spinning.



GRAPHIC: C. BICKEL/*SCIENCE*

Sandwich earthquakes that repeated in the same spot over the span of decades, lined up with Vidale's reduced estimate.

The discovery of the inner core's superrotation shocked many geophysicists, who had assumed it spun at the same rate as the mantle. It also tantalized them. The rotation could offer clues to how the inner core couples to the outer core and influences the magnetic dynamo. Some thought it could even help explain why Earth's magnetic poles wander and flip from time to time.

But almost as rapidly as this picture of the inner core's spin emerged, it grew more complicated and more mysterious. "What we thought 10 years ago isn't holding together," Vidale says.

RECENTLY, SONG, now at Peking University, decided to revisit his rotation work. His postdoc, Yi Yang, compiled the world's most extensive database of repeating earthquakes, with sources not just in the South Sandwich Islands, but also in places like Chile and Kazakhstan. Analyzing more than 500 source-detector pairs with a range of paths through the core, Song and Yi found that the superrotation stopped all at once a decade ago, and since then the inner core has rotated at the same speed as the mantle. The changes "all disappear at the same time," says Song, who presented the work at a meeting of the American Geophysical Union (AGU) late last year.

Meanwhile, Vidale was trying to push his trend further back in time using LASA data. He focused on two bomb-induced earthquakes, both set off by the U.S. government underneath the far end of Alaska's Aleutian Islands, in 1969 and 1971. The tests were controversial; the second, Cannikin, at 5 megatons, was the largest ever U.S. underground test, and it faced opposition from environmental activists who chartered a fishing ship, christened it *Greenpeace*, and sailed it to the island in protest. Despite appeals to the Supreme Court, the test went as planned, creating a crater lake at the island's surface even though the detonation was 1800 meters down.

The two tests created another, much delayed splash last year at the AGU meeting. Vidale reported that waves from the detonations revealed not superrotation, but subrotation: During the time between the two U.S. tests, the inner core rotated more slowly than the rest of the planet, by some 0.05° per year. Yet by the time of the Soviet tests, the inner core had somehow reversed course and sped up. The "observations are really amazing," says Barbara Romanowicz,

a seismologist at the University of California (UC), Berkeley.

For Vidale, the pattern from 1969 to 1974, from slow to fast, may indicate a fundamental rhythm of the inner core. For decades, radio astronomers have tracked minute changes in Earth's surface rotation—the length of a day—against a cosmic reference frame: the fixed position of distant cosmic beacons called quasars. Although most of the yearly jitter is due to events like hurricanes and earthquakes, a tiny-but-regular 6-year wobble in day length has emerged. "Nobody has been able to say what causes it," says Benjamin Chao, a geodesist at Academia Sinica. "But everybody bets on the core."

Chao says one possible explanation for the 6-year cycle is gravitational interactions between the mantle and inner core. The inner core is likely to be lumpy, with hills hundreds of meters high, and at the bottom of the mantle, seismologists have discovered two ultradense, continent-size blobs. The tugs of the blobs on the hills could create a loose coupling between the mantle and the inner core—enough to "pull the inner core back and forth" in cycles of superrotation and subrotation, Chao says.

Song, however, only sees a slowdown, with no sign of an oscillation. He ties his record to a longer term trend in the length of a day, which saw the planet spin progressively faster from the 1970s before settling down in the early 2000s. Song thinks gravitational tugs from the mantle might have pulled the inner core along, but with a lag.

Given that neither finding has yet been published, it's hard to say how they fit together. "Is everybody right? Is everybody wrong?" Romanowicz asks. Either way, varying rotation seems more plausible than constant superrotation, says Miaki Ishii, a seismologist at Harvard University. "It makes more sense than what we have right now."

THE INNER CORE is the most metal place on Earth—even more so than the outer core. Both are made mostly of iron, along with a smattering of nickel. But the iron is thought to also contain traces of lighter elements like oxygen, carbon, and silicon. As the iron crystallizes on the growing surface of the inner core, it spits out some of those elements, leaving behind almost pure iron, much as ice freezing from a bucket of saltwater expels the salt and becomes largely fresh. The expelled elements, lighter than iron, rise and sweep along the surrounding liquid, driving up to 80% of the convection that generates Earth's magnetic field.

The nature of the iron left behind is the subject of ongoing debate. Iron atoms at Earth's surface—in your cast iron skillet, for example—pack themselves in cubic arrangements. But when tiny samples of iron are compressed between two diamonds to inner core–like pressures, the atoms rearrange into hexagons. The hard question is what happens when iron is simultaneously squashed and heated to thousands of degrees, says Lidunka Vočadlo, a computational mineral physicist at University College London. These conditions are difficult to re-create in the lab, because carbon in the diamonds often contaminates the iron when the apparatus is heated. But in computer models, Vočadlo says, "There's no limit to the pressure and temperature you can get."



In 1971, a 5-megaton nuclear bomb was lowered into a borehole in Alaska. Seismic waves from the blast bounced off the inner core, helping gauge its spin.

PHOTO: UNITED STATES ATOMIC ENERGY COMMISSION

Modeling by Vočadlo and her collaborators suggests hexagonal packing is the most stable arrangement under inner core conditions. The models also find that pure iron grows soft when it sits at 98% of its melting point, as it may throughout much of the inner core. This "premelting effect," as it is called, could explain why S waves travel much slower than expected in the supposedly solid inner core.

The story isn't closed for cubic iron, however. Just as water must cool below freezing before ice can nucleate, researchers have suggested iron can't solidify directly into its hexagonal form unless it is nearly 1000 K cooler than the inner core. Atom-scale modeling published early this year by a team led by Yang Sun, a mineral physicist at Columbia, suggests a solution: Iron accreting onto the inner core could first crystallize into its cubic form before transitioning into a hexagonal end state.

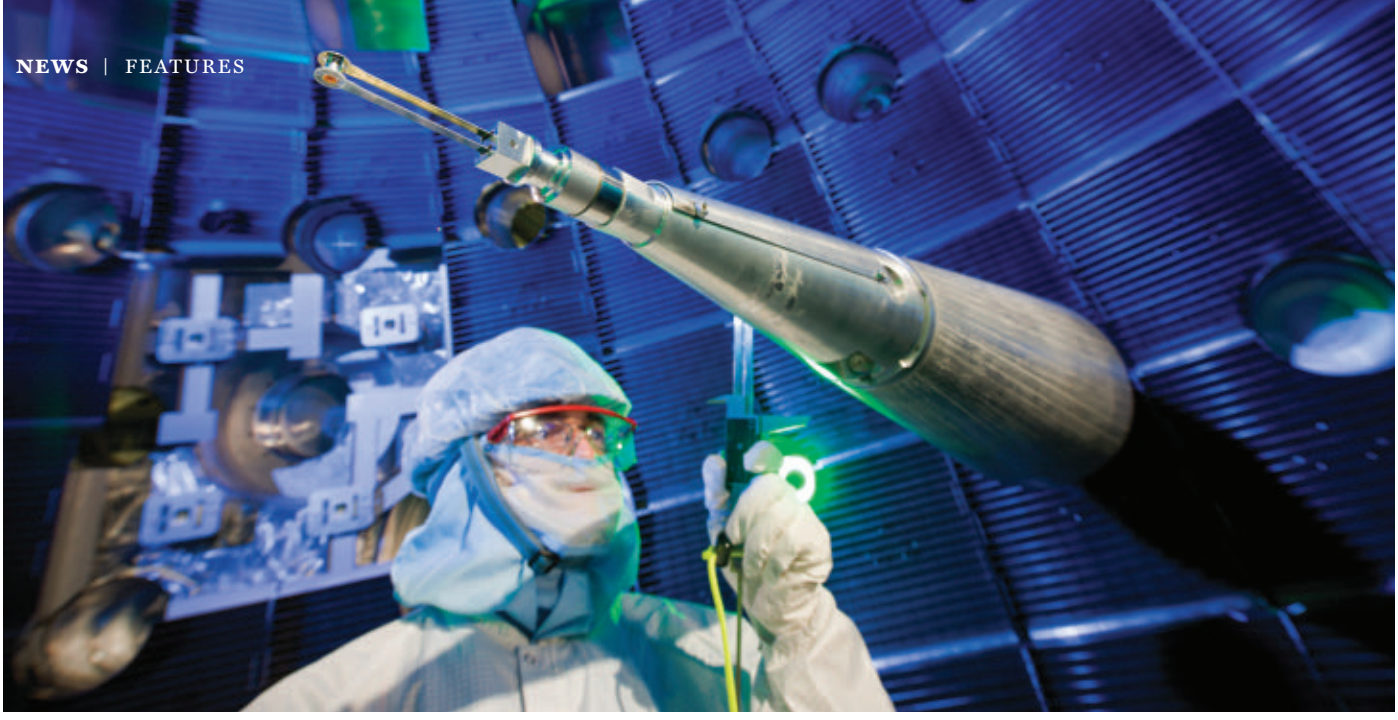
Although the cubic versus hexagonal debate may seem academic, the structure may determine how the iron crystals align, how much nickel and other light elements can mix with the iron, how much heat it releases on crystallization, and even its melting point. "The fundamental properties of iron change depending on what phase you're in," Vočadlo says.

A new wave of lab studies may help settle the question. After years of halting progress, researchers are on the verge of regularly re-creating and observing inner core conditions. One strategy is to press and heat iron in diamond anvil cells, as before—but to glimpse its structure, quickly, before it is contaminated with carbon. New, powerful x-ray light sources such as the Extremely Brilliant Source at the European Synchrotron Radiation Facility, which turned on in 2020, can take that kind of flash photo.

Another is to harness the massive lasers of the National Ignition Facility (NIF) at Lawrence Livermore National Laboratory (LLNL), which are typically aimed at pellets of hydrogen isotopes to spark tiny nuclear fusion reactions. In a study published earlier this year, NIF researchers instead turned some of those beams on iron, heating and pressurizing it to levels far beyond those seen in Earth's core. Each time they examined the iron's structure with an x-ray, it came out the same—as hexagonal iron, says Richard Kraus, an LLNL research scientist who led the study.

A third tack to re-create the inner core is through shock wave experiments. Jung-Fu Lin, an experimental mineral physicist at the University of Texas, Austin, has partnered with researchers in China who use bursts of gas to fire projectiles into iron at speeds 10 times faster than a rifle bullet, generating corelike temperatures and pressures. They are already seeing hints of the premelting effect identified by Vočadlo and predicted by others. If the results hold up, they may suggest the "solid" inner core isn't so solid after all. "It's like a smoothie," Lin says. "Very soft."

If the inner core is a mystery, then the "innermost" inner core is a riddle wrapped in a mystery. Since the 1980s, seismologists have known that seismic waves run faster through the inner core along a north-south axis, perhaps because the iron crystals have a common alignment, presumably along the prevailing direction of Earth's



Experiments at the National Ignition Facility have focused high-power lasers on small samples of iron to re-create inner core conditions.

magnetic field. But in 2002, Ishii and Adam Dziewoński, also at Harvard, discovered that within a sphere roughly 600 kilometers across, that fast lane is tilted by 45°. Ishii says that anomaly could be a relic of an ancient, tilted magnetic field or a kernel of cubic rather than hexagonal iron. No matter what, she says, "There's something different going on at the center of the Earth."

Researchers are poised to turn these hints into something more rigorous. Over the past decade, a clutch of high-quality seismometers has been erected in Antarctica, allowing researchers to catch far more earthquake waves that pass through the inner core's north-south fast lanes. Armed with the improved resolutions provided by these waves and many others globally, Utrecht's Deuss and her graduate student Henry Brett used a supercomputing-based technique to create the first 3D view of the inner core—a bit like a CT scan in the hospital.

This work, set for publication soon, confirms the existence of the innermost core, but finds it is slightly offset from the planet's center. It also reveals speed differences between the fast lanes seen in the inner core's western and eastern hemispheres. That suggests the story of the fast lanes is more complicated than iron crystals aligning with the dominant magnetic field, which would have a more uniform signal. It's still early days, similar to where imaging of the mantle was in the 1980s, but Brett says more detailed models are coming soon. "We're going to be able to ask more interesting questions."

ALL THIS COMPLEXITY appears to be geologically recent. Scientists once placed the inner core's birth back near the planet's formation. But a decade ago, researchers found, using diamond anvils at outer core conditions, that iron conducts heat at least twice as fast as previously thought. Cooling drives the growth of the inner core, so the rapid heat loss combined with the inner core's current size meant it was unlikely to have formed more than 1 billion years ago, and more than likely came even later. "There's no way around a relatively recent appearance of the inner core," says Bruce Buffett, a geodynamicist at UC Berkeley.

> **"The dynamo could have been close to dying."**
>
> **Peter Driscoll**,
> Carnegie Institution for Science

Tarduno realized rocks from the time might record the dramatic magnetic field changes expected at the inner core's birth. Until recently, the paleomagnetic data from 600 million to 1 billion years ago were sparse. So Tarduno went searching for rocks of the right age containing tiny, needle-shaped crystals of the mineral titanomagnetite, which record the magnetic field's strength at the time of their crystallization. In a 565-million-year-old volcanic formation on the north bank of the St. Lawrence River in Quebec, his team found the crystals—and convincing evidence that the magnetic field of the time was one-tenth the present day strength, they reported in 2019. The fragility of the field at the time has since been confirmed by multiple studies.

It was probably a sign that rapid heat loss from the outer core was weakening the convective motions that generate the magnetic field, says Peter Driscoll, a geodynamicist at the Carnegie Institution for Science. "The dynamo could have been close to dying," he says. Its death could have left Earth's developing life—which mostly lived in the ocean as microbes and protojellyfish—exposed to far more radiation from solar flares. In Earth's atmosphere, where oxygen levels were rising, the increased radiation could have ionized some of this oxygen, allowing it to escape to space and depleting a valuable resource for life, Tarduno says. "The potential for loss was gaining."

Just 30 million years later, the tide had turned in favor of life. Tarduno's team went to quarries and roadcuts in the Wichita Mountains of Oklahoma and harvested 532-million-year-old volcanic rocks. After analyzing the field strength frozen in the tiny magnetic needles, they found that its intensity had already jumped to 70% of present values, they reported at the AGU meeting. "That kind of nails it now," Tarduno says. He credits the growth of the inner core for the field jump, which he says is "the true signature of inner core nucleation."

Around the same time, life experienced its own revolution: the Cambrian explosion, the rapid diversification of life that gave rise to most animal groups and eventually led to the first land animals, protomillipedes that ventured onto land some 425 million years ago.

It just may be that the clement world they found owes much to the inner iron planet we'll never see, 5000 kilometers below. ■

PHOTO: NATIONAL IGNITION FACILITY

CALL FOR NOMINATIONS FOR THE 2023 HFSP NAKASONE AWARD

The HFSP Nakasone Award is awarded to scientists in recognition of pioneering work that has moved the frontier of the life sciences. This may encompass conceptual, experimental, or technological breakthroughs. The award recognizes the vision of former Prime Minister Nakasone of Japan in the creation of the Human Frontier Science Program (HFSP).

The competition is open; it is not limited to HFSP awardees and there is no age limit for candidates. In selecting the awardee, the HFSPO Council of Scientists will pay particular attention to recent breakthroughs by early career scientists. The awardee will receive a commemorative gold medal, a small unrestricted research grant and an invitation to deliver the HFSP Nakasone Lecture at the 2023 HFSP Awardees Meeting.

The deadline for nominations is 6 May 2022.

The nomination form and further information are available at:

www.hfsp.org/awardees/hfsp-nakasone-award

International Human Frontier Science Program Organization

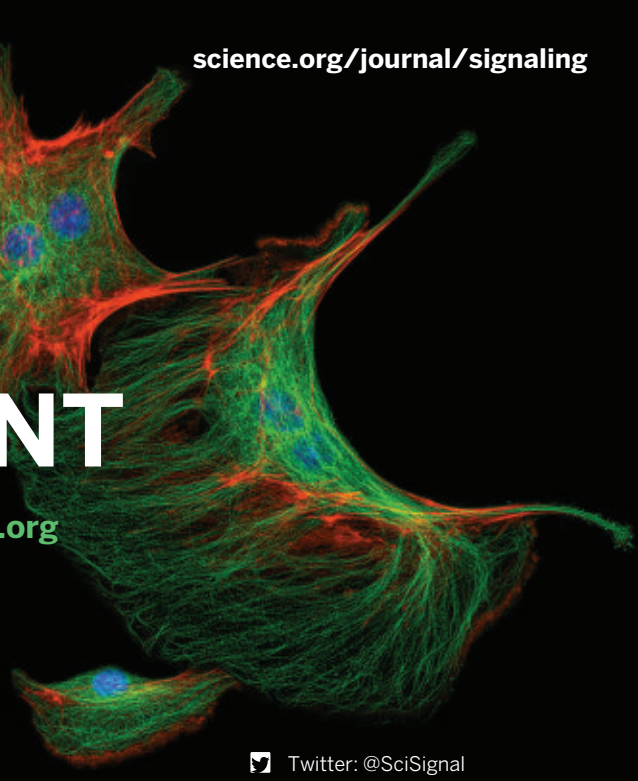
PREVIOUS WINNERS

Aviv Regev (2022)
F. Ulrich Hartl &
Arthur Horwich (2022)
Anthony Hyman &
Clifford Brangwynne (2021)
Angelika Amon (2020)
Michael Hall (2019)
Svante Pääbo (2018)
David Julius (2017)
Emmanuelle Charpentier &
Jennifer Doudna (2016)
James Collins (2015)
Uri Alon (2014)
Stephen Quake (2013)
Gina Turrigiano (2012)
Michael Elowitz (2011)
Karl Deisseroth (2010)

science.org/journal/signaling

PUT YOUR RESEARCH OUT IN FRONT

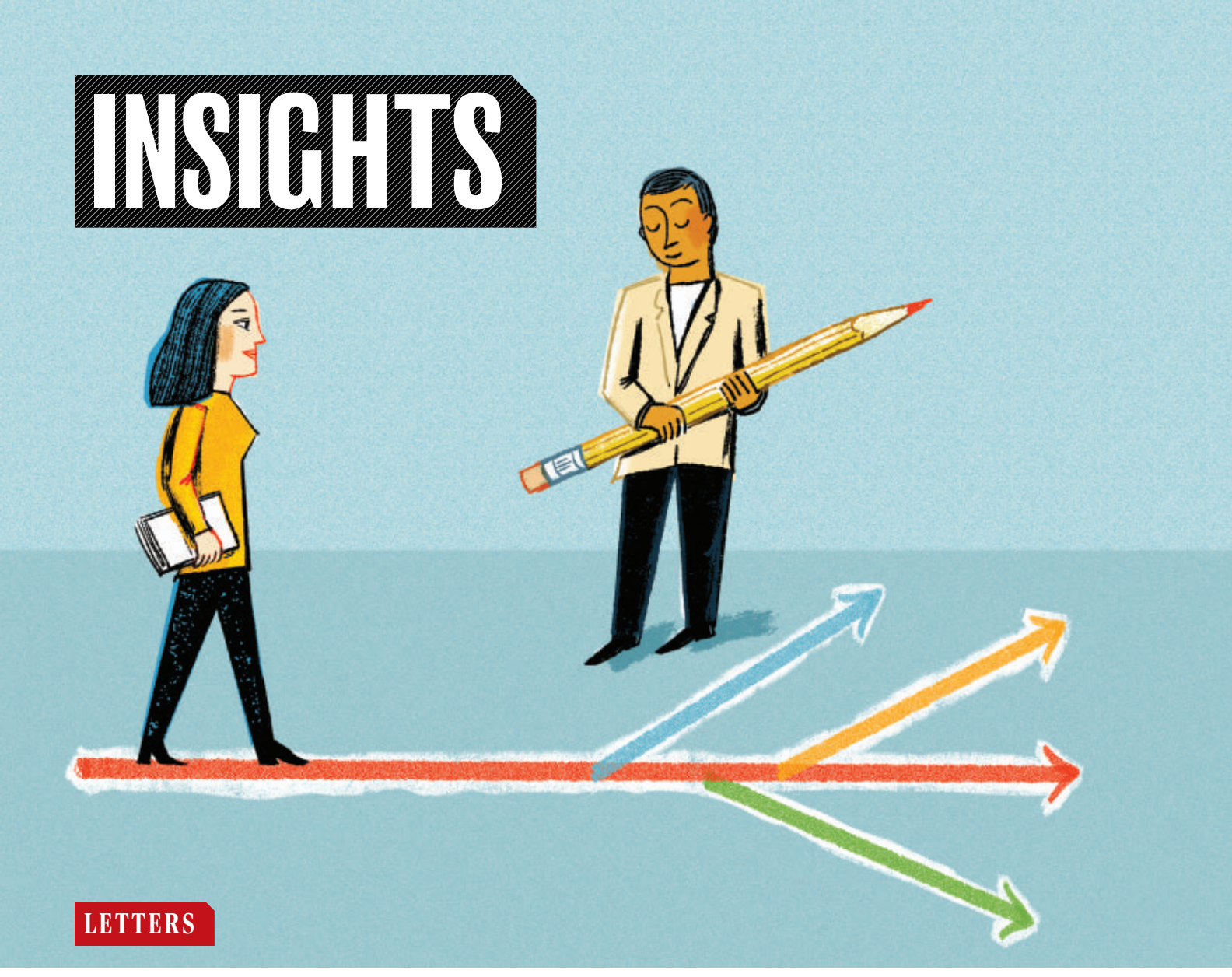
Submit your research: **cts.ScienceMag.org**



Science Signaling
AAAS

Twitter: @SciSignal
Facebook: @ScienceSignaling

INSIGHTS



LETTERS

NEXTGEN VOICES

Rules all PIs should follow

We asked young scientists to write a rule that all principal investigators (PIs) should be required to follow to improve the experience of young scientists in their lab. Read a selection of their suggestions below. Follow NextGen Voices on Twitter with hashtag #NextGenSci. **—Jennifer Sills**

Encourage debate

Admit when you are wrong and encourage trainees to do the same. Science provides many opportunities to be wrong, whether it be disproving a favorite hypothesis or forgetting a step in a protocol. But young scientists often have a fear of failure or worry about facing repercussions for their mistakes. When PIs openly admit their missteps, they model how disproven hypotheses and mistakes are actually opportunities for learning and growth. Transparency about setbacks also empowers trainees to speak up with opposing opinions, making for richer discourse.

Jennifer S. Chen
Yale School of Medicine, New Haven, CT 06510, USA. Email: jennifer.s.chen@yale.edu

Take students' questions seriously and answer them constructively. PIs who show students this respect will stimulate their research motivation by demonstrating that the lab is a safe environment to express opinions.

Chih Ying Huang
Department of Physics, National Cheng Kung University, Tainan 70101, Taiwan (Republic of China). Email: magic0910@gmail.com

Give students the freedom to express their views. Encouraging all lab members to share their ideas will help establish a friendly environment in the lab, allowing young scientists to enjoy the research process and grow intellectually.

Shantanu Lanke
Maharashtra 422101, India.
Email: slanke@asu.edu

Be humble. When PIs interpret data, trainees may not feel comfortable sharing an alternative opinion. By reminding trainees that a PI's intuition may be wrong, the

ILLUSTRATION: ROBERT NEUBECKER

PI can encourage young scientists to cultivate their curiosity, imagine new possibilities, develop original ideas, and judge results objectively.

Michael Sanjay Fernandopulle
Medical Scientist Training Program, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA.
Email: michael.fernandopulle@northwestern.edu

Facilitate learning

Invite at least one expert from another field to give a lecture to the research group every month. Learning about other disciplines will broaden young researchers' horizons, stimulate their thinking, and introduce them to different research methods.

Yongsheng Ji
Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230026, China. Email: jiys2020@ustc.edu.cn

After an experiment, model the correct way to clean the lab and document safety records. By doing this work together, PIs can increase young scientists' safety awareness, cultivate careful and responsible methods, and foster a team spirit among the group.

Yuan Zhi
School of Economics, Guizhou University, Guiyang, Guizhou 550025, China.
Email: yzhi@gzu.edu.cn

Hold monthly lab meetings during which PhD students present their work. These presentations will allow young researchers to gain experience in data representation and science communication and give the PI an opportunity to monitor progress and offer relevant feedback.

Sara Granado Rodríguez
Department of Biology, Universidad Autónoma de Madrid, Madrid, Madrid, 28049, Spain.
Email: sara.granado12@gmail.com

Promote cooperation instead of competition between students within your own lab and between labs. This strategy will create a supportive learning environment for students, enhance their productivity, and expand their skill set through shared knowledge and equipment.

Andrea Y. Frommel
Faculty of Land and Food Systems, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. Email: andrea.frommel@ubc.ca

Foster independence

When making a strategic project decision, demonstrate how the decision is in the best interest of the young researchers involved. Shifting the burden of proof to the PI would automatically give leverage to the young researcher in negotiating the project's direction and give them the power to veto decisions that only benefit the PI. If respect for the young researcher's career prospects were a standard criterion in decision-making, exploitative practices would become untenable.

Martin Lukačišin
Faculty of Medicine, Technion, Haifa 3525433, Israel. Twitter: @CombinationBio

If you move from one institution to another, help lab members find new positions. When PIs accept job offers, young scientists in the lab they are leaving face substantial challenges. PIs should provide support to all lab members as they find new positions, even if they don't plan to follow the PI to the new institution.

Yida Zhang
Department of Biomedical Informatics, Harvard University, Brookline, MA 02445, USA. Email: yida_zhang@hms.harvard.edu

Encourage young scientists to do what's best for their budding career, not what's best for your established career. Supervisors should support and mentor their students and postdocs to help them build the skills required to get jobs and navigate their own career path.

Christina N. Zdenek
School of Biological Sciences, The University of Queensland, St Lucia, QLD 4072, Australia.
Twitter: @CNZdenek

Bring your students to academic conferences. Building a professional network is crucial for young scientists, and PIs can facilitate connections to senior scientists.

Xiao-Yu Wu
Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. Twitter: @XiaoYuWu10

Provide students with opportunities to think creatively. Giving lab members the freedom to explore their unique strengths and encouraging them to think about problems from different perspectives will empower them and improve the quality of their research.

Senthilkumar Seenuvasaragavan
Mumbai 400076, India.
Email: shrisenjohn@gmail.com

Provide positive reinforcement

Encourage your students unconditionally. Young scientists may give up if they receive no positive reinforcement for their hard work. PIs should let them try new things. If they succeed, PIs can give them affirmation. If they fail, PIs can remind them that the experience is a lesson that will help them improve.

Yan Zhuang
Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.
Email: zhuangyanisme@pku.edu.cn

Offer honest words of encouragement. Research is difficult, failure frequent, and feelings of inadequacy rampant among young scientists. Words of encouragement from a PI can be the difference between a student giving up or pushing through yet another experiment.

Cathrine Bergh
KTH Royal Institute of Technology, 11428 Stockholm, Sweden. Email: cabergh@kth.se

Recognize young scientists' work both publicly and privately. By publicly giving lab members credit as well as privately communicating appreciation for their contributions, PIs can ensure that everyone in the lab feels respected.

Jaime Coulbois
Departamento de Ciencia Política y Relaciones Internacionales, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid, Spain. Twitter: @CoulboisJaime

Create a supportive environment

Support the individual goals and aspirations of each lab member. When PIs acknowledge each person's unique perspective, they can encourage cooperating and celebrating each other's successes, creating an environment where everyone can flourish.

Salam Salloum-Asfar
Qatar Biomedical Research Institute, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar.
Twitter: @Dr_SalamSalloum

Organize a homemade group dinner at least once a month. My mentor showed us that cooking dinner is just like doing experiments. While cooking, our group learned the spirit of collaboration. While eating, we talked freely about academic ideas or the progress we had made. These valuable experiences allowed us to relax and feel closer to one another.

Bo Cao
Core Research Laboratory, The Second Affiliated Hospital, School of Medicine, Xi'an Jiaotong University, Xi'an, Shaanxi 710004, China.
Email: bocao@vip.qq.com

Treat others as you would have liked to be treated—not how you were treated! Sometimes PIs believe that if they survived an unpleasant training environment, then their students should, too. However, the best PIs try to improve the training experience by working together

with young scientists to overcome challenges. If everyone were required to support members of their lab, the culture of academia would improve.

Katherine Davis
Imperial College London, London, UK.
Twitter: @kd_katdavis

Be responsive to the unique cultural needs, such as ethnicity, nationality, gender identity, sexual orientation, health, and socioeconomic status, of your mentees. Culturally responsive mentorship provides mentees with more opportunities to succeed and a greater sense of comfort and safety.

Fernanda Oda
Department of Applied Behavioral Science, University of Kansas, Lawrence, KS 66045, USA.
Email: oda@ku.edu

Solicit feedback and act to address it. Once a year, the PI should request anonymous feedback from all lab members about the lab's culture, atmosphere, science, and integrity. The next year, the PI should report to both the lab group and an independent committee of peers what steps were taken to address concerns.

Nikos Konstantinides
Université Paris Cité, Centre National de la Recherche Scientifique, Institut Jacques Monod, F-75013 Paris, France. Twitter: @nkonst4

Respect young scientists' time

Avoid contacting lab members by instant messaging applications outside of business hours. After work hours, PIs should respect their students' time by using email, which alleviates the pressure to respond right away.

Liping Zhang
Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China.
Email: zhanglp@sustech.edu.cn

Encourage students to finish their academic degree on schedule without extensions. PIs have an incentive to keep their PhD candidates for extra time in hopes of getting more data or finishing a project while paying a lower stipend than a postdoc would demand. Such delays adversely affect students' careers and may make them feel frustrated. Moreover, desperation to generate more experimental data could potentially lead to inappropriate experiments or fraud.

Name withheld
Calgary, Canada.

Refrain from contacting lab members on weekends. Young scientists need a chance to reflect, rejuvenate, and contextualize their work. Good science is driven by inspiration, and time away from the work environment is important for both the young scientist and the PI to plan their work, identify gaps and opportunities, and continually assess long-term goals.

Divyansh Agarwal
Harvard Medical School, Massachusetts General Hospital, Boston, MA 02114, USA.
Email: dagarwal@mgh.harvard.edu

Encourage attainable thesis milestones. PIs should help young scientists determine when abandoning a project is in their best interest.

Joseph Nicholas Rainaldi
Department of Biomedical Science, University of California, San Diego, San Diego, CA 92122, USA.
Email: jrainaldi@ucsd.edu

Communicate regularly

Talk to your students regularly and listen to their concerns. Frequent communication between PIs and young scientists facilitates immediate feedback, prevents studies from getting stuck, and helps students cope with failures. The PI must listen carefully, acknowledge the students' opinions, and give them the space to have their own ideas. Ongoing dialogue will help students develop critical thinking skills and allow them to start thinking independently.

Jan Kadlec
Department of Brain Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel.
Email: jan.kadlec@weizmann.ac.il

Recognize that tasks you find trivial or routine, such as publishing a paper or finding out whether funding will be approved, may seem crucial to the people you supervise. Instead of letting emails about such issues languish, PIs should acknowledge the importance of the subject and provide a status update to prevent unnecessary stress.

Jelle Vekeman
Center for Molecular Modeling, Ghent University, 9052 Zwijnaarde, Oost-Vlaanderen, Belgium.
Email: jelle.vekeman@ugent.be

Treat each of the young scientists in your lab to lunch once a month. Food is a great equalizer; sharing a meal can make a PI more approachable. The meeting can also serve as a check-in for the PI to see how each young scientist is doing, in an informal environment where information and conversation flow more freely.

Vishal Anirudh Kanigicherla
Department of Chemistry, College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: vkanigicherla@gmail.com

Schedule regular one-to-one meetings with your PhD students. Such a routine formality will provide students with advance notice and allow them to practice presenting their work in an allotted time frame. Spontaneity does not work for everyone; students deserve the time and space to analyze, interpret, and articulate their own data before a meeting.

Kathryn Oi
Department of Chemistry, University of Zurich, ZH 8052 Zurich, Switzerland.
Email: kathryn.oi@chem.uzh.ch

Set realistic expectations

Critically read a copy of your own PhD dissertation before evaluating the work of young scientists. This exercise will remind PIs that they also had a lot to learn early in their career, allowing them to set reasonable expectations for their students. Fostering a more understanding environment could greatly improve mental health for everyone in the lab.

Kyle J. Isaacson
Ike Scientific, South Weber, UT 84405, USA.
Twitter: @kjisaacson

Actively engage in bench work. PIs who perform day-to-day experiments alongside their mentees will motivate young scientists and ensure that the PI is aware of the challenges they face. The daily presence of PIs will also break down the barriers between the mentor and mentee, which will lead to unhindered exchange of ideas and a more enjoyable and productive work environment free of unrealistic expectations from either party.

Rakesh Ganji
Department of Developmental, Molecular, and Chemical Biology, Tufts University, Boston, MA 02111, USA. Email: rakesh.ganji@tufts.edu

Clarify boundaries early. For young scientists, the "getting to know you" phase of working with a PI can be stressful. Specifying details such as when it is okay to email, how much time is needed when requesting recommendation letters, and whether unprompted doorway chats are welcome can be critical to making young scientists comfortable in a lab. Moreover, this sets a good example for young scientists to confidently set their own boundaries, which, in the long term, can help to fight burnout and imposter syndrome.

Emma Dawson-Glass
Research Department, Holden Arboretum, Kirtland, OH 44094, USA.
Email: edawson-glass@holdenfg.org

10.1126/science.abp9887



Fossils of Morganucodon—an extinct genus from the Mesozoic—revealed a major step in brain evolution in mammaliaform animals.

PALEONTOLOGY

The road to a larger brain

Ecological opportunities in the early Cenozoic favored larger, not smarter, mammals

By **Felisa A. Smith**

Modern mammals have the largest brains relative to their body size among vertebrates (*1*, *2*). Because brain size is correlated with cognitive ability, complex behavior, and puzzle-solving abilities, this suggests that mammals are better able than other animals to interface with their environment (*3*, *4*). But how and why did mammals evolve large brain sizes relative to their body mass? The question of whether the rapid morphological evolution of mammals after the Cretaceous-Paleogene extinction was accompanied by equally rapid changes in brain size speaks to the role of ecology in macroevolution. On page 80 of this issue, Bertrand *et al*. (*5*) address the idea of "brawn before brain" by characterizing the timing and pattern of mammal brain development across the Early Jurassic to the middle Cenozoic (~200 million to 30 million years ago).

The phylogenetic encephalization quotient (PEQ) is a measure of relative brain size: the ratio of actual to expected brain volume for a given body size and phylogeny. Encephalization refers to an increase in this ratio. Although the mammal lineage experienced several sudden increases in encephalization during the Mesozoic—perhaps driven by selection for enhanced hearing, smell, taste, and feeding abilities—the range of body mass and PEQ was relatively limited compared with that of modern mammals (*6*–*10*). This changed with the extinction of nonavian dinosaurs at the Cretaceous-Paleogene boundary (66 million years ago) as surviving mammals rapidly diversified to occupy newly vacated ecological niches (*10*, *11*). But was this rapid increase in body mass accompanied by an equally rapid and proportional increase in brain size (*8*)? Or did encephalization occur at a different pace? That is, did mammals first become "smarter" then larger, or larger then smarter?

Across contemporary mammals, there is a highly nonlinear scaling of body and brain size, with a strong phylogenetic signal (*2*, *4*). Slopes and intercepts vary substantially among mammalian orders, with groups such as carnivores exhibiting steeper slopes than those of ungulates (*2*, *4*). Recent work demonstrates that the allometric scaling of brain to body size appears to have been repatterned multiple times during the Cenozoic, especially near the Cretaceous-Paleogene and Paleogene-Neogene boundaries (~23 million years ago), when changing environmental conditions may have precipitated evolutionary innovations and ecological radiations (*4*). However, it is difficult to determine how and when mammals encephalized to near-modern levels because well-preserved skulls dating to the Mesozoic and early Cenozoic are fairly rare. Moreover, investigating brain development is complicated because relative or absolute brain size may

Department of Biology, University of New Mexico, Albuquerque, NM 87131-0001, USA.
Email: fasmith@unm.edu

ILLUSTRATION: MARK P. WITTON/SCIENCE SOURCE

not be the best measure of cognition (*4*). Ideally, studies should examine brain complexity and the relative proportions of different brain regions—the volume of brain devoted to higher cognition versus mediating basic autonomic and sensory functions (*4*).

Bertrand *et al.* combined previously reported endocranial data with high-resolution computed tomography (CT) images of newly discovered early Cenozoic mammal skulls. Not only did they measure brain size, they also characterized the individual sensory components for many fossils. These included the neocortex, which is related to higher-order brain functions such as sensory perception and cognition; the olfactory bulbs, which are related to smell; and the cerebellar petrosal lobules, which are related to eye movement. The authors examined 124 extinct species dating from early in mammal evolution to the middle Cenozoic. Their study was particularly data-rich for the crucial period immediately after the Cretaceous-Paleogene extinction, which marked the onset of rapid morphological and ecological diversification in terrestrial mammals and an expansion in mass of more than four orders of magnitude (*10*). They found that immediately after that extinction, mammals were actually less encephalized than earlier, or later, in their evolutionary history. Mammals expanded in body size much faster than brain size as they diversified after the end-Cretaceous extinction—favoring "brawn" before "brain" (*8*). The reduction in PEQ was relatively short-lived, however. By the early Eocene (~50 million years ago), relative brain size had substantially expanded in all clades, although more so in crown orders (earliest members of extant placental orders) than in stem taxa (archaic placentals). The rate of encephalization slowed by 40 million years ago, about when morphological diversification plateaued (*10*).

Notably, changes in relative brain volume over the Cenozoic were not uniform across the different sensory regions; larger brains were not simply scaled-up versions of smaller brains (*4*). Rather, there was a large increase in the neocortex, whereas the proportion of the brain devoted to olfaction decreased. And after an initial decrease in the relative size of the petrosal lobules after the Cretaceous-Paleogene, they too increased over time. These changes in brain composition were more marked in crown orders than in stem taxa, supporting the idea that differences contributed to the eventual decline and extinction of the latter. Overall, encephalization after the initial drop in the Paleocene was driven by the expansion of brain regions mediating balance, vision and eye movement, head control, and sensory integration, rather than olfaction.

Brains are energetically expensive, using almost an order of magnitude more energy per gram than that of other body tissues (*12*). Evolving a larger brain relative to body size can necessitate higher basal metabolic rates and/or reduced allocation to other essential physiological processes (*12*–*14*), although trade-offs remain unclear (*2*, *4*). Given this cost, why have mammals undergone such remarkable encephalization over the Cenozoic? Although numerous hypotheses have been proposed—including selection related to sociality, improved hearing, feeding, taste, olfaction, miniaturization, parental care, endothermy, and nocturnality [for example, (*6*, *8*)], the ecological niche is likely important (*2*, *4*, *15*). Bertrand *et al.* posit that as mammals evolved larger sizes and ecosystems began saturating, interspecific competition intensified selection for larger relative brains. For example, predators developed substantially larger PEQ than that of prey groups by the Eocene. Earlier work reported a transition between the growth and saturating phase of body size evolution at this time, which is consistent with the saturation of ecological niches (*10*).

Understanding the trade-offs between the high energetic demands of brain tissue and the potential value of increased brain size for survival and reproductive success remains a challenge. Studies such as Bertrand *et al.* highlight the utility of the fossil record for disentangling fundamental allometric scaling among contemporary mammals, as well as highlighting the central role of ecology in mammal macroevolution. ■

REFERENCES AND NOTES

1. M. Tsuboi *et al.*, *Nat. Ecol. Evol.* **2**, 1492 (2018).
2. J. R. Burger, M. A. George Jr., C. Leadbetter, F. Shaikh, *J. Mammal.* **100**, 276 (2019).
3. S. Benson-Amram, B. Dantzer, G. Stricker, E. M. Swanson, K. E. Holekamp, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2532 (2016).
4. J. B. Smaers *et al.*, *Sci. Adv.* **7**, eabe2101 (2021).
5. O. C. Bertrand *et al.*, *Science* **376**, 80 (2022).
6. T. B. Rowe, T. E. Macrini, Z.-X. Luo, *Science* **332**, 955 (2011).
7. O. C. Marsh, *Ann. Mag. Nat. Hist.* **14**, 167 (1874).
8. H. J. Jerison, *Evolution of the Brain and Intelligence* (Academic Press, 1973).
9. G. R. Hurlburt, *J. Vertebr. Paleontol.* **15**, 37A (1995).
10. F. A. Smith *et al.*, *Science* **330**, 1216 (2010).
11. J. Alroy, *Syst. Biol.* **48**, 107 (1999).
12. J. W. Mink, R. J. Blumenschine, D. B. Adams, *Am. J. Physiol.* **241**, R203 (1981).
13. R. D. Martin, *Nature* **293**, 57 (1981).
14. K. Isler, C. P. van Schaik, *Biol. Lett.* **2**, 557 (2006).
15. M. D. Pagel, P. H. Harvey, *Am. Nat.* **132**, 344 (1988).

ACKNOWLEDGMENTS

E. A. Elliott Smith and R. E. Elliott Smith provided helpful comments on the manuscript.

10.1126/science.abo1985

PHOTOVOLTAICS

Perovskite solar cells can take the heat

Layered surface structure adds durability to packaged perovskite cells

By **Joseph M. Luther** *and* **Laura T. Schelhas**

Solar panels must endure extreme conditions during their service lifetime. There are few other products, let alone pristine semiconductor electronics, that are manufactured today and expected to withstand decades of abuse from full exposure to outdoor elements, including harsh ultraviolet sunlight, rain and hail, and daily and seasonal thermal cycles. Over the past decade, a new class of semiconductors, known as metal halide perovskites, has emerged, which constitutes the light-absorbing material in solar panels and has made incredible advancements, potentially taking photovoltaics in new directions. Perovskite materials have been touted as a potential game changer for solar energy production (*1*). On page 73 of this issue, Azmi *et al.* (*2*) present a reproducible layered surface structure of the perovskite for improved durability. The resulting perovskite cells can maintain performance under high heat and humidity.

Perovskite solar panels promise an efficient, low-cost, and simple-to-manufacture solution that is on the cusp of commercialization, as either a stand-alone technology or an add-on to silicon in a tandem configuration. However, naysayers of perovskite's future potential often point to the lack of studies demonstrating durability in packaged cells and modules. In a commercial product, arrays of solar cells are packaged to create a solar module. Packaging is used to contain the cells and can protect them from the environment. Commercial solar panels are often subjected to accelerated stress testing to quickly assess their durability in the field. When developing new technologies, researchers cannot simply place panels outside and wait multiple decades to assess stability; rather, they are forced to develop accelerated test procedures and protocols (*3*, *4*).

National Renewable Energy Laboratory, Golden, CO, USA. Email: joey.luther@nrel.gov; laura.shelhas@nrel.gov

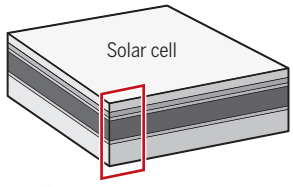
One common testing protocol is the International Electrotechnical Commission's 61215 series of standardized tests (*5*). This test series is a pass-fail protocol that virtually every commercial solar module must pass. One of the tests in this protocol is the so-called damp heat test, which places the module in a chamber set to 85°C (185°F) with 85% relative humidity for 1000 hours. It is designed to expose any weakness of the product against the combined stresses of heat and moisture. Perovskite solar materials have relatively low formation energy—in other words, it takes little energy to form or break atomic bonds within the material. Although this is beneficial for the fabrication of perovskites, it also makes them prone to degradation in high heat. Additionally, perovskites can be sensitive to moisture, further making this test one of the most difficult for perovskite technology to pass. A conventional way to survive the damp heat test is to physically limit the ingress of water into the working components of the panel, thus decoupling and eliminating the humidity portion of the stress. It has been demonstrated that a glass sandwich, with the solar cell material in the middle and an edge seal containing desiccant, can prevent water from reaching the absorber and metal contacts (*6*, *7*). Azmi *et al.* used this approach on highly efficient solar cells but found that they still degraded after damp heat exposure. Although their glass sandwich barrier can keep the moisture out, the device is still subjected to extreme thermal stress, suggesting that the perovskite was degrading from high heat and not moisture. To put this in context, the hottest air temperature ever recorded on earth is 58°C, and the solar panel itself may regularly reach temperatures of 65°C (*8*). The 85°C temperature during the damp heat test is beyond the maximum temperature experienced, but the benchmark has proved to be a good substitute for assessing temperature-induced degradation processes.

Azmi *et al.* improved the thermal stability of perovskite cells by treating the films with an oleylammonium iodide $[(C_{18}H_{35})NH_3]I$ solution and modifying the thermal processing during fabrication to alter the perovskite crystals' surface structure (see the figure). The authors show that this treatment creates a layered surface structure, causing passivation on each microscopic perovskite grain. Surfaces are often the most problematic portion of semiconductors, and passivation refers to terminal binding species that remove what could become electronic defects. This transformation has been shown to improve various aspects of performance and stability (*9*, *10*). Layered perovskite-inspired materials are at the forefront of materials science research and are being considered for various applications, including solar cells, light-emitting diodes (LEDs), and other devices (*11*, *12*). Microscopically speaking, layered perovskites are composed of alternating lead halide octahedra and self-assembled molecules that are typically too large to allow the formation of a three-dimensional perovskite lattice. These layers can form various crystal structures depending on how the octahedra of one layer align with those of the adjacent layer. For example, the added oleylammonium cations from the treatment can be controlled to orient between each lead halide layer or less frequently. Amzi *et al.* observed that films processed without annealing during fabrication created two octahedral units per layer adjacent to the perovskite rather than one, which is shown to be a superior barrier. Another major win for this approach is the reproducibility of the process. Consistent results were repeatedly obtained by many device makers within the research team.

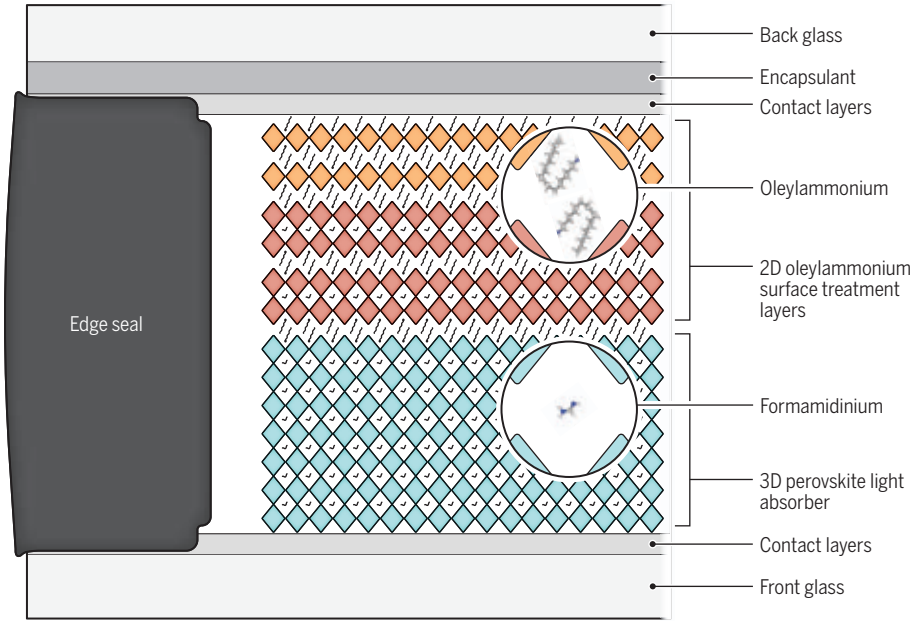
The 24% efficient perovskite solar cells that are stable under damp heat tests demonstrate a step in the right direction for perovskite solar panels. Thoughtful selection of the package can prevent some degradation pathways, and informed materials engineering can open routes to improving the inherent thermal stability of perovskite solar materials. Despite all the excitement with perovskites, accelerated testing protocols such as this are no replacement for long-duration field studies that existing technologies have completed over the years. There is a strong push to begin collecting outdoor performance and durability data now. As the community pushes this technology forward, more will be known regarding whether perovskites can contribute as much to solar energy as anticipated. Continued dedication to these approaches can help enable commercialization. ■

Perovskite solar cell packaging

The oleylammonium iodide surface treatment creates a thin layer of two dimensional (2D) perovskite, which aids in heat tolerance. The treated perovskite material can then be used to build the photovoltaic device, where glass and an edge seal protect the stack from humidity.



Packaged perovskite module structure



GRAPHIC: A. MASTIN/*SCIENCE*

REFERENCES AND NOTES

1. Y. Rong *et al.*, *Science* **361**, eaat8235 (2018).
2. R. Azmi *et al.*, *Science* **376**, 73 (2022).
3. H. J. Snaith, P. Hacke, *Nat. Energy* **3**, 459 (2018).
4. M. V. Khenkin *et al.*, *Nat. Energy* **5**, 35 (2020).
5. IEC 61215-1-1, revision 2.0, *Terrestrial photovoltaic (PV) modules – Design qualification and type approval – Part 1-1: Special requirements for testing of crystalline silicon photovoltaic (PV) modules* [International Electrotechnical Commission (IEC), 2021].
6. R. Cheacharoen *et al.*, *Sustain. Energy Fuels* **2**, 2398 (2018).
7. M. D. Kempe, A. A. Dameron, M. O. Reese, *Prog. Photovolt. Res. Appl.* **22**, 1159 (2014).
8. S. Kurtz *et al.*, *Prog. Photovolt. Res. Appl.* **19**, 954 (2011).
9. G. Grancini *et al.*, *Nat. Commun.* **8**, 15684 (2017).
10. J. Tong *et al.*, *Science* **364**, 475 (2019).
11. Y.-H. Kim *et al.*, *Science* **371**, 1129 (2021).
12. J.-C. Blancon *et al.*, *Science* **355**, 1288 (2017).

ACKNOWLEDGMENTS

This work was authored by the National Renewable Energy Laboratory, operated by the Alliance for Sustainable Energy, LLC, for the US Department of Energy (DOE) under contract no. DE-AC36-08GO28308. The authors acknowledge the US DOE Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office and the Operational Energy Capability Improvement Fund of the Department of Defense for support. The views expressed in the article do not necessarily represent the views of the DOE or the US Government.

10.1126/science.abo3368

ECOLOGY

Eco-evolutionary effects of keystone genes

The rapid evolution of specific genes within species can drive ecological changes

By **Patrik Nosil**[1] *and* **Zach Gompert**[2]

There is increasing evidence that genetic evolution can occur rapidly enough to affect the ecological dynamics of populations and communities (*1–3*). To better predict the future of ecosystems, it is necessary to understand how evolutionary changes within species influence and interact with ecological changes through processes known as "eco-evolutionary dynamics" (*4*). On page 70 of this issue, Barbour *et al.* (*5*) demonstrate that a gene affecting a plant's resistance to herbivory also influences the persistence of the food web through the gene's effect on plant growth (see the figure). Subsequent studies of natural selection in the wild can help explain how variations of such "keystone genes" can be maintained (*6, 7*). The maintenance of genetic variation in keystone genes is required for eco-evolutionary dynamics to be perpetual rather than transient.

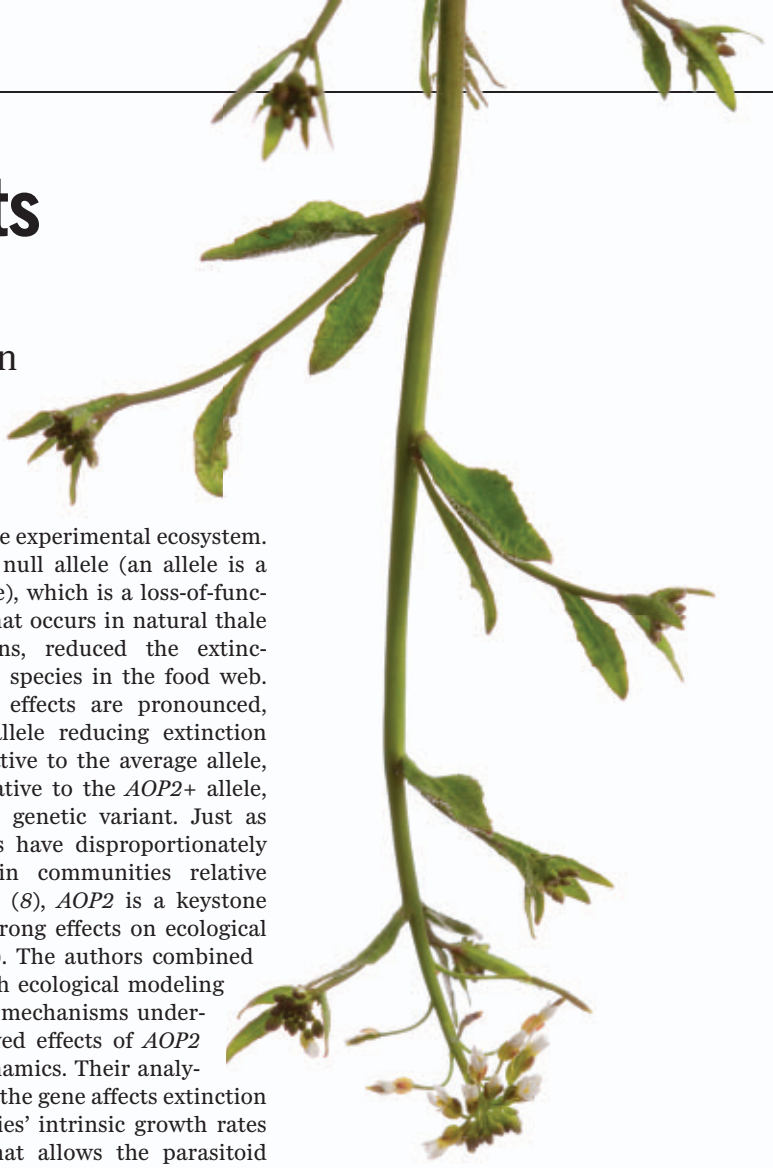
It is important to determine the genes that underlie eco-evolutionary dynamics because genetic details, such as the number and biological effect of genes that affect traits, are expected to influence evolution. In this context, by identifying a gene with marked ecological effects, the study by Barbour et al. offers several key insights. The authors constructed an experimental food web in the laboratory, which comprised a predator (a parasitoid wasp that attacks aphids), two herbivores (two species of aphids that eat plants), and the plant *Arabidopsis thaliana*—also known as thale cress, a model genetic organism. They studied three genes in the cress that showed variation in the wild. The genes were chosen because of their role in controlling chemical biosynthesis that could affect resistance to herbivory.

Barbour *et al.* observed that one of the genes, *AOP2*, affects the persistence of the food web. That is, this gene affects extinction risk and the collapse of the wider community in the experimental ecosystem. Specifically, the null allele (an allele is a variant of a gene), which is a loss-of-function mutation that occurs in natural thale cress populations, reduced the extinction risk for the species in the food web. Critically, these effects are pronounced, with the null allele reducing extinction risk by 16% relative to the average allele, and by 29% relative to the *AOP2*+ allele, a specific *AOP2* genetic variant. Just as keystone species have disproportionately strong effects in communities relative to other species (*8*), *AOP2* is a keystone gene that has strong effects on ecological dynamics (*9, 10*). The authors combined their results with ecological modeling to elucidate the mechanisms underlying the observed effects of *AOP2* on food web dynamics. Their analysis revealed that the gene affects extinction by altering species' intrinsic growth rates in a manner that allows the parasitoid and the dominant aphid species to coexist. Thus, the results also inform the mechanisms by which species coexist—a major theme in ecology (*11*).

Given the revelation brought by these results, perhaps the most pressing question that needs answering is whether the variation at keystone genes is maintained over the long term, and if so, what are the mechanisms that preserve this variation over time (*6, 7*)? These are important questions because genetic variation is the fuel for evolution. If the variation is lost, eco-evolutionary dynamics will not occur except when following variation introduced by the occasional mutation or gene flow. In classic population-genetic models of adaptation, genetic variation is expected to be lost as natural selection acts in a single, consistent direction that favors beneficial alleles in the population (*6*). Through this process, natural selection causes one allele to replace all others, eliminating genetic variation from the population.

In contrast to this view of directional selection and the loss of variation, the ecological genetics literature emphasizes "balancing selection," where heterozygotes (i.e., individuals with two different allele copies at a gene) have higher fitness than homozygotes or where selection fluctuates over time and is dependent on the environmental or genetic context (*7*). In such instances, the selection is not uniform, and genetic variation can be maintained. Determining how selection acts on keystone genes in the wild is thus required to predict whether eco-evolutionary dynamics will be brief and transient (as predicted by directional selection), or more pervasive and perpetual (as predicted by balancing selection) (*1–3*).

Another future direction concerns the genetic architecture of traits that exhibit ecological effects (*4, 12, 13*). Although traits controlled by single genes of large effect certainly exist, such as the *Agouti* gene affecting coat color in mice (*12*),



Barbour *et al.* demonstrate in a laboratory setting that a gene affecting the herbivory resistance of a plant (pictured here as the thale cress) also influences extinction risks of other organisms in the community.

PHOTO: NIALL BENVIE/MINDEN PICTURES

[1]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Université Paul Valéry Montpellier 3, Montpellier, France. [2]Department of Biology, Utah State University, Logan, UT, USA. Email: patrik.nosil@cefe.crns.fr

many traits exhibit continuous phenotypic variation underlain by many genes (*14*). In such cases, individual genes have minor effects on trait variation. Moreover, not only genes but also the environment influences trait expression.

Additionally, even apparent cases of single genes with strong effects on a trait may represent multiple genes that are tightly linked (i.e., physically located) on the same chromosome, as occurs in "supergenes" (*13*). Studying such complex genetic architectures is more challenging than studying single genes, as evidenced by difficulties in genetic mapping of human disease and complex behavioral traits (*14*). Testing how traits underlain by many genes affect ecological dynamics is a challenging, yet important, avenue for future work. A caveat is that even if major effect loci are relatively rare, they could be more likely than minor effect loci to exert marked ecological effects (*1*).

Further studies that combine disciplines such as ecology, genetics, and mathematical modeling are likely to invigorate the field of eco-evolutionary dynamics. Although simple systems are a powerful and useful starting point for such work, most eco-evolutionary systems are more complex because of the interactions and feedback among and within ecological and evolutionary processes, and complex communities and trait genetics (*1*–*3*). This complexity of eco-evolutionary systems must be unraveled to elucidate if these dynamics will be gradual or abrupt, and how the dynamics can be characterized by tipping points in ecosystems (*15*). Such knowledge will inform the importance of evolution for ecological dynamics and biodiversity. ■

REFERENCES AND NOTES

1. A. P. Hendry, *Eco-evolutionary Dynamics* (Princeton Univ, Press, 2017).
2. T. E. Farkas *et al.*, *Curr. Biol.* **23**, 1835 (2013).
3. J. N. Thompson, *Relentless Evolution* (Univ. of Chicago Press, 2013).
4. L. Becks *et al.*, *Ecol. Lett.* **15**, 492 (2012).
5. M. A. Barbour *et al.*, *Science* **376**, 70 (2022).
6. H. A. Orr, *Evolution* **52**, 935 (1998).
7. E. B. Ford, *Ecological Genetics* (Chapman and Hall, ed. 3, 1971).
8. R. T. Paine, *Am. Nat.* **103**, 91 (1969).
9. L. H. Skovmand *et al.*, *Trends Ecol. Evol.* **33**, 689 (2018).
10. T. G. Whitham *et al.*, *Ecology* **84**, 559 (2003).
11. J. M. Levine *et al.*, *Nature* **546**, 56 (2017).
12. C. R. Linnen *et al.*, *Science* **339**, 1312 (2013).
13. M. Wellenreuther, L. Bernatchez, *Trends Ecol. Evol.* **33**, 427 (2018).
14. S. E. Hyman, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, 20170031 (2018).
15. M. Scheffer *et al.*, *Science* **338**, 344 (2012).

ACKNOWLEDGMENTS

This work is part of a project that has received funding from the European Research Council (ERC) to P.N., under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 770826 EE-Dynamics). Z.G. is supported by the US National Science Foundation (DEB-1638768).

10.1126/science.abo3575

EVOLUTION AND COGNITION

Selective cultural processes generate adaptive heuristics

Less intuitive, hard-to-learn cognitive heuristics can thrive

By **Joseph Henrich**

Long before the arrival of astrolabes, compasses, or marine chronometers, Micronesian navigators guided canoes across hundreds of miles of open ocean by integrating information from stars, waves, coral reefs and more into a rich conceptual framework. Looking into the heavens, they saw a celestial compass and knew which rising and setting stars to follow. They used three distinct types of waves—originating from the south, east, and northeast—as directional references (*1*, *2*). Equipped with such cognitive tools, the ancestors of these Micronesians spread across Oceania, sailing from the bustling islands of Southeast Asia to the remote reaches of Hawaii, Easter Island, and South America. How do such complex cognitive technologies emerge? On page 95 of this issue, Thompson *et al.* (*3*) describe taking an experimental approach to the question of how opportunities to selectively learn from successful role models can favor the spread of more adaptive, but less intuitive, cognitive heuristics over more intuitive and memorable alternatives.

To hunt, gather, farm, fish, and tackle countless other challenges, humans have relied on a dizzying array of complex, locally adaptive heuristics (*4*, *5*). Yet the origins of such heuristics present a puzzle because the best protocols and practices are often not the easiest to learn, remember, and teach. Even when one manages to master a more effective, but less intuitive, technique, it will deteriorate as it's transmitted and transformed by the minds of others over generations (*6*). Researchers have long considered how ensembles of psychological biases, preferences, and inferences make certain ideas, stories, songs, and concepts catchier—easier to acquire, recall, and retransmit—and have deployed these "cognitive attractors" to account for the recurrent patterns found across societies in domains such as religion, literature, art, and folk biology (*7*–*10*). But given the pervasiveness of such cognitive attractors, how can the impressive assemblages of nonintuitive heuristics and hard-to-learn cognitive abilities that have permitted humans to dominate Earth's major ecosystems be accounted for?

Approaching this puzzle, cultural evolutionary theorists were inspired by ethnographic accounts such as those for Micronesia, where all young males aspired to become master navigators—the pinnacle of local prestige—by apprenticing under the most respected masters (*1*). Few aspirants, however, succeeded in being initiated as masters because some of the most important skills were nonintuitive, difficult to learn, and unforgiving of mediocrity—miscalculating by a few degrees could mean sailing past an atoll and dehydrating or starving, lost in the Pacific's vastness. Using mathematical models, theorists have shown that if learners selectively attend to models or teachers based on cues of prestige, success, and skill, this can drive the spread of less intuitive heuristics by compensating for errors that creep in when complex heuristics are transmitted from person to person (*6*).

To assess the role of selective cultural learning in creating adaptive heuristics, Thompson *et al.* paid experimental participants according to how efficiently they could complete a sorting task. Participants had to correctly order six tiles using the fewest number of paired comparisons and were told that the tiles themselves were uninformative—only the paired comparisons revealed ordering information. Participants had to sort nine different sextets of tiles. After completing their assignment, they each wrote up what they had learned for the next generation of sorters and passed it along with a demonstration of their sorting strategy. Sorters were further paid according to the success of those who tapped them as "teachers." After the first generation, which served as an asocial treatment in which sorters had no opportunity to learn from others, participants in generations 2 to 12 were randomly placed into either the selective social learning or random mixing treatments. In both treatments, naïve sorters could select up to three teachers from a set of eight partici-

Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA.
Email: henrich@fas.harvard.edu



An illustration from the early 1800s depicts canoes near the Mariana Islands in Micronesia.

pants from the prior generation. The only difference was that sorters in the selective social learning treatment could observe the monetary payoffs ("success") of their eight potential teachers, whereas those in the random mixing treatment observed only their teachers' arbitrary identification numbers. When learners selected a particular model, they had an opportunity to read their teacher's write-up, observe their demonstration, and practice their strategy.

When individuals could preferentially learn from more successful sorters, their average performances increased across time such that, by the final generations, their scores exceeded those in the random mixing and asocial treatments by 24 and 70%, respectively. By the end, the randomly chosen participants who could selectively learn from others seemed smarter: They had better cognitive algorithms for solving this kind of problem. This not only confirms prior work (*11*, *12*), illustrating how selective cultural learning can foster the evolution of more efficient artifacts (e.g., kayaks and bows), but also shows that it can generate better information-processing algorithms.

Closer inspections of the data reveal even more notable results. Although there are a vast number of possible ways to go about this sorting process, only eight distinct and identifiable algorithms emerged in the social treatments, and only two—selection sort and gnome sort—dominated the final three generations. Meanwhile, in the asocial treatment, most sorters used idiosyncratic approaches that didn't approximate any known algorithm and generally performed poorly. Thus, when social learning was possible, a method's transmissibility, fidelity, and memorability rapidly shaped the variation in sorting algorithms, creating quasi-discrete cultural units for selective learning to act on (*6*).

Of the two leading algorithms, selection sort was substantially easier to learn and/or transmit—and thus represents a stronger cognitive attractor—but was 30% less efficient than gnome sort at the sorting task. The difficulty of transmitting and learning the gnome algorithm is seen by the more complex written instructions provided by gnome teachers and the more frequent errors made by those trying to imitate gnome during their practice round. Consequently, in the random mixing treatment, when gnome did sometimes emerge and begin to spread, errors introduced during the transmission process degraded its effectiveness, allowing selection sort to out compete it. By contrast, when selective learning was possible, gnome not only survived but often diffused across the population and was sustained at sizable frequencies. Here, selective processes favored the more efficient, but harder to learn, cognitive attractor by allowing learners to pick models who hadn't inadvertently introduced errors, thus filtering out errors each generation. Indeed, only in the selective social learning treatment did a stable cadre of gnome "master sorters" emerge. As with Micronesian navigators, nearly all master sorters (>85%) had master sorters as teachers, though most students of master sorters did not become master sorters themselves. This experimental evidence elegantly converges with well-established ethnographic patterns (*13*, *14*).

These results highlight a deeper point: Humans don't have culture because we're smart, we're smart because we have culture (*3*). The selective processes of cultural evolution not only generate more sophisticated practices and technologies but also produce new cognitive tools—algorithms—that make humans better adapted to the ecological and institutional challenges that we confront. Thompson *et al.*'s results underline the need for the psychological sciences to abandon their implicit reliance on a digital computer metaphor of the mind (hardware versus software) and transform into a historical science that considers not just how cultural evolution shapes what we think (our mental contents) but also how we think [our cognitive processes (*15*)]. ■

REFERENCES AND NOTES

1. T. Gladwin, *East is a Big Bird: Navigation and Logic on Puluwat Atoll* (Harvard Univ. Press, 2009).
2. W. H. Goodenough, *Native Astronomy in the Central Carolinas* (University of Pennsylvania, University Museum, Anthropological Publications, 1953).
3. B. Thompson, B. van Opheusden, T. Sumers, T. L. Griffiths, *Science* **376**, 95 (2022).
4. J. Henrich, *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter* (Princeton Univ. Press, 2016).
5. R. Boyd, *A Different Kind of Animal: How Culture Transformed Our Species* (Princeton Univ. Press, 2017).
6. R. Boyd, J. Henrich, *J. Cogn. Cult.* **2**, 87 (2002).
7. O. Morin, *Evol. Hum. Behav.* **34**, 222 (2013).
8. S. Atran, D. L. Medin, *The Native Mind and the Cultural Construction of Nature* (MIT Press, 2008).
9. P. Boyer, *Religion Explained: The Evolutionary Origins of Religious Thought* (Basic Books, 2001).
10. M. Singh, *Emot. Rev.* **13**, 183 (2021).
11. M. Derex *et al.*, *Nature* **503**, 389 (2013).
12. M. Muthukrishna *et al.*, *Proc. R. Soc. London Ser.* B **281**, 20132511 (2014).
13. J. Henrich, N. Henrich, *Proc. R. Soc. London Ser. B* **277**, 3715 (2010).
14. J. Henrich, J. Broesch, *Philos. Trans. R. Soc. London Ser. B* **366**, 1139 (2011).
15. M. Muthukrishna, J. Henrich, E. Slingerland, *Annu. Rev. Psychol.* **72**, 717 (2021).

10.1126/science.abo0713

IMAGE: SCIENCE & SOCIETY PICTURE LIBRARY/GETTY IMAGES

NEUROSCIENCE

Pain-resolving microglia

An emergent subgroup of spinal cord microglia mediates recovery from persistent pain

By **George Sideris-Lampretsas** *and* **Marzia Malcangio**

Neuropathic pain after peripheral nerve damage is a lasting condition that generally persists even when the cause of damage disappears (*1*). The immune system is integral to the development of neuropathic pain: In the spinal cord, microglia—the central nervous system–resident macrophages—respond to neuronal activity and set up a positive feedback loop with neurons that promotes pain onset. Thus, disruption of microglia-neuron communication is being considered as a strategy to produce analgesia. On page 86 of this issue, Kohno *et al.* (*2*) describe a spinal cord–resident pool of microglia that emerges during pain maintenance and contributes to the resolution of neuropathic pain in mice. Microglial heterogeneity is a well-accepted concept that is underexplored in the context of chronic pain (*3*). The finding that spinal cord microglia acquire spatial and temporal transcriptional heterogeneity that affects pain could identify new therapeutic strategies to relieve pain.

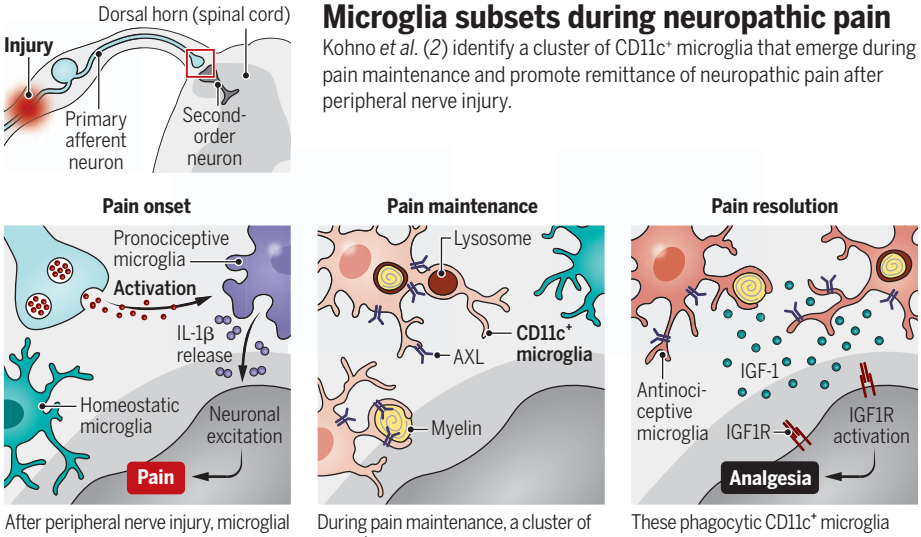
In response to peripheral nerve trauma, the nociceptive system is activated. Nociception is the response of the nervous system, including the spinal cord, to noxious stimulation and tissue damage. Painful stimuli are detected by peripheral nerve fibers and transmitted to their central terminal in the dorsal horn of the spinal cord. There, microglia promptly proliferate and up-regulate the expression of various genes that are implicated in pronociceptive pathways (*4*). Microglia play a pronociceptive role at the onset of pain, in a sex-dependent manner: The involvement of microglial signaling is reliable in male mice, whereas some pathways appear not to be effective in female mice (*5*). However, these cells are considered less relevant during the maintenance of neuropathic pain (*4*).

By contrast, the study of Kohno *et al.* shows that in the maintenance phase, an emergent CD11c$^+$ microglial pool engages in clearance of myelin. The prevention of CD11c$^+$ microglia emergence is sufficient for pain to be maintained in a mouse model of neuropathic pain, indicating that they are essential for pain remission. These microglia develop an antinociceptive function through the release of anti-inflammatory mediators, such as insulin-like growth factor 1 (IGF-1). Thus, under neuropathic conditions, microglia can serve diverse roles in a temporal manner alongside pain development and remittance. Some microglia initially respond to neuronal activation and contribute to pain onset; then, with time, this distinct pool of microglia appears, expands, and eventually establishes a new microglia-to-neuron communication that is associated with pain resolution (see the figure).

Wolfson Centre for Age-Related Diseases, King's College London, London, UK. Email: george.sideris@kcl.ac.uk; marzia.malcangio@kcl.ac.uk

Microglia subsets during neuropathic pain

Kohno *et al.* (*2*) identify a cluster of CD11c$^+$ microglia that emerge during pain maintenance and promote remittance of neuropathic pain after peripheral nerve injury.



After peripheral nerve injury, microglial activation in the dorsal horn engages pronociceptive communication with neurons by releasing cytokines [such as interleukin-1β (IL-1β)] that facilitate neuronal activity.

During pain maintenance, a cluster of CD11c$^+$ microglia emerge, express the receptor AXL, and engulf myelin debris from degenerating processes that affect injured neurons.

These phagocytic CD11c$^+$ microglia release insulin-like growth factor 1 (IGF-1), which exerts antinociceptive action through activation of neuronal IGF-1 receptor (IGF1R).

What is the local cue that triggers antinociceptive microglia emergence during neuropathic pain? In brain development and disease, CD11c$^+$ microglia appear in response to disturbance of homeostasis, and they check neuronal health by increasing phagocytic potentials (*6*). Kohno *et al.* found that at 2 weeks after nerve injury in mice, CD11c$^+$ microglia express the receptor tyrosine kinase AXL, which alongside the receptor tyrosine kinase MER detects and engulfs amyloid-β aggregates in the brains of people with Alzheimer's disease (*7*). Kohno *et al.* observe the CD11c$^+$ microglial population engulfing myelin basic protein, which suggests that they are devoted to taking up debris resulting from loss of myelin integrity in some fiber terminals. At this time, CD11c$^+$ microglia are not yet antinociceptive. Loss of myelin integrity may be the result of degenerative processes affecting injured neurons that terminate in the spinal cord. Alternatively, loss of myelin integrity may be mediated by increased neuronal excitation in response to injury; hence, microglial modulation of myelin integrity may indirectly control neural circuit functionality (*8*).

By 5 weeks after nerve injury, when pain has fully resolved in the mouse model, CD11c$^+$ microglia express IGF-1, which confers a sex-independent antinociceptive effect to microglia in these neuropathic conditions. Thus, throughout the pain-maintenance phase, CD11c$^+$ phagocytic microglia gradually become antinociceptive and eventually contribute to the resolution of nociceptive hypersensitivity through release of IGF-1. Because the receptor for IGF-1 is expressed by neurons in the dorsal horn but also by oligodendrocytes, astrocytes, endothelial cells, and microglia, there remains scope for further studies on cellular targets of CD11c$^+$ microglia-derived IGF-1.

Microglia can change function in response to regional cues through fast and effective regulation of gene expression. For example, transcriptomics changes in neurodegenera-

GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

tive conditions show that subsets of microglia shift from a dynamic homeostatic profile and acquire a specific disease-associated microglia (DAM) transcriptional signature (*9*). In pain remission, CD11c$^+$ microglia show some similarities to DAMs because they differentially express genes such as apolipoprotein E (*Apoe*) and *Axl*. Nevertheless, CD11c$^+$ microglia in the neuropathic spinal cord of mice have a distinct transcriptional profile from the DAM subset. This transcriptional profile is associated with antinociceptive microglial functional output. A cluster of microglia reminiscent of the antinociceptive CD11c$^+$ microglia that express *Apoe*, *Axl*, and *Igf1* are also detected in the spinal cord of mice 5 months after peripheral nerve injury (*10*). These results support the suggestion that a subset of microglia expressing a well-preserved network of transcripts appears weeks after pain onset, with Kohno *et al.* finding that these microglia have antinociceptive properties.

One of the most valuable implications of this work is how this new mechanism that stimulates the emergence of pain-killing microglia can be translated into effective treatments. Better understanding of the role of IGF1 and APOE in antinociceptive microglia would facilitate the design of therapeutic paradigms that promote analgesia. Furthermore, definition of the context required for the emergence of antinociceptive microglia would be invaluable to determine how to induce protective microglia in chronic pain conditions. Therefore, studies on pain remission states that are not characterized by loss of myelin integrity, and hence are not associated with enhanced microglial phagocytic activity, would illustrate whether microglia can then acquire an antinociceptive phenotype. It remains to be shown whether the constant uptake of myelin might eventually cause phagocytic exhaustion and microglia dystrophy, which will affect neuronal health. Overall, the study of Kohno *et al.* highlights that it is time to reassess the role of microglia in pain as a therapeutic target for the effective treatment of chronic pain. ■

REFERENCES AND NOTES

1. M. Costigan *et al.*, *Annu. Rev. Neurosci.* **32**, 1 (2009).
2. K. Kohno *et al.*, *Science* **376**, 86 (2022).
3. G. Sideris-Lampretsas, M. Malcangio, *Brain Behav. Immun.* **96**, 279 (2021).
4. G. Chen *et al.*, *Neuron* **100**, 1292 (2018).
5. R. E. Sorge *et al.*, *Nat. Neurosci.* **18**, 1081 (2015).
6. A. Benmamar-Badel *et al.*, *Front. Immunol.* **11**, 430 (2020).
7. Y. Huang *et al.*, *Nat. Immunol.* **22**, 586 (2021).
8. E. N. Santos, R. D. Fields, *Sci. Adv.* **7**, eabk1131 (2021).
9. H. Keren-Shaul *et al.*, *Cell* **169**, 1276 (2017).
10. S. Tansley *et al.*, *Nat. Commun.* **13**, 843 (2022).

ACKNOWLEDGMENTS

We acknowledge funds from the European Union Horizon 2020 research and innovation programme "TOBeATPAIN" under Marie Skłodowska-Curie grant agreement 764860.

10.1126/science.abo5592

GENOMICS

A next-generation human genome sequence

A near-complete sequence outlines a path for a more inclusive reference

By **Deanna M. Church**

Twenty-one years ago, two initial versions of a human genome sequence were published by Celera Genomics (*1*) and the Human Genome Project (HGP) (*2*). These assemblies were incomplete and replete with errors, but despite these flaws, the value of having a human genomic reference assembly was clear. The assembly produced by the HGP went on to be "finished" (*3*) and has been continually updated over the past decade (*4*). Despite the scientific and economic value of the reference (*5*), it has many shortcomings, including that it is not actually finished. On page 44 of this issue, Nurk *et al.* (*6*) provide the most complete reference assembly for any mammal. This new human reference is poised to have its own substantial impact on genome analysis and represents an important step to assembly models that represent all humans, which will better support personalized medicine, population genome analysis, and genome editing.

The current version of the human genome reference assembly, GRCh38.p14 (GRCh38), has millions of bases represented by the letter "N," which means that the actual base residing at that location is unknown. There are also 169 sequences that cannot confidently be ordered or oriented within the assembly, typically owing to their repetitive nature, and that get carried along as a bag of sequences that are hard to analyze. Biologically important regions—such as the short arms of acrocentric chromosomes, centromeres, and several duplicated euchromatic regions—are not represented, are represented by model sequences, or are incorrectly represented. These sequences account for ~8% of the human genome. Until recently, limitations of sequencing technology, primarily that the sequencers could read no more than about 1000 bases at a time, it was impossible to correctly assemble the sequence reads of these regions. But they have important biological functions and, in some cases, these regions are associated with human disease.

Inscripta, Inc., Boulder, CO, USA.
Email: deanna.church@inscripta.com

When the idea of sequencing the human genome was initiated, it was not a universally popular idea, nor was there a clear technological blueprint of how the project would be completed (*7*). The initial focus was on mapping-based approaches to better understand the human genome and to further develop sequencing technology. There were public debates on the best way to proceed, with some pushing for an approach that required no mapping, just chopping up the whole genome and sequencing it in a "whole-genome shotgun" sequencing approach (*8*); others argued against this (*9*). The HGP opted for a more structured approach. This involved cloning genomic DNA into pieces that could be grown in bacteria (clones) and indexed in 96-well plates. Clones from these libraries were first mapped to chromosome regions, then individually sequenced. The finished clones were assembled to create the chromosome sequences found in the assembly. Celera Genomics opted for the shotgun approach, which was also used by the Telomere-to-Telomere (T2T) Consortium and is the dominant method today. The comparison of the two assemblies (*10*) led to a better understanding of both approaches and of the human genome sequence.

An important attribute of the human reference assembly is that the source DNA was derived from multiple individuals. One of the first lessons in genetics is that humans inherit one set of chromosomes from each parent. This leads to a duplication of each chromosome, except for the sex chromosomes in individuals harboring an XY configuration. Each pair of chromosomes are not exact duplicates; there can be hundreds of thousands of differences between them, including the presence or absence of large stretches of sequence on one chromosome. The sequence on an individual chromosome is called a haplotype. One reason the bacterial cloning approach was preferred by the HGP was "because each clone represents a single haplotype, problems caused by the presence of polymorphisms are eliminated" [(*9*), p. 411]. That DNA from many donors was pooled in both approaches suggests that the extent of individual variation was not fully appreciated at the time. For example, when two clones

from different haplotypes of an individual are adjacent in the reference assembly, this can create sequence representations that are not normally found in the population. In some cases, the sequence of the two haplotypes can be very different, leading to artificial sequence gaps because correct overlaps between sequence reads cannot be obtained.

Nurk *et al.*, and other studies from the T2T Consortium, eliminated the problem of allelic diversity by sequencing the genome of a cell line derived from a complete hydatidiform mole (CHM). This abnormal product of conception arises when only one copy of a parental genome is retained after fertilization. This is duplicated so that the cell contains two copies of the same parental genome, which is incompatible with fetal viability and leads to placental overgrowth, called a molar pregnancy. Cell lines derived from CHM eliminate allelic variation so that an assembly approach can focus on resolving sequences that have more than two copies in a genome (paralogous duplication). By focusing on one cell line that contains a single haplotype, and thus no allelic differences between the chromosomes that can confuse assembly algorithms, many inaccuracies of GRCh38 could be corrected. The one deficit of this assembly is the lack of a Y chromosome, because the CHM13 cell line does not contain one.

The new assembly provides access to centromeric sequences (*11*), ribosomal sequences (*6*), and an improved representation of euchromatic segmental duplications (SDs) (*12*). SDs are long, often >100-kb sequences with >90% sequence identity that are often associated with chromosome rearrangements, many of which lead to neurodevelopmental disorders. The assembly of these complex regions is only now possible because of the advent of accurate long-read (~20 kb) and accurate-enough ultra-long–read (>100 kb) sequencing. Having reads that can span the length of repetitive sequences allows for their correct placement and representation in the assembly. These sequences are complex and dynamic in the human population, and although there is more complexity in the population than is represented in the T2T-CHM13 assembly, having one correctly assembled human haplotype facilitates analysis of these biomedically important regions in other humans and nonhuman primates (*12*). However, an assembly of a single haplotype, even one as high quality as this, is insufficient for representing the full diversity of the human population. An exciting aspect of the study by Nurk *et al.* is that it lays the framework for how more high-quality assemblies can be achieved, and thus more insight into population- and individual-level diversity can be obtained.

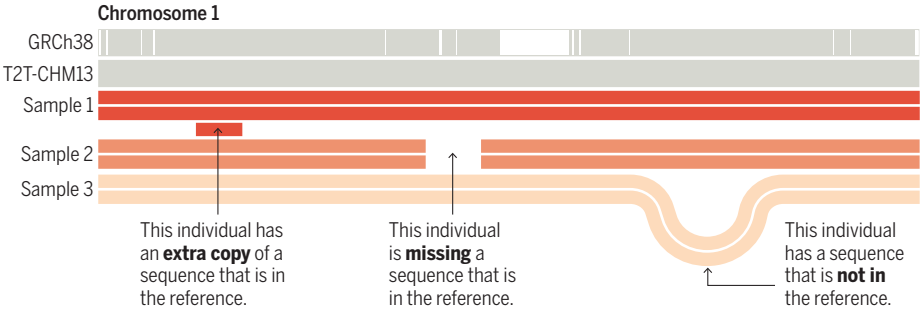
The quality and completeness of the T2T-CHM13 assembly suggests that this should replace GRCh38, with a representation of the Y chromosome added. An immediate solution would be to add the Y chromosome from GRCh38. Longer-term solutions involve improved methods that allow assemblies of diploid genomes, a goal of the T2T Consortium. Considerable challenges remain in achieving this transition. A reference assembly is not just a substrate for sequence alignment and variant calling, it is also the framework on which biological information is annotated. Although the T2T Consortium provided gene-level annotation for the new assembly, it is still not possible to take full advantage of the improvements in T2T-CHM13. Remapping annotations from GRCh38 to T2T-CHM13 is insufficient because of regions in T2T-CHM13 that have no correlate in GRCh38. The reannotation of large datasets—such as DNA variants in gnomAD (*13*), which remains split between annotating GRCh37 and GRCh38, and regulatory information found in ENCODE (*14*)—will be substantial endeavors but critical for the full adoption of T2T-CHM13.

It is an exciting advance to have a human assembly this complete and accurate. But this is only a first step in developing better assembly models that allow the representation of any sequence found in a human genome. Despite the rich array of human diversity, the reference assembly has always been represented as a set of single, linear sequences, one for each chromosome. This single linear representation limits our ability to analyze all genomes because sequences that are present on some chromosomes, but not in the reference, cannot be analyzed. It is necessary to continue developing assembly models that allow representation of the rich sequence diversity found in the human population (see the figure). Understanding this diversity will help dissect disease risks observed in different populations. The work to date by the T2T Consortium is an important part of this vision, because it provides the blueprint by which routine, de novo assembly of individual genomes will be possible. The continued work of the T2T Consortium, along with the Human Pangenome Reference Consortium (*15*), which aims to produce high-quality assemblies from diverse human populations, will provide additional protocols for routine, diploid assemblies as well as the data structures and tools needed to produce a reference assembly that can represent all possible sequences in a population. This type of representation is important not only for population-based analysis but also for individual-genome analysis. Better resolution of both sets of chromosomes in an individual will give a full picture of the genome and undoubtedly help improve personalized medicine. In addition, these data models will allow better representation of the programmed variation that is now possible using genome-editing tools. Assemblies that can also represent the diversity of the human population will better serve science and humanity. ■

A more complete reference

The new human genome assembly, T2T-CHM13 from the Telomere-to-Telomere Consortium, includes complex and repetitive regions of chromosomes that had not been included in previous versions of the human genome assembly (GRCh38). Although the Y chromosome remains to be completed, this new reference could be annotated with regulatory regions, variants, and sequence diversity to give a fuller picture of human genomic variation.





GRAPHIC: V. ALTOUNIAN/*SCIENCE*

REFERENCES AND NOTES

1. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
2. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
3. International Human Genome Sequencing Consortium, *Nature* **431**, 931 (2004).
4. D. M. Church et al., *PLOS Biol.* **9**, e1001091 (2011).
5. S. Tripp, M. Grueber, "Economic impact of the Human Genome Project" (Battelle Memorial Institute, 2011); https://bit.ly/3JpwH4r.
6. S. Nurk *et al.*, *Science* **376**, 44 (2022).
7. M. V. Olson, *Science* **270**, 394 (1995).
8. J. L. Weber, E. W. Myers, *Genome Res.* **7**, 401 (1997).
9. P. Green, *Genome Res.* **7**, 410 (1997).
10. X. She *et al.*, *Nature* **431**, 927 (2004).
11. N. Altemose *et al.*, *Science* **376**, eabl4178 (2022).
12. M. R. Vollger *et al.*, *Science* **376**, eabj6965 (2022).
13. K. J. Karczewski *et al.*, *Nature* **581**, 434 (2020).
14. J. E. Moore *et al.*, *Nature* **583**, 699 (2020).
15. K. H. Miga, T. Wang, *Annu. Rev. Genomics Hum. Genet.* **22**, 81 (2021).

ACKNOWLEDGMENTS

I thank C. Gunter for critical reading of this work.

10.1126/science.abo5367

RETROSPECTIVE

Paul Farmer (1959–2022)

Humanitarian and public health pioneer

By **Bill Gates**

In 2005, I had the opportunity to visit Paul Farmer, who dedicated his life to helping people access medical care in the world's poorest countries. As Melinda and I walked around the clinic he had helped build in Cange, a small town in central Haiti, it was clear how much Paul's patients loved him. The facility was filled with people who had traveled many miles to receive treatment for diseases such as cholera, tuberculosis, and HIV. Paul seemed to know the names of everyone we passed, often pausing to ask how they were feeling.

At one point, we stopped to talk to a woman who only spoke Haitian Creole. Paul, who was fluent in the language, translated for us. The woman launched into what was obviously a long and effusive story about him—I could hear the words "Doktè Paul" spoken several times. When she finished, I asked Paul what she had said. He blushed as he replied, "Just some obligatory praise for me."

Paul was never comfortable being the center of attention, but it is hard to imagine anyone being more worthy of praise. He was born in North Adams, Massachusetts, in 1959. As a teenager, he and his family spent a summer picking citrus fruit in Florida alongside Haitian immigrants. A few years later, he got to know Haitian farm workers in North Carolina's tobacco fields. These experiences sparked an interest in Haiti that would determine the course of his life.

After graduating from Duke University in 1982 with a degree with medical anthropology, Paul moved to Cange to work in a local hospital. He was horrified to learn that people had to pay up front—an insurmountable barrier for many Haitians—for even the most basic supplies before receiving treatment. Resolved to create a better system, he began studying at Harvard Medical School, splitting his time between Cange and Cambridge. He earned an MD and a PhD in anthropology in 1990. In addition to pursuing his global public health work, he became a professor and later a department chair at Harvard Medical School, positions he held until his death.

In 1987, Paul cofounded the incredible organization that will be his legacy: Partners in Health (PIH). PIH was built on Paul's core belief that everyone deserves good health and access to high-quality services. Today, the organization helps run medical clinics in 12 countries, providing phenomenal care in everything from ophthalmology to women's health to infectious diseases. Paul spent much of his later years traveling from clinic to clinic, working with staff to ensure their patients received the best treatment possible.



A big believer in bringing care directly to people in their homes, Paul structured PIH accordingly. In addition to the medical facilities that PIH runs—each one staffed by locals, providing stable jobs to the community—it oversees a robust network of community health workers. Last year alone, PIH arranged for 2.1 million in-home health visits around the world.

The Gates Foundation has been a proud funder of PIH for more than 16 years, and it has been amazing to watch the organization grow under Paul's leadership. In 2014, Melinda and I traveled back to Haiti to see how Paul's work had changed and grown since the devastating earthquake 4 years earlier. PIH had recently opened a new hospital in Mirebalais that was solar-powered and had all the equipment you would expect to see in a first-class facility, including incubators, fully equipped operating theaters, and a CT scanner. It was remarkable to see Paul's vision come to life: a world where your economic status does not limit your access to high-quality care.

Although Paul's most lasting impact was on the patients he loved so dearly, it is impossible to ignore his broader influence on the global health community. One of the best examples is his work treating AIDS in Haiti. After the first antiretroviral treatments for the disease became available in the late 1980s, many experts assumed they would not be a feasible solution in poor countries. Paul proved them wrong by going door to door to deliver the medicine and help his patients keep up with their treatment. His work paved the way for future international efforts such as the US President's Emergency Plan for AIDS Relief (PEPFAR), which has saved 20 million lives and prevented millions of HIV infections since 2003.

Paul wrote and published a dozen books throughout his life, including *Reimagining Global Health*, which was based on a course he cotaught at Harvard. He was the recipient of many awards, including the Berggruen Prize for Philosophy and Culture, the American Medical Association's Outstanding International Physician Award, and a John D. and Catherine T. MacArthur Foundation fellowship. But the only time Paul actively sought the spotlight was when he knew he had an opportunity to highlight inequality. He was often asked to give commencement speeches, and I suspect he is the reason a lot of young people have entered careers in public health. He was one of the most inspirational people I have ever met.

No one was a better or more tireless advocate for the health of the poor than Paul. In Tracy Kidder's book *Mountains Beyond Mountains*, Paul is quoted as saying, "There's always somebody not getting treatment. I can't stand that." Paul died in Butaro, Rwanda, where he was teaching students at the University of Global Health Equity, which is run by PIH. He spent his last days treating patients at the nearby hospital and educating the next generation of global health leaders. His work will continue through them and through Partners in Health and the many people he trained and inspired over the years. ■

Bill and Melinda Gates Foundation, Seattle, WA 98109, USA.
Email: info@gatesfoundation.org

10.1126/science.abq0061

PHOTO: VICTOR BOYKO/GETTY IMAGES FOR AURORA HUMANITARIAN INITIATIVE

POLICY FORUM

DIVERSITY

A push for inclusive data collection in STEM organizations

Professional societies could better survey, and thus better serve, underrepresented groups

By **Nicholas P. Burnett[1], Alyssa M. Hernandez[2], Emily E. King[3], Richelle L. Tanner[4], Kathryn Wilsterman[5,6]**

Professional organizations in science, technology, engineering, and mathematics (STEM) are well-positioned to improve the recruitment and retention (R&R) of underrepresented groups (*1*, *2*) by providing targeted professional development, networking opportunities, and political advocacy (*3*, *4*). Tailoring these initiatives to specific underrepresented groups can enhance their impact (*5*), but this is predicated on organizations knowing their demographic make-up (*6*). Here, we report patterns in STEM organizations' collection and usage of demographic data from members and conference attendees, based on information from 73 professional societies representing 712,000 constituents. In light of inconsistencies and limitations that we observed, we suggest survey programs that can serve as models for inclusive survey designs by organizations and, where possible, provide demographic information for benchmarking relative to the general population. With improved surveys, organizations can leverage demographic data to prioritize and evaluate R&R efforts, and share effective strategies for R&R of underrepresented groups across STEM.

Baseline demographic data, when compared to the general population (across STEM or across a country), can help organizations set and prioritize R&R goals and, when monitored over time, help organizations evaluate the effectiveness of R&R efforts. Government agencies often provide the most relevant demographic data for a general population in STEM (e.g., the US National Science Foundation) and in a country (e.g., the US Census Bureau) because of their surveys' large sample sizes and broad distributions. As a result, organizations may be compelled to use government surveys as a model for demographic survey design and for benchmarking (i.e., comparing their organization's demographic diversity to the general population). Although these agencies survey many categories of demographic information (e.g., gender identity, family status, citizenship, abilities, race and ethnicity), they do not necessarily collect all the demographic information that is considered meaningful to describe the STEM community—i.e., treating some groups as homogeneous (*7*) and ignoring other groups completely (*8*). By contrast, inclusive demographic surveys acknowledge the full diversity of identities that are meaningful among members of the STEM community. Thus, organizations seeking to describe their demographic composition are pressured to choose between following the examples of government agencies versus creating new, inclusive surveys to recognize additional (and evolving) identities within the STEM community.

A SURVEY OF SURVEYS

We surveyed 164 STEM organizations (73 responses, rate = 44.5%) between December 2020 and July 2021 with the goal of understanding what demographic data each organization collects from its constituents (i.e., members and conference attendees) and how the data are used. See supplementary materials (SM) for more details on the questionnaire. Organizations were sourced from a list of professional societies affiliated with the American Association for the Advancement of Science (AAAS, the publisher of *Science*) (n = 156) or from social media (n = 8). The survey was sent to the elected leadership and management firms for each organization, and follow-up reminders were sent after 1 month. Although these organizations can have international memberships of up to 40% (*6*), we focus our study on the demographic data that are culturally important in the US, under the assumption that US-affiliated organizations usually hold events in the US and thus the relevant demographic context is US based.

The responding organizations represented a wide range of fields: 31 life science organizations (157,000 constituents), 5 mathematics organizations (93,000 constituents), 16 physical science organizations (207,000 constituents), 7 technology organizations (124,000 constituents), and 14 multidisciplinary organizations spanning multiple branches of STEM (131,000 constituents). A list of the responding organizations is available in the SM.

Our survey may have selected for organizations that are committed to, or are interested in, collecting demographic data. However, based on the AAAS-affiliated recruitment of the organizations and the similar sizes of constituencies across STEM fields, we conclude that the responding organizations are likely a reasonably representative cross-section of the most prominent STEM organizations in the US. Each organization was asked about the demographic information that they collect from their constituents, the response rates to their surveys, and how the data were used. Organizations participated under the condition that no data would be associated with them directly. Hence we report aggregated findings and de-identified organization-level data, and do not report any group- or discipline-specific patterns to minimize potential for identification.

Most STEM organizations (59 out of 73, or 80.8%) collect demographic information from their constituents. Commonly surveyed demographic categories included sexual orientation, disability status, racial and ethnic identity, and gender identity (see the top panel of the figure). The number of options offered for each demographic category varied among organizations, resulting in datasets with different resolution and validity (i.e., how accurately the response options reflect the identities of the respondents). Specific response options for each demographic category are shown (see the box) and provided in the SM.

Of the organizations that provided response rates to their surveys (n = 22), the average response rate was 36.1% (SD = 30.4%), which is close to response rates reported by other organizations (*9*). When asked the year of their most recent demographic survey, 29 out of 59 organizations (49.2%) indicated either 2020 or 2021, 7 (11.9%) indicated a year from 2012 to 2019, and 15 (25.4%) indicated that they collect demographic information on a rolling basis with member registration.

Of the organizations that collected demographic data and shared details of their data usage (n = 48), 87.5% reported using demo-

[1]Department of Neurobiology, Physiology, and Behavior, University of California at Davis, Davis, CA, USA. [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [3]Department of Integrative Biology, University of California at Berkeley, Berkeley, CA, USA. [4]Environmental Science and Policy Program, Chapman University, Orange, CA, USA. [5]Division of Biological Sciences, University of Montana, Missoula, MT, USA. [6]Department of Biology, Colorado State University, Fort Collins, CO, USA. Email: burnettnp@gmail.com

graphic data for one or more purposes that fell in the general categories of temporal monitoring, resource planning (e.g., for conferences), publishing reports (e.g., internal or external reports summarizing organizational growth), writing grant proposals, and contributing to third-party research (see the bottom panel of the figure). Some respondents provided specific examples, which included using data to create statistical reports for an organization's board of directors, writing proposals and progress reports to federal funding agencies, and ensuring diverse representation on organizational service committees and speaking panels. Thus, there are disparities between STEM organizations in the underlying design of demographics questions, the administration of surveys, and the usage of the collected data.

Our results indicate that some STEM organizations do not recognize entire groups in STEM, including individuals in sexual minorities (i.e., LGBTQ+ people) or individuals with disabilities (see the top panel of the figure). This observation is surprising given the well-documented discrimination and underrepresentation of these groups in STEM (*8*, *10*). Furthermore, variation in response options on demographic questions—e.g., for racial and ethnic identity and gender identity—signals that only a fraction of organizations aim to give a voice to distinctive identities that are frequently relegated to broader demographic classifications. For instance, numerous Asian American and Pacific Islander (AAPI) ethnic groups are often consolidated into a single "Asian" or "AAPI" grouping (*7*).

In response to survey designs that ignore or obfuscate demographic identities, individuals from underrepresented groups may elect not to respond to certain questions or elect not to complete the survey, introducing a nonresponse bias into the data. As a result, the collected data do not directly reflect the true demographic composition of the organizations. Nonresponse biases may become even more prominent if surveys lack anonymity—e.g., data collected during member registration are likely linked to a respondent's identity (*11*, *12*). Cumulatively, these findings suggest that the bulk of professional organizations in STEM are not collecting demographic data that are representative of the true diversity within STEM, which misinforms any subsequent use of the data for supporting or guiding organizational operations such as R&R.

GUIDES FOR SURVEY DESIGN

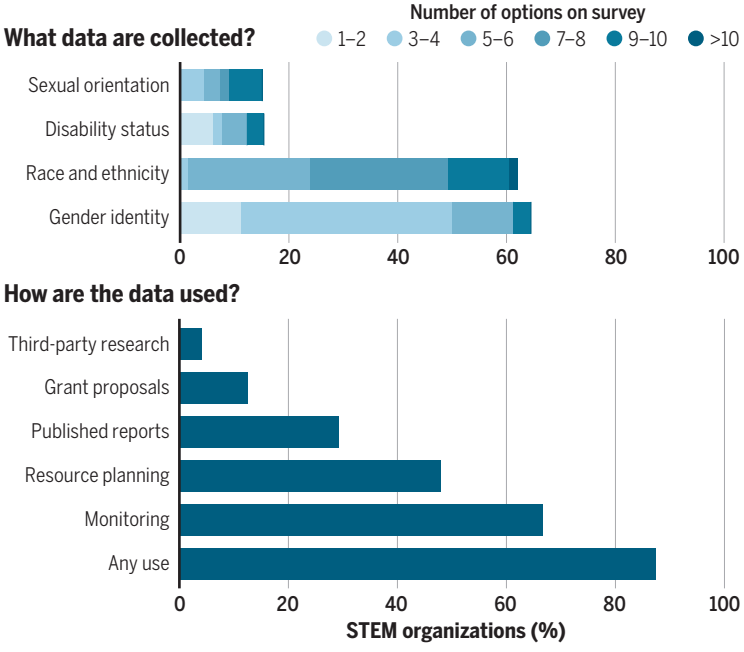
Our data and the recommendations made here are based on a subsample of US-affiliated STEM organizations, and thus they are limited in part by the demographic categories that are prominent and culturally important in the US. Nonetheless, we believe that the principles outlined below are applicable to both US and international organizations. STEM organizations can look to national surveying programs with publicly available data to model demographic survey designs and provide benchmarking of organizational diversity relative to the general population (*13*), though these national programs may have flaws. In the US, three such programs are the National Science Foundation's Survey of Earned Doctorates (SED), the US Census Bureau's American Community Survey (ACS), and the Centers for Disease Control and Prevention's National Health Interview Survey (NHIS).

The SED targets individuals receiving research doctorates in the US (~55,000 per year), whereas the ACS targets the general US population (~3.5 million households per year). Thus, the SED and ACS can provide relevant benchmarking data for demographic surveys of STEM organizations, but the diversity of identities recognized in their survey questions pales in comparison to the questions of the smaller NHIS (~32,000 to 59,000 households per year, 2015 to 2020). For example, in questions on race, the NHIS and ACS each recognize seven distinctive Asian identities, whereas the SED recognizes only a singular Asian group. For ethnicity, the NHIS recognizes seven distinctive Hispanic or Latinx identities, whereas the ACS and SED each recognize only four. Organizations wishing to describe international constituents should avoid describing racial groups as "American" (e.g., "Asian American") and consider asking for country of residence (*6*). However, "residence" can have several definitions, such as legal versus historical, so surveys should be explicit in their use of this term.

In addition to race and ethnicity, the NHIS asks sexual orientation (SED and ACS do not explicitly do this), and the survey includes a series of questions regarding cognitive, motor, visual, and auditory abilities that are more comprehensive in scope and response options than the abilities-related questions of the SED or ACS. Thus, for racial and ethnic identity, sexual orientation, and disability status, the NHIS is likely the most effective, all-in-one guide for question design and benchmarking by STEM organizations. Unfortunately, the SED, ACS, and NHIS do not ask questions explicitly related to gender or transgender identity. Programs that survey these categories do not, to our knowledge, publicly release data that would be helpful for benchmarking by STEM organizations, but their survey questions can still act as guides. For instance, Indiana University's National Survey of Student Engagement provides a model for gender identity questions (with additional, nonbinary response options), and a survey from the advocacy group, National Center for Transgender Equality, provides a model for transgender identity questions. Outside of the US, orga-

Demographic data collection and usage in STEM

We obtained information about demographic data collection and usage from 73 STEM organizations, representing 712,000 members and conference attendees. Organizations most commonly collected race, ethnicity, and gender identity information but with different resolution and validity (top). Demographic data were commonly used for monitoring and resource planning (bottom). "Any use" refers to one or more of the individually listed uses.



GRAPHIC: K. FRANKLIN/*SCIENCE*

nizations can follow these same guidelines by identifying large-scale, local survey programs that collect and report meaningful data from the general population and modifying the survey as needed to provide inclusive questions and options.

STEM organizations may wish to survey demographic categories beyond those discussed here, which were limited to the four most surveyed categories observed in our dataset (see the top panel of the figure). We encourage organizational leaders to find inclusive guides and benchmarking data for additional categories using reputable sources, such as the national survey programs described above. And just as national surveys evolve over time, organizational leaders should openly include new questions and response options to reflect and recognize levels of diversity that are important to the STEM community.

Adding new, inclusive survey elements does not necessarily compromise long-term studies of organizational diversity (i.e., continuity between surveys). In many cases, the inclusive suggestions described here involve recognizing distinctive identities from within groups historically treated as homogeneous. If needed, data from the updated (inclusive) surveys may be compared to historical values by statistically aggregating these newly recognized groups. In other cases, new identities and dimensions of diversity are recognized. Continuity between surveys is then less important because historical surveys were not collecting information that is now considered correct or meaningful to the community.

Inclusive survey questions and response options alone do not guarantee representative demographic data. Other aspects of surveys can prompt or prevent entire groups from responding, resulting in nonresponse biases. Response rates in general can be improved—and nonresponse bias reduced—by ensuring anonymity (e.g., delinking surveys from member registration), sending reminders, minimizing survey length, and providing incentives (*11, 14*). Furthermore, for sensitive questions, response rates may also benefit from providing a justification for data collection, such as that the results will influence specific R&R initiatives.

It is our hope that the guides presented here will enable more STEM organizations to quantify demographic diversity among their constituencies and use the data to inform and evaluate R&R efforts. To improve capacity to compare the efficacy of R&R efforts, partnering STEM organizations may wish to develop compatible frameworks for data collection by first identifying a common survey program for survey design and benchmarking data, and then agreeing on the demographic questions and response options to ask constituents. With this collaborative approach to survey design, organizations can then share successful strategies for the R&R of specific groups, resulting in impactful and widespread support of underrepresented groups across STEM.

Future work in this field should seek to understand how organizational recognition of diverse identities influences the R&R of underrepresented groups, and how this relationship varies across disciplines, organization types, and world regions. For instance, the analysis presented here is an aggregate snapshot of US-affiliated organizations across all areas of STEM. Do some STEM disciplines stand out in their approach to demographic data collection and R&R—e.g., in ways that could be good examples for other disciplines? Do for-profit organizations approach demographic diversity in ways that could inform R&R strategies in other organizations? Detailed assessments in these areas could identify and refine key strategies for demographic data collection and usage, and the use of these surveys for assessing R&R more broadly. Disseminating these findings to all STEM organizations will facilitate society-led support of underrepresented groups on a broad scale.

Different surveys that ask about disability status offer different response options

Surveyed STEM organizations provided varying sets of options for responding to a question about disability status on demographic surveys. See supplementary materials for response options provided for other questions regarding racial and ethnic identity, gender identity, and sexual orientation.

Survey A (ten options)
Sensory impairment (vision or hearing) • Learning disability (e.g., ADHD, dyslexia) • Autism spectrum disorder • Long-term medical illness (e.g., epilepsy, cystic fibrosis) • Speech or language impairment • Mobility limitation or orthopedic impairment • Mental health disorder (e.g., Major depressive disorder) • Temporary impairment due to illness or injury • Disability or impairment not listed - diagnosed or undiagnosed • I do not identify with a disability or impairment

Survey B (ten options)
Deaf/hard of hearing • Visual impairment • Mobility impairment • Cognitive or learning disability • Mental health diagnosis • Neuroatypical • Autoimmune or pain disorder • Moderate to severe allergies/asthma/environmental sensitivities • Other • None of the above

Survey C (six options)
Difficulty seeing • Hearing • Speaking • Walking • Lifting/carrying • Concentrating/remembering

Survey D (five options)
I have serious difficulty walking or climbing stairs • I am deaf or have serious difficulty hearing • I have been diagnosed by a health professional as having permanent memory loss or learning disability • I am blind or have difficulty seeing even when wearing glasses • None of the above

Survey E (five options)
Hearing impairment • Vision impairment • Learning disability • Mobility/orthopedic impairment • Other

Survey F (three options)
Yes • No • Specify if you wish

Survey G (two options)
Yes • No

REFERENCES AND NOTES

1. N. Sellami, B. Toven-Lindsey, M. Levis-Fitzgerald, P. H. Barber, T. Hasson, *CBE Life Sci. Educ.* **20**, ar11 (2021).
2. National Science Board, "Science and engineering indicators: The state of U.S. science and engineering 2020" (NSB-2020-1, National Science Foundation, Alexandria, VA, 2020); https://ncses.nsf.gov/pubs/nsb20201/.
3. J. M. Scott, J. L. Rachlow, R. T. Lackey, *Bioscience* **58**, 865 (2008).
4. N. P. Burnett *et al.*, *Nat. Ecol. Evol.* **4**, 1283 (2020).
5. M. A. Armstrong, J. Jovanovic, *J. Women Minor. Sci. Eng.* **21**, 141 (2015).
6. C. A. Rushworth *et al.*, *Evolution* **75**, 208 (2021).
7. K. H. Nguyen *et al.*, *Bull. Ecol. Soc. Am.* **103**, e01939 (2022).
8. J. B. Freeman, *Science* **374**, 1333 (2021).
9. Y. Baruch, B. C. Holtom, *Hum. Relat.* **61**, 1139 (2008).
10. National Center for Science and Engineering Statistics, "Women, minorities, and persons with disabilities in science and engineering: 2021" (Special Report NSF 21-321, National Science Foundation, Alexandria, VA, 2021); https://ncses.nsf.gov/wmpd.
11. P. Roth, *J. Manage.* **24**, 97 (1998).
12. J. B. Yoder, A. Mattheis, *J. Homosex.* **63**, 1 (2016).
13. C. M. Snipp, *Annu. Rev. Sociol.* **29**, 563 (2003).
14. S. J. Sills, C. Song, *Soc. Sci. Comput. Rev.* **20**, 22 (2002).

ACKNOWLEDGMENTS

We thank T. Burnette, S. Cheng, R. Mehta, and M. Salcedo for constructive comments on earlier versions of this manuscript. Prior to this study, N.P.B. and A.M.H. assisted, unpaid, with developing the demographic survey for the Society for Integrative and Comparative Biology (SICB), which is included in the anonymous dataset. N.P.B. and A.M.H. currently hold unpaid leadership positions in SICB.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo1599

10.1126/science.abo1599



Benjamin Franklin poses in the studio of French sculptor Jean Antoine Houdon.

SCIENCE LIVES

The scientist-statesman

A new documentary capably navigates the life and legacy of Benjamin Franklin

By **R. B. Bernstein**

Based on a well-crafted, thoughtful script by Dayton Duncan and ably narrated by the actor Peter Coyote, Ken Burns's new documentary, *Benjamin Franklin*, features enlightening interviews with historians and biographers including Walter Isaacson, Erica Armstrong Dunbar, Christopher Brown, Sheila Skemp, Stacy Schiff, Joyce Chaplin, Joseph Ellis, Clay Jenkinson, Gordon Wood, William Leuchtenburg, and the late Bernard Bailyn. The actor Mandy Patinkin brings Franklin's words to life, with other well-known actors such as Liam Neeson and Paul Giamatti voicing supporting characters. An effective musical score drawing on 18th-century European composers undergirds the documentary, and its equally rich visuals depict 18th-century Boston, Philadelphia, London, and Paris, along with other places linked with Franklin's life.

Benjamin Franklin
Ken Burns, executive producer
Florentine Films, 2022.
240 minutes.

Benjamin Franklin has two parts, the first tracing Franklin's life in colonial Boston and Philadelphia (with European interludes) and the second following him through the American Revolution, his diplomacy efforts in the 1770s and 1780s, his service at the Federal Convention in Philadelphia in 1787, and his final years, concluding with his death in 1790 at the age of 84. This project blends its narrative of Franklin's life with probing, thoughtful explanations of why his life mattered—to him, to his contemporaries, and to posterity.

Part 1 begins with Benjamin's birth in Boston in 1706, the 16th child and youngest son of Josiah Franklin, a Boston candlemaker, and Josiah's second wife, Abiah Folger Franklin. A precocious child, Benjamin was apprenticed to his older brother James as a printer and, early on, demonstrated a talent for writing and an irrepressible independence.

At age 17, Franklin ran away from his dictatorial brother, eventually settling in Philadelphia, where he pursued an increasingly busy and successful career as a printer. With only 2 years of formal schooling, he became the leading printer in colonial America, a successful craftsman and man of business, and one of the leading American writers of his time. He was known in particular for his perennially successful *Poor Richard's Almanack*, which he published from 1732 to 1758.

As the first distinguished American scientist, he pioneered research in electricity and frequently invented or adapted technological advances, including the Franklin stove, the lightning rod, and bifocal spectacles. He mapped and explained scientifically the Gulf Stream and also wrote a phenomenally successful autobiography (*1*).

Franklin broadened his public spirit to include activism in colonial and British imperial politics, a part of his life bridging parts 1 and 2 of this documentary. A committed imperialist, he grew frustrated and saddened by the widening rift between American colonists and the mother country. Burns is especially effective in telling this part of Franklin's story and in charting Franklin's abrupt shift from supporting the British Empire to embracing the American cause, ultimately rupturing beyond repair his relationship with his illegitimate son William, who had become the royal governor of New Jersey. Franklin never forgave William for his loyalism, the most anguished of Franklin's failures in the realm of family.

As part 2 shows, Franklin was far more successful as an American politician. He played a key role in the American declaration of independence, advising Thomas Jefferson's writing of the Declaration. The first American diplomat, he was American minister to France, securing the new nation's alliance with France and then, along with John Jay and John Adams, negotiating the Treaty of Paris with Great Britain that recognized American independence and doubled the size of the new nation. And, despite his age and frailty, Franklin was a highly effective broker of compromises and voice of reasoned moderation during the Federal Convention's struggles to frame the US Constitution.

Like the 2002 Middlemarch documentary on Franklin (*2*), Burns's film is most successful in challenging older scholarship's celebration of Franklin as what his classic biographer Carl Van Doren called "a harmonious human multitude." Burns's Franklin, by contrast, is evolving and conflicted, a man of conventional white supremacist views who radically pivots in his perceptions of African Americans and the evils of slavery, a man who shows the conflicts and inconsistencies of many of his contemporaries on matters of sex and gender, a proud Briton who in his sixties and seventies becomes an equally proud American founding father. Burns gives us a Benjamin Franklin for our own troubled times, an extraordinary and complicated man who never lost sight of what it means to be human. ■

The reviewer is at the Colin Powell School for Civic and Global Leadership, The City College of New York, New York, NY 10031, USA. Email: rbernstein@ccny.cuny.edu

REFERENCES AND NOTES

1. B. Franklin, *The Autobiography of Benjamin Franklin* (J. Parson's, 1793).
2. M. Meyer, E. Hovde, dirs., *Benjamin Franklin* (Middlemarch Films, 2002).

10.1126/science.abo2776

IMAGE: COURTESY OF DIPLOMATIC RECEPTION ROOMS, US DEPARTMENT OF STATE

SCIENCE AND SOCIETY

War in an automated, data-driven world

An anthropologist misses the mark on a topic of great concern

By **George Lucas**

The topics addressed by Roberto González in his new book, *War Virtually*, are ones that ought to concern us all. These include the intersections of Big Data and Big Tech and the battlefield, lethal autonomous weapons systems, cyber conflict, and the enhancement of such technologies using increasingly sophisticated artificial intelligence.

González is an anthropologist specializing in the Zapotec culture in the northern regions of Oaxaca. He has also been a vocal critic, along with many other social scientists, of attempts by the US Army and NATO allies to exploit the expertise of social scientists with knowledge of local cultures during the counterinsurgency and antiterrorist campaigns in Iraq and Afghanistan by means of a program known as the Human Terrain System (HTS).

Authors of works on military technology need not be physical scientists or engineers. The political scientist and 21st-century warfare expert Peter Singer, whom González cites, for example, is neither. And yet Singer is remarkably adept at providing a narrative framework that encompasses both the promises and perils of the efforts of such individuals and that is intelligible and informative for a wide audience.

Meanwhile, the uses and misuses of social science and its practitioners is a controversial topic, on which the author is an unquestioned authority. However, his decision to examine what the National Academy of Sciences recently termed "exotic" new military technologies through an interpretive framework constructed by examining the work of anthropologists Margaret Mead and Gregory Bateson during World War II and the Phoenix Program—a project intended to infiltrate and ultimately destroy the Viet Cong during the Vietnam War—is puzzling. One is left to wonder how such reflections, interwoven with the author's discussions of other historical collaborations between social scientists and government and military entities, intersect with emerging technologies.

Applying anthropological methodologies to contemporary issues such as cyber conflict, the social impacts of artificial intelligence, and military applications of social media is a promising, but nascent, area of research. However, apart from a discussion of the Facebook Cambridge Analytica scandal—an egregious violation of personal privacy, but hardly one that constituted war or the "militarization" of data—the author neglects such topics and gives short shrift to the work of fellow anthropologists such as Gabriella Coleman, Mizuko Ito, Daniel Miller, and Genevieve Bell, who have done pioneering work on these subjects.

There is already a spirited and contentious discussion concerning the wisdom (not to mention legality and morality) of the growing international reliance on emerging military technologies. Noel Sharkey—a world-renowned computer scientist and roboticist whom the author unfortunately fails to cite—is, like González, highly critical of the naïve reductionism inherent in much artificial intelligence research. He, too, is deeply skeptical of our ability to equip armed autonomous weapons with the requisite capacities to operate reliably in the battlespace and, together with philosopher Peter Asaro, cofounded and leads the International Committee for Robot Arms Control, whose membership has worked tirelessly for well over a decade to pressure the United Nations and the International Red Cross to ban outright any further development or deployment of lethal autonomous military technologies.

This vitally important debate has also been thoroughly and impartially documented by tech-savvy journalists such as Ellen Nakashima and Noah Shachtman. Rather than seeking common cause with like-minded experts and organizations, however, González leans into his deep mistrust of all things military and his apparent nostalgia for HTS debates. Readers interested in the topic of tech-enabled warfare would be better served by consulting the scholarship of subject-matter experts such as Sharkey or Paul Scharre (*1*, *2*). ■

The reviewer is a professor emeritus at the US Naval Academy, Annapolis, MD 21402, USA.
Email: george.r.lucas.jr@gmail.com

War Virtually: The Quest to Automate Conflict, Militarize Data, and Predict the Future
Roberto J. González
University of California Press, 2022. 270 pp.

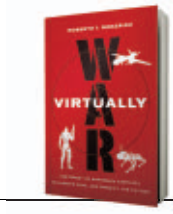




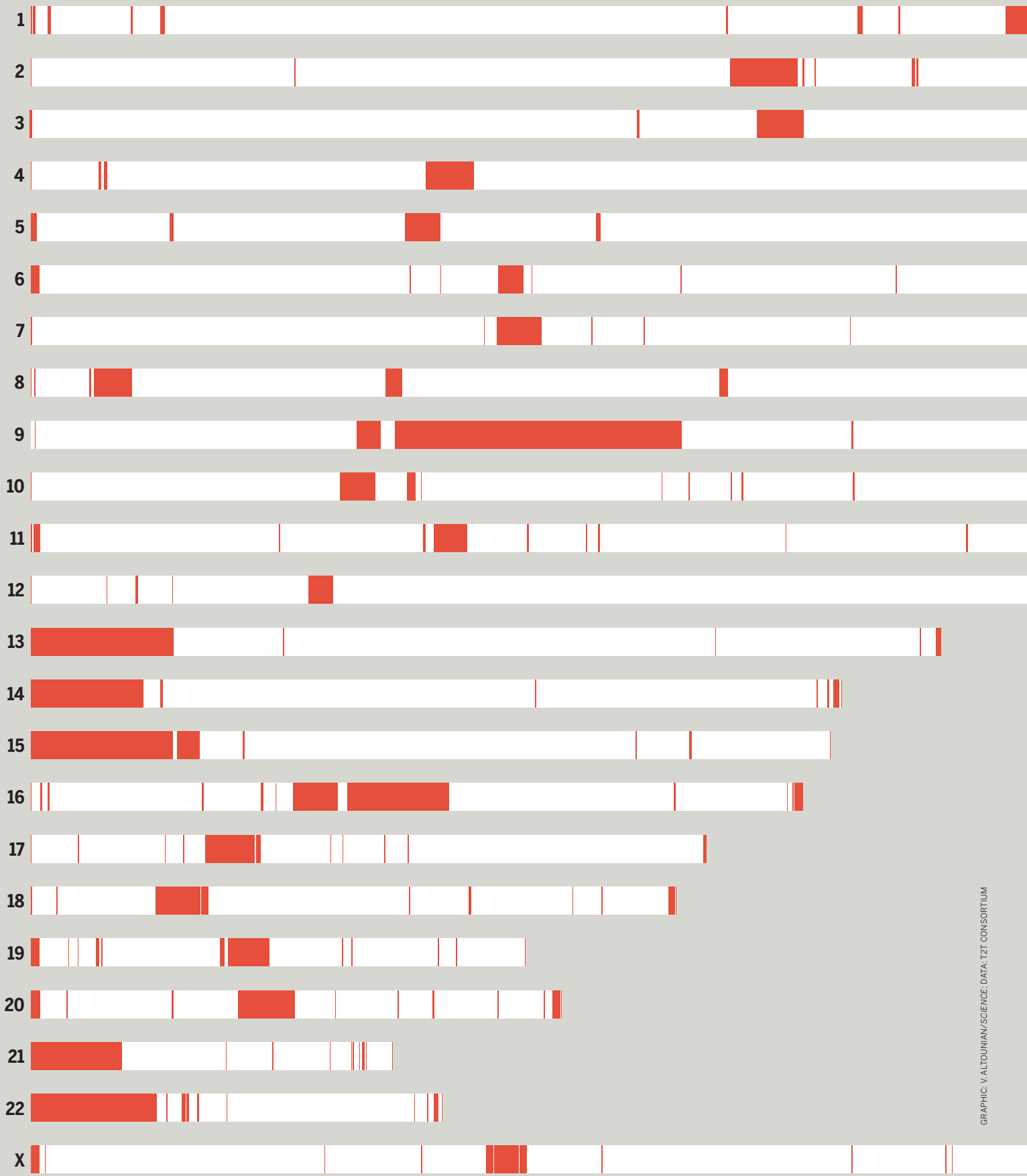
PHOTO: COURTESY OF US MARINE CORPS/SARAH DIETZ

A Legged Squad Support System (LS3) accompanies US Marines on patrol.

REFERENCES AND NOTES

1. N. Sharkey, in *Military Ethics and Emerging Technologies*, T. J. Demy, G. R. Lucas Jr., B. J. Strawser, eds. (Routledge, 2014), p. 132.
2. P. Scharre, *Army of None: Autonomous Weapons and the Future of War* (Norton, 2018).

10.1126/science.abo4591



GRAPHIC: V. ALTOUNIAN/*SCIENCE*; DATA: T2T CONSORTIUM



FILLING THE GAPS

By **Laura M. Zahn**

A fully sequenced human genome was triumphantly announced more than 20 years ago. However, owing to technological limitations, some genomic regions remained unresolved. Here, *Science* presents research by the Telomere-to-Telomere (T2T) Consortium, reporting on the endeavor to complete a comprehensive human reference genome. Generated primarily by long-read sequencing of a hydatidiform mole, a doubly haploid growth, this effort adds ~200 megabases of genetic information—a full chromosome's worth—to the human genome.

Through the resolution of previously unsequenceable and unalignable regions, mostly composed of highly repetitive sequences, this reference genome allows for a detailed characterization of the centromeric satellite repeats, transposable elements, and segmental duplications. Mapping of genomic sequences, including those from previously published studies, resolves aspects of human genetic diversity, including evolutionary comparisons with our primate relatives. Furthermore, it allows for identification of how changes in methylation density differ within and among centromeres and how epigenetics can affect the transcription of repeat sequences.

These investigations have only begun to tease apart how the T2T reference genome influences the detection of biomedically relevant variants and the evolution of genomic regions that determine human traits. Although much remains to be discovered, the T2T reference genome provides another celebratory benchmark to observe as we continue to delve into the genetics that underlie our complete selves.

Access T2T papers published in *Science* and companion papers at science.org/collections/completing-human-genome.

RESEARCH ARTICLES

The complete sequence of a human genome p. 44

A complete reference genome improves analysis of human genetic variation p. 54

Segmental duplications and their variation in a complete human genome p. 55

Complete genomic and epigenetic maps of human centromeres p. 56

From telomere to telomere: The transcriptional and epigenetic state of human repeat elements p. 57

Epigenetic patterns in a complete human genome p. 58

RELATED ITEMS

NEWS STORY p. 15
PERSPECTIVE p. 34 PODCAST

Each bar is a linear visualization of a chromosome, with the chromosome number shown at left. Red segments denote previously missing sequences that the T2T Consortium resolved.

RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

Sergey Nurk[1]†, Sergey Koren[1]†, Arang Rhie[1]†, Mikko Rautiainen[1]†, Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9]‡, Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28], Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[9,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[41], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[44], Ivan Sovi[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[44], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,53]*, Karen H. Miga[11,54]*, Adam M. Phillippy[1]*

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

The current human reference genome was released by the Genome Reference Consortium (GRC) in 2013 and most recently patched in 2019 (GRCh38.p13) (*1*). This reference traces its origin to the publicly funded Human Genome Project (*2*) and has been continually improved over the past two decades. Unlike the competing Celera effort (*3*) and most modern sequencing projects based on "shotgun" sequence assembly (*4*), the GRC assembly was constructed from sequenced bacterial artificial chromosomes (BACs) that were ordered and oriented along the human genome by means of radiation hybrid, genetic linkage, and fingerprint maps. However, limitations of BAC cloning led to an underrepresentation of repetitive sequences, and the opportunistic assembly of BACs derived from multiple individuals resulted in a mosaic of haplotypes. As a result, several GRC assembly gaps are unsolvable because of incompatible structural polymorphisms on their flanks, and many other repetitive and polymorphic regions were left unfinished or incorrectly assembled (*5*).

The GRCh38 reference assembly contains 151 mega–base pairs (Mbp) of unknown sequence distributed throughout the genome, including pericentromeric and subtelomeric regions, recent segmental duplications, ampliconic gene arrays, and ribosomal DNA (rDNA) arrays, all of which are necessary for fundamental cellular processes (Fig. 1A). Some of the largest reference gaps include human satellite (HSat) repeat arrays and the short arms of all five acrocentric chromosomes, which are represented in GRCh38 as multimegabase stretches of unknown bases (Fig. 1, B and C). In addition to these apparent gaps, other regions of GRCh38 are artificial or are otherwise incorrect. For example, the centromeric alpha satellite arrays are represented as computationally generated models of alpha satellite monomers to serve as decoys for resequencing analyses (*6*), and sequence assigned to the short arm of chromosome 21 appears falsely duplicated and poorly assembled (*7*). When compared with other human genomes, GRCh38 also shows a genome-wide deletion bias that is indicative of incomplete assembly (*8*). Despite finishing efforts from both the Human Genome Project (*9*) and GRC (*1*) that improved the quality of the reference, there was limited

[1]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [2]Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA. [3]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia. [4]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [5]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA. [6]Sirius University of Science and Technology, Sochi, Russia. [7]Vavilov Institute of General Genetics, Moscow, Russia. [8]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. [9]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [10]Institute for Systems Genomics and Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [11]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [12]University of Geneva Medical School, Geneva, Switzerland. [13]Stowers Institute for Medical Research, Kansas City, MO, USA. [14]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [15]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [16]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. [17]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [18]DNAnexus, Mountain View, CA, USA. [19]Wellcome Sanger Institute, Cambridge, UK. [20]Department of Genetics, University of Cambridge, Cambridge, UK. [21]Inscripta, Boulder, CO, USA. [22]Laboratory of Neurogenetics of Language and The Vertebrate Genome Lab, The Rockefeller University, New York, NY, USA. [23]Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. [24]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [25]University of Tennessee Health Science Center, Memphis, TN, USA. [26]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO, USA. [27]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA. [28]Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [29]Pacific Biosciences, Menlo Park, CA, USA. [30]Department of Computational and Data Sciences, Indian Institute of Science, Bangalore KA, India. [31]Reservoir Genomics LLC, Oakland, CA, USA. [32]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [33]Undiagnosed Diseases Program, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [34]Heinrich Heine University Düsseldorf, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany. [35]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA. [36]Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA. [37]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [38]Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA. [39]Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia. [40]Cancer Institute of New Jersey, New Brunswick, NJ, USA. [41]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [42]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [43]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [44]Genome Center, MIND Institute, Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA. [45]Institute for Systems Biology, Seattle, WA, USA. [46]Digital BioLogic d.o.o., Ivanić-Grad, Croatia. [47]Chan Zuckerberg Biohub, San Francisco, CA, USA. [48]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC, USA. [49]Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [50]Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA. [51]Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia. [52]Department of Biochemistry and Molecular Biology, University of Kansas Medical School, Kansas City, MO, USA. [53]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [54]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA.
***Corresponding author. Email: eee@gs.washington.edu (E.E.E.); khmiga@ucsc.edu (K.H.M.); adam.phillippy@nih.gov (A.M.P.)**
†These authors contributed equally to this work. ‡Present address: Oxford Nanopore Technologies Inc., Lexington, MA, USA.

progress toward closing the remaining gaps in the years that followed (Fig. 1D).

Long-read shotgun sequencing overcomes the limitations of BAC-based assembly and bypasses the challenges of structural polymorphism between genomes. PacBio's multikilobase, single-molecule reads (*10*) proved capable of resolving complex structural variation and gaps in GRCh38 (*8*, *11*), whereas Oxford Nanopore's >100-kbp "ultralong" reads (*12*) enabled complete assemblies of a human centromere (chromosome Y) (*13*) and, later, an entire chromosome (chromosome X) (*14*). However, the high error rate (>5%) of these technologies posed challenges for the assembly of long, near-identical repeat arrays. PacBio's most recent "HiFi" circular consensus sequencing offers a compromise of 20-kbp read lengths with an error rate of 0.1% (*15*). Whereas ultralong reads are useful for spanning repeats, HiFi reads excel at differentiating subtly diverged repeat copies or haplotypes (*16*).

To finish the last remaining regions of the genome, we leveraged the complementary aspects of PacBio HiFi and Oxford Nanopore ultralong-read sequencing to assemble the uniformly homozygous CHM13hTERT cell line (hereafter, CHM13) (*17*). The resulting T2T-CHM13 reference assembly removes a 20-year-old barrier that has hidden 8% of the genome from sequence-based analysis, including all centromeric regions and the entire short arms of five human chromosomes. Here, we describe the construction, validation, and initial analysis of a truly complete human reference genome and discuss its potential impact on the field.

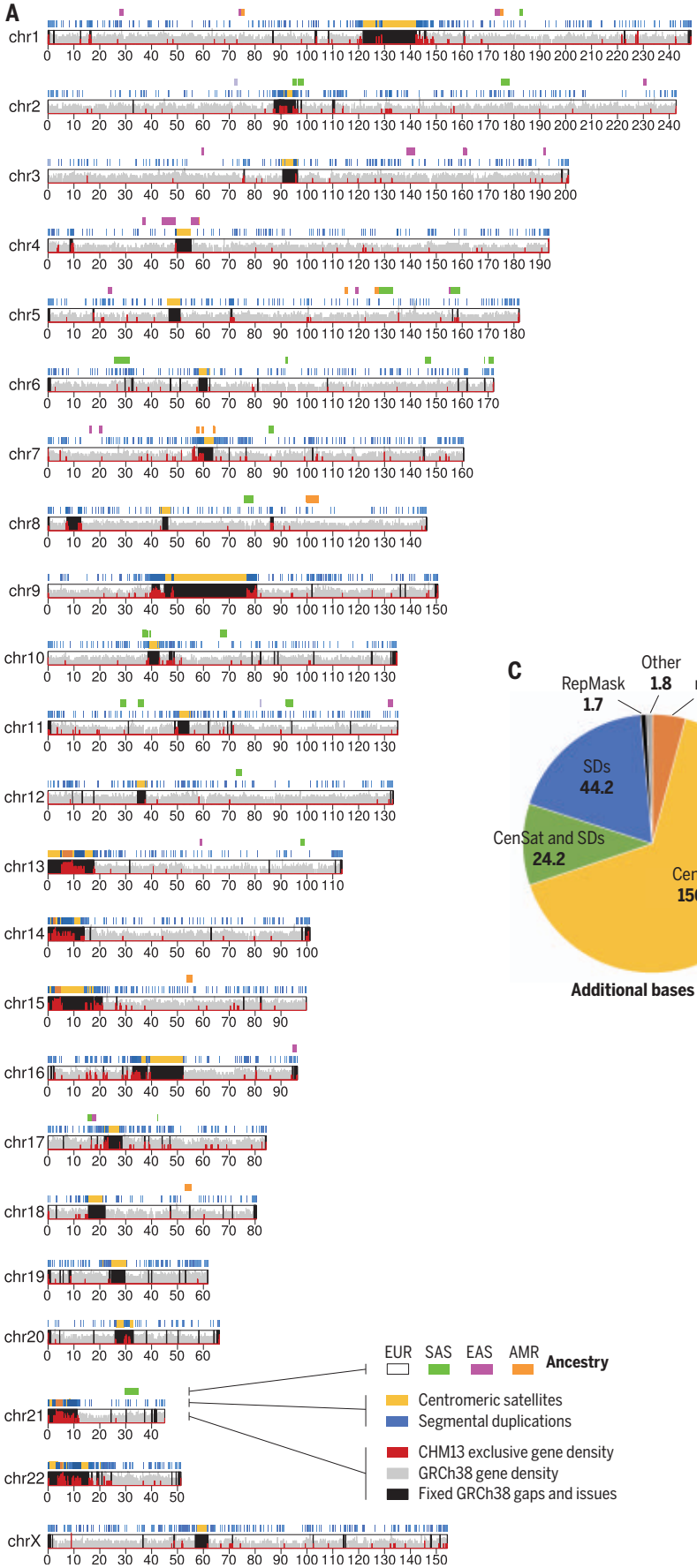
Cell line and sequencing

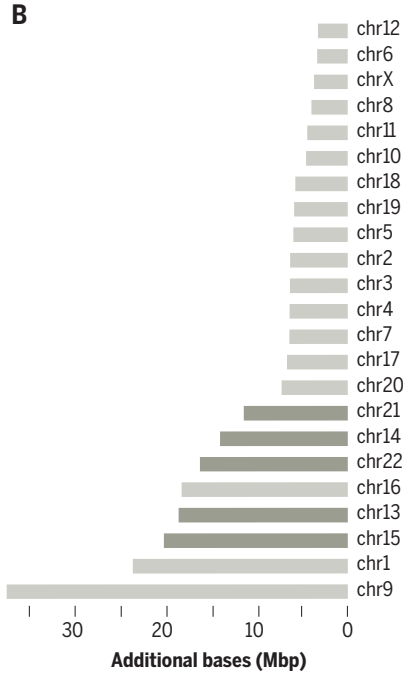
As with many prior reference genome improvement efforts (*1*, *8*, *17*–*20*), including the T2T assemblies of human chromosomes X (*14*) and 8 (*21*), we targeted a complete hydatidiform mole (CHM) for sequencing. Most CHM genomes arise from the loss of the maternal complement and duplication of the paternal complement postfertilization and are, therefore, homozygous with a 46,XX karyotype (*22*). Sequencing of CHM13 confirmed nearly uniform homozygosity, with the exception of a few thousand heterozygous variants and a megabase-scale heterozygous deletion within the rDNA array on chromosome 15 (*23*) (figs. S1 and S2). Local ancestry analysis shows that most of the CHM13 genome is of European origin, including regions of Neanderthal introgression, with some predicted admixture (*23*) (Fig. 1A). Compared with diverse samples from the 1000 Genomes Project (1KGP) (*24*), CHM13 possesses no apparent excess of singleton alleles or loss-of-function variants (*25*).

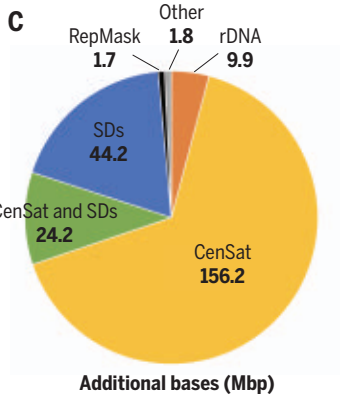
We extensively sequenced CHM13 with multiple technologies (*23*), including 30× PacBio circular consensus sequencing (HiFi) (*16*, *20*), 120× Oxford Nanopore ultralong-read sequencing (ONT) (*14*, *21*), 100× Illumina PCR-Free sequencing (ILMN) (*1*), 70× Illumina Arima Genomics Hi-C (Hi-C) (*14*), BioNano optical maps (*14*), and single-cell DNA template strand sequencing (Strand-seq) (*20*) (table S1). To enable assembly of the highly repetitive centromeric satellite arrays and closely related segmental duplications, we developed methods

Table 1. Comparison of GRCh38 and T2T-CHM13v1.1 human genome assemblies. GRCh38 summary statistics exclude "alts" (110 Mbp), patches (63 Mbp), and chromosome Y (58 Mbp). Assembled bases include all non-N bases. Unplaced bases are those not assigned or positioned within a chromosome. GRCh38 scaffolds were split at three consecutive Ns to obtain the number of contigs. Contig NG50 is the largest value such that contigs of at least this size total more than half of the 3.05-Gbp genome size. The number of exclusive genes or transcripts is as follows: for GRCh38, GENCODE genes and transcripts not found in CHM13; and for CHM13, extra putative paralogs that are not in GENCODE. Segmental duplication analysis is from (*42*). RepeatMasker analysis is from (*49*). Blank spaces indicate not applicable.

STATISTICS	GRCH38	T2T-CHM13	DIFFERENCE (±%)
Summary			
Assembled bases (Gbp)	2.92	3.05	+4.5
Unplaced bases (Mbp)	11.42	0	−100.0
Gap bases (Mbp)	120.31	0	−100.0
Number of contigs	949	24	−97.5
Contig NG50 (Mbp)	56.41	154.26	+173.5
Number of issues	230	46	−80.0
Issues (Mbp)	230.43	8.18	−96.5
Gene annotation			
Number of genes	60,090	63,494	+5.7
Protein coding	19,890	19,969	+0.4
Number of exclusive genes	263	3,604	
Protein coding	63	140	
Number of transcripts	228,597	233,615	+2.2
Protein coding	84,277	86,245	+2.3
Number of exclusive transcripts	1,708	6,693	
Protein coding	829	2,780	
Segmental duplications			
Percentage of segmental duplications (%)	5.00	6.61	
Segmental duplication bases (Mbp)	151.71	201.93	+33.1
Number of segmental duplications	24097	41528	+72.3
RepeatMasker			
Percentage of repeats (%)	51.89	53.94	
Repeat bases (Mbp)	1,516.37	1,647.81	+8.7
Long interspersed nuclear elements	626.33	631.64	+0.8
Short interspersed nuclear elements	386.48	390.27	+1.0
Long terminal repeats	267.52	269.91	+0.9
Satellite	76.51	150.42	+96.6
DNA	108.53	109.35	+0.8
Simple repeat	36.5	77.69	+112.9
Low complexity	6.16	6.44	+4.6
Retroposon	4.51	4.65	+3.3
rRNA	0.21	1.71	+730.4







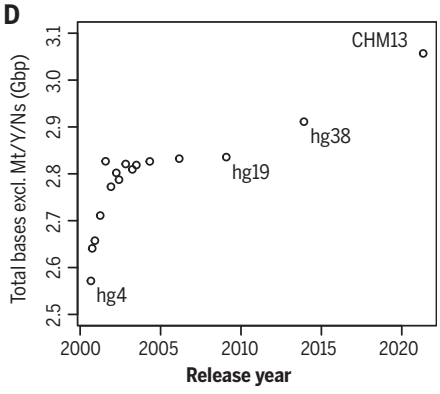


Fig. 1. Summary of the complete T2T-CHM13 human genome assembly. (**A**) Ideogram of T2T-CHM13v1.1 assembly features. For each chromosome (chr), the following information is provided from bottom to top: gaps and issues in GRCh38 fixed by CHM13 overlaid with the density of genes exclusive to CHM13 in red; segmental duplications (SDs) (*42*) and centromeric satellites (CenSat) (*30*); and CHM13 ancestry predictions (EUR, European; SAS, South Asian; EAS, East Asian; AMR, ad-mixed American). Bottom scale is measured in Mbp. (**B** and **C**) Additional (nonsyntenic) bases in the CHM13 assembly relative to GRCh38 per chromosome, with the acrocentrics highlighted in dark gray (B) and by sequence type (C). (Note that the CenSat and SD annotations overlap.) RepMask, RepeatMasker. (**D**) Total nongap bases in UCSC reference genome releases dating back to September 2000 (hg4) and ending with T2T-CHM13 in 2021. Mt/Y/Ns, mitochondria, chrY, and gaps.

for assembly, polishing, and validation that better utilize these available datasets.

Genome assembly

The basis of the T2T-CHM13 assembly is a high-resolution assembly string graph (*26*) built directly from HiFi reads. In a bidirected string graph, nodes represent unambiguously assembled sequences, and edges correspond to the overlaps between them, owing to either repeats or true adjacencies in the underlying genome. The CHM13 graph was constructed using a purpose-built method that combines components from existing assemblers (*16*, *27*) along with specialized graph processing (*23*). Most HiFi errors are small insertions or deletions within homopolymer runs and simple sequence repeats (*16*), so homopolymer runs were first "compressed" to a single nucleotide (e.g., $A_1 \ldots A_n$ becomes A_1 for $n > 1$). All compressed reads were then aligned to one another to identify and correct small errors, and differences within simple sequence repeats were masked. After compression, correction, and masking, only exact read overlaps were considered during graph construction, followed by iterative graph simplification (*23*).

In the resulting graph, most components originate from a single chromosome and have an almost linear structure (Fig. 2A), which suggests that few perfect repeats greater than roughly 10 kbp exist between different chromosomes or distant loci. Two notable exceptions are the five acrocentric chromosomes, which form a single connected component in the graph, and a recent multimegabase HSat3 duplication on chromosome 9, consistent with the 9qh+ karyotype of CHM13 (fig. S3). Minor fragmentation of the chromosomes into multiple components resulted from a lack of HiFi sequencing coverage across GA-rich sequences (*16*). These gaps were later filled with a prior ONT-based assembly (CHM13v0.7) (*14*).

Ideally, the complete sequence for each chromosome should exist as a walk through the string graph where some nodes may be traversed multiple times (repeats) and some not at all (errors and heterozygous variants). To help identify the correct walks, we estimated coverage depth and multiplicity of the nodes (*23*), which allowed most tangles to be manually resolved as unique walks visiting each node the appropriate number of times (Fig. 2B and fig. S4). In the remaining cases, the correct path was ambiguous and required integration of ONT reads (Fig. 2, C and D). Where possible, ONT reads were aligned to candidate traversals or directly to the HiFi graph (*28*) to guide the correct walk (fig. S5), but more elaborate strategies were required for recent satellite array duplications on chromosomes 6 and 9 (*23*). Only the five rDNA arrays, constituting about 10 Mbp of sequence, could not be resolved with the string graph and required a specialized approach (described later). An accurate consensus sequence for the selected graph walks was computed from the uncompressed HiFi reads (*23*), resulting in the CHM13v0.9 draft assembly.

For comparative genomics of the centromere (*29*, *30*), we repeated this process on an additional X chromosome from the Coriell GM24385 cell line [National Institute of Standards and Technology (NIST) ID: HG002]. The resulting T2T-HG002-ChrX assembly shows comparable accuracy to T2T-CHM13 (*23*) (figs. S6 to S8).

rDNA assembly

The most complex region of the CHM13 string graph involves the human rDNA arrays and their surrounding sequence (Fig. 2D). Human rDNAs are 45-kbp near-identical repeats that encode the 45*S* rRNA and are arranged in large, tandem repeat arrays embedded within the short arms of the acrocentric chromosomes. The length of these arrays varies between individuals (*31*) and even somatically, especially with aging and certain cancers (*32*). A typical diploid human genome has an average of 315 rDNA copies, with a standard deviation of 104 copies (*31*). We estimate that the diploid CHM13 genome contains about 400 rDNA copies based on ILMN depth of coverage (*23*) (fig. S9) or 409 ± 9 (mean ± SD) rDNA copies by droplet digital polymerase chain reaction (ddPCR) (fig. S10).

To assemble these highly dynamic regions of the genome and overcome limitations of the string graph construction (*23*) (fig. S11), we constructed sparse de Bruijn graphs for each of the five rDNA arrays (*33*) (fig. S12). ONT reads were aligned to the graphs to identify a set of walks, which were converted to sequence, segmented into individual rDNA units, and clustered into "morphs" according to their sequence similarity. The copy number of each morph was estimated from the number of supporting ONT reads, and consensus sequences were polished with mapped HiFi reads. ONT reads spanning two or more rDNA units were used to build a morph graph representing the structure of each array (fig. S12).

The shorter arrays on chromosomes 14 and 22 consist of a single primary morph arranged in a head-to-tail array, whereas the longer arrays on chromosomes 13, 15, and 21 exhibit a more mosaic structure involving multiple, interspersed morphs. In these cases, the ONT reads were not long enough to fully resolve the ordering, and the primary morphs were artificially arranged in consecutive blocks reflecting their estimated copy number. These three arrays capture the chromosome-specific morphs but should be treated as model sequences. The final T2T-CHM13 assembly contains 219 complete rDNA copies, totaling 9.9 Mbp of sequence.

Assembly validation and polishing

To evaluate concordance between the reads and the assembly, we mapped all available primary data—including HiFi, ONT, ILMN, Strand-seq, and Hi-C—to the CHM13v0.9 draft assembly to identify both small and structural variants [see (*34*) for a complete description]. Manual curation corrected 4 large and 993 small errors, resulting in the CHM13v1.0 assembly, and identified 44 large and 3901 small heterozygous variants (*34*). Further telomere polishing and addition of the rDNA arrays (*23*) resulted in a complete, telomere-to-telomere assembly of a human genome, T2T-CHM13v1.1.

The T2T-CHM13 assembly is consistent with previously validated assemblies of chromosomes X (*14*) and 8 (*21*), and the sizes of assembled satellite arrays match ddPCR copy-number estimates for those tested (fig. S10 and tables S2 and S3). Mapped Strand-seq (figs. S13 and S14) and Hi-C (fig. S15) data show no signs of misorientations or other large-scale structural errors. The assembly correctly resolves 644 of 647 previously sequenced CHM13 BACs at >99.99% identity, with the three others reflecting errors in the BACs themselves (figs. S16 to S19).

Mapped sequencing read depth shows uniform coverage across all chromosomes (Fig. 3A), with 99.86% of the assembly within three standard deviations of the mean coverage for either HiFi or ONT (HiFi coverage 34.70 ± 7.03 and ONT coverage 116.16 ± 16.96, excluding the mitochondrial genome). Ignoring the 10 Mbp of rDNA sequence, where most of the coverage deviation resides, 99.99% of the assembly is within three standard deviations (*23*). Alignment-free analysis of ILMN and HiFi copy-number data also shows concordance with the assembly (figs. S20 and S21). This is consistent with uniform coverage of the genome and confirms both the accuracy of the assembly and the absence of aneuploidy in the sequenced CHM13 cells.

Coverage increases or decreases were observed across multiple satellite arrays (Fig. 3, B to D). However, given the uniformity of coverage across these arrays, association with specific satellite classes, and the sometimes opposite effect observed for HiFi and ONT, we hypothesize that these anomalies are related to biases introduced during sample preparation, sequencing, or base calling, rather than assembly error (*23*) (figs. S22 to S26 and table S4). Although the specific mechanisms require further investigation, prior studies have noted similar biases within certain satellite arrays and sequence contexts for both ONT and HiFi (*35*, *36*).

Because they are the most difficult regions of the genome to assemble, we performed targeted validation of long tandem repeats to

identify any errors missed by the genome-wide approach. The assembled rDNA morphs, being only 45 kbp each, were manually validated by inspection of the read alignments used for polishing. Alpha satellite higher-order repeats (HORs) were validated using a purpose-built method (*37*) (fig. S27 and table S5) and compared with independent ILMN-based HOR copy-number estimates (fig. S28). All centromeric satellite arrays, including beta satellite (BSat) and HSat repeats, were further validated by measuring the ratio of primary to secondary variants identified by HiFi reads (*38*) (fig. S29).

The consensus accuracy of the T2T-CHM13 assembly is estimated to be about one error per 10 Mbp (*23*, *34*), which exceeds the historical standard of "finished" sequence by orders of magnitude. However, regions of low HiFi coverage were found to be associated with an enrichment of potential errors, as estimated from both HiFi and ILMN data (*34*). To guide future use of the assembly, we have cataloged all low-coverage, low-confidence, and known heterozygous sites identified by the above validation procedures (*34*). The total number of bases covered by potential issues in the T2T-CHM13 assembly is just 0.3% of the total assembly length compared with 8% for GRCh38 (Fig. 3A).

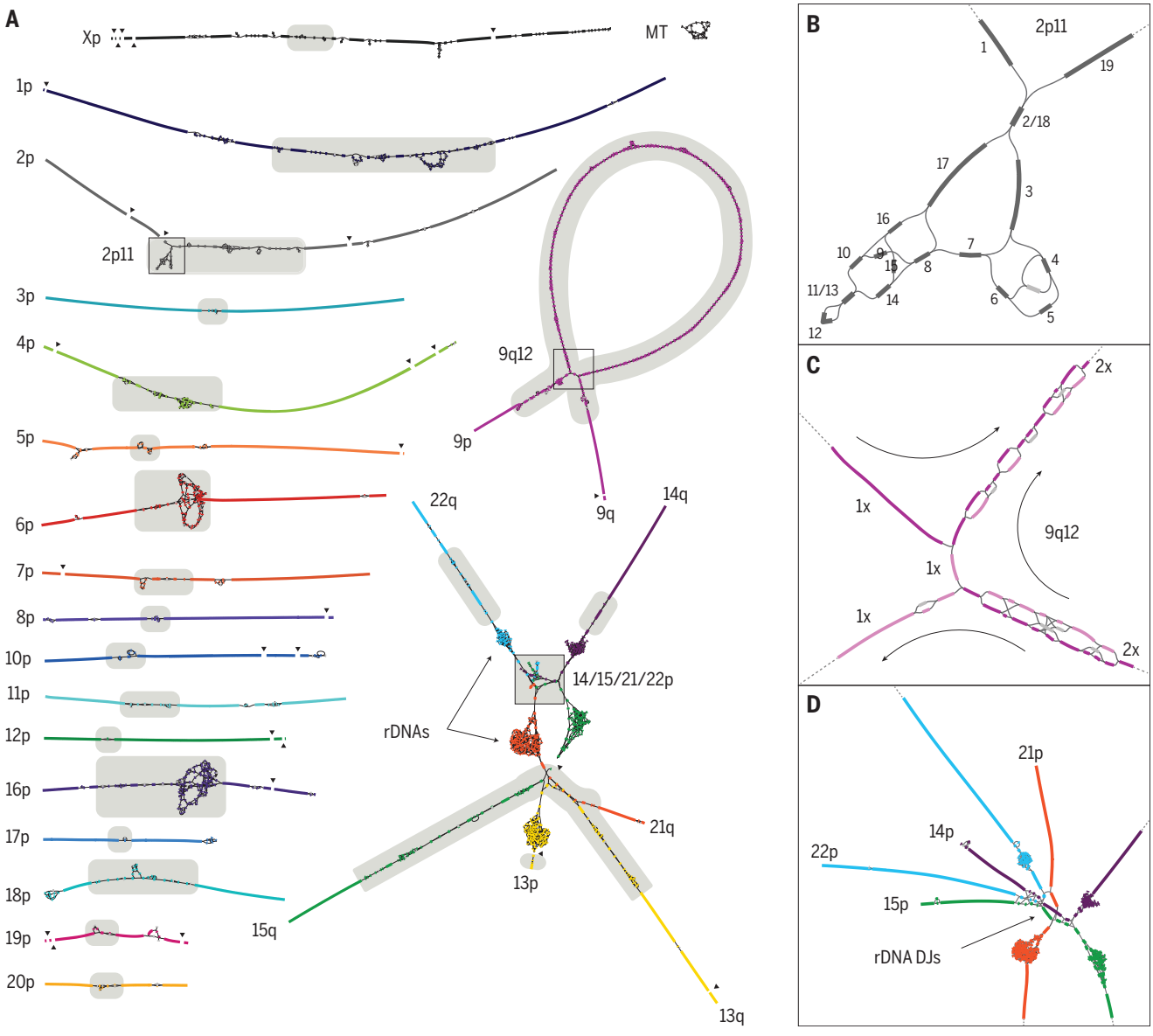


Fig. 2. High-resolution assembly string graph of the CHM13 genome. (**A**) Bandage (*60*) visualization, where nodes represent unambiguously assembled sequences scaled by length and edges correspond to the overlaps between node sequences. Each chromosome is both colored and numbered on the short (p) arm. Long (q) arms are labeled where unclear. The five acrocentric chromosomes (bottom right) are connected owing to similarity between their short arms, and the rDNA arrays form five dense tangles because of their high copy number. The graph is partially fragmented because of HiFi coverage dropout surrounding GA-rich sequence (black triangles). Centromeric satellites (*30*) are the source of most ambiguity in the graph (gray highlights). MT, mitochondria. (**B**) The ONT-assisted graph traversal for the 2p11 locus is given by numerical order. Based on low depth of coverage, the unlabeled light gray node represents an artifact or heterozygous variant and was not used. (**C**) The multimegabase tandem HSat3 duplication (9qh+) at 9q12 requires two traversals of the large loop structure. (The size of the loop is exaggerated because graph edges are of constant size.) Nodes used by the first traversal are in dark purple, and nodes used by the second traversal are in light purple. Nodes used by both traversals typically have twice the sequencing coverage. (**D**) Enlargement of the distal short arms of the acrocentrics, showing the colored graph walks and edges between highly similar sequences in the distal junctions (DJs) adjacent to the rDNA arrays.

A truly complete genome

T2T-CHM13 includes gapless telomere-to-telomere assemblies for all 22 human autosomes and chromosome X, comprising 3,054,815,472 bp of nuclear DNA, plus a 16,569-bp mitochondrial genome. This complete assembly adds or corrects 238 Mbp of sequence that does not colinearly align to GRCh38 over a 1-Mbp interval (i.e., is nonsyntenic), primarily comprising centromeric satellites (76%), nonsatellite segmental duplications (19%), and rDNAs (4%) (Fig. 1C). Of this, 182 Mbp of sequence has no primary alignments to GRCh38 and is exclusive to T2T-CHM13. As a result, T2T-CHM13 increases the number of known genes and repeats in the human genome (Table 1).

To provide an initial annotation, we used both the Comparative Annotation Toolkit (CAT) (*39*) and Liftoff (*40*) to project the GENCODE v35 (*41*) reference annotation onto the T2T-CHM13 assembly. Additionally, CHM13 full-length isoform sequencing (Iso-seq) transcriptome reads were assembled into transcripts and provided as complementary input to CAT. A comprehensive annotation was built by combining the CAT annotation with genes identified only by Liftoff (*23*).

The draft T2T-CHM13 annotation totals 63,494 genes and 233,615 transcripts, of which 19,969 genes (86,245 transcripts) are predicted to be protein coding, with 683 predicted frameshifts in 385 genes (469 transcripts) (Table 1, fig. S30, and tables S6 to S8). Only 263 GENCODE genes (448 transcripts) are exclusive to GRCh38 and have no assigned ortholog in the CHM13 annotation (tables S9 and S10). Of these, 194 are due to a lower copy number in the CHM13 annotation (fig. S31), 46 do not align well to CHM13, and 23 correspond to known false duplications in GRCh38 (*25*) (fig. S32). Most of these genes are noncoding and associated with repetitive elements. Only four are annotated as being medically relevant (*CFHR1*, *CFHR3*, *OR51A2*, *UGT2B28*), all of which are absent owing to a copy number difference, and the only protein-coding genes that align poorly are immunoglobulin and T cell receptor genes, which are known to be highly diverse.

In comparison, a total of 3604 genes (6693 transcripts) are exclusive to CHM13 (tables S11 and S12). Most of these genes represent putative paralogs and localize to pericentromeric regions and the short arms of the acrocentrics, including 876 rRNA transcripts. Only 48 of the CHM13-exclusive genes (56 transcripts) were predicted solely from de novo assembled transcripts. Of all genes exclusive to CHM13, 140 are predicted to be protein coding based on their GENCODE paralogs and have a mean of 99.5% nucleotide and 98.7% amino acid identity to their most similar GRCh38 copy (table S13). Although some of these additional paralogs may be present (but unannotated) in GRCh38 (*23*), 1956 of the genes exclusive to CHM13 (99 protein coding) are in regions with no primary alignment to GRCh38 (table S11). A broader set of 182 multi-exon protein-coding genes fall within nonsyntenic regions, 36% of which were confirmed to be expressed in CHM13 (*42*).

Compared with GRCh38, T2T-CHM13 is a more complete, accurate, and representative reference for both short- and long-read variant calling across human samples of all ancestries

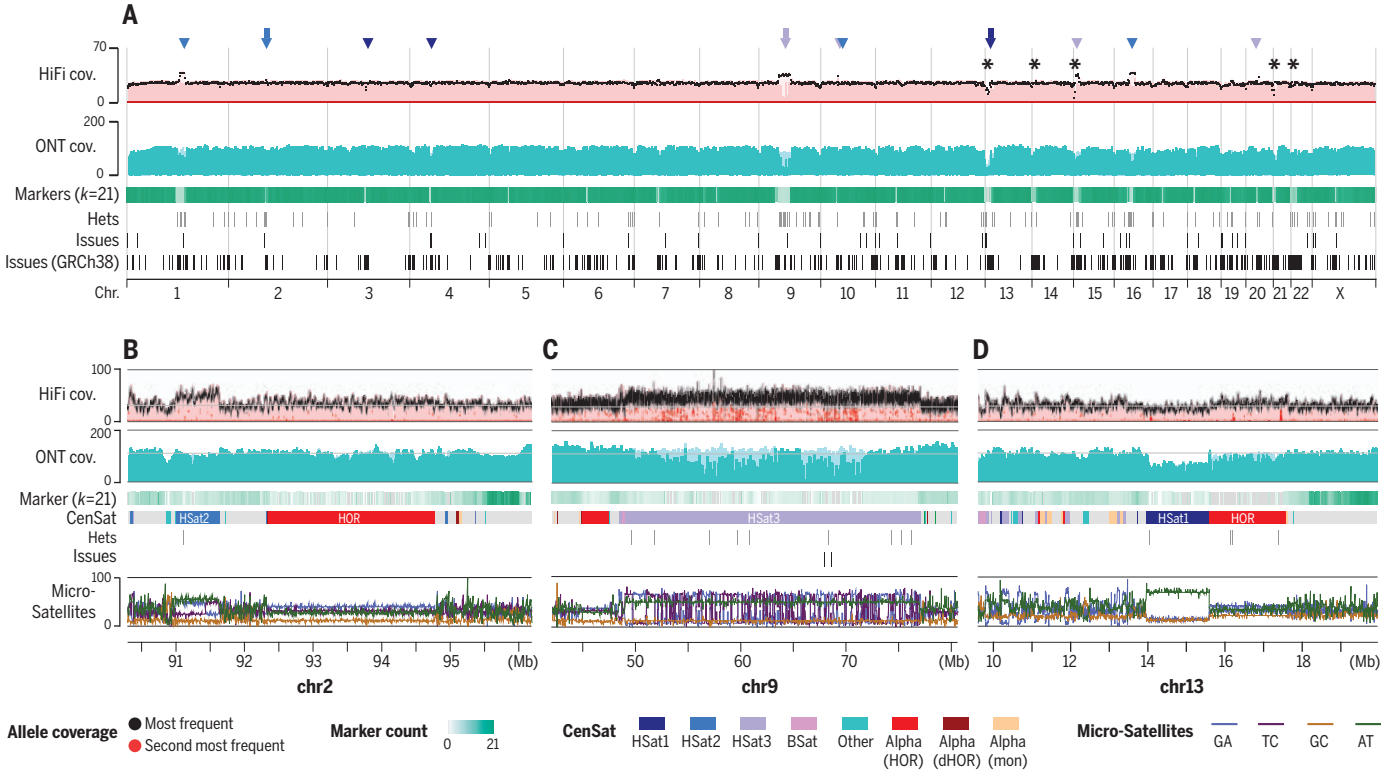


Fig. 3. Sequencing coverage and assembly validation. (**A**) Uniform whole-genome coverage of mapped HiFi and ONT reads is shown with primary alignments in light shades and marker-assisted alignments overlaid in dark shades. Large HSat arrays (*30*) are noted by triangles, with inset regions marked by arrowheads and the location of the rDNA arrays marked with asterisks. Regions with low unique marker frequency (light green) correspond to drops in unique marker density but are recovered by the lower-confidence primary alignments. Annotated assembly issues are compared for T2T-CHM13 and GRCh38. Hets, heterozygous variants; *k*, marker size. (**B** to **D**) Enlargements corresponding to regions of the genome featured in Fig. 2, B to D, respectively. Uniform coverage changes within certain satellites are reproducible and likely caused by sequencing bias. Identified heterozygous variants and assembly issues are marked below and typically correspond with low coverage of the primary allele (black) and increased coverage of the secondary allele (red). The percentage of microsatellite repeats for every 128-bp window is shown at the bottom. dHOR, divergent HOR; mon, monomeric.

(*25*). Reanalysis of 3202 short-read datasets from the 1KGP showed that T2T-CHM13 simultaneously reduces both false-negative and false-positive variant calls because of the addition of 182 Mbp of missing sequence and the exclusion of 1.2 Mbp of falsely duplicated sequence in GRCh38. These improvements, combined with a lower frequency of rare variants and errors in T2T-CHM13, eliminate tens of thousands of spurious variants per 1KGP sample (*25*). In addition, the T2T-CHM13 reference was found to be more representative of human copy-number variation than GRCh38 when compared against 268 human genomes from the Simons Genome Diversity Project (SGDP) (*42*, *43*). Specifically, within non-syntenic segmentally duplicated regions of the genome, T2T-CHM13 is nine times more predictive of SGDP copy number than GRCh38 (*42*). These results underscore both the quality of the assembly and the genomic stability of the cell line from which it was derived.

Acrocentric chromosomes

T2T-CHM13 uncovers the genomic structure of the short arms of the five acrocentric chromosomes, which, despite their importance for cellular function (*44*), have remained largely unsequenced to date. This omission has been due to their enrichment for satellite repeats and segmental duplications, which has prohibited sequence assembly and limited their characterization to cytogenetics, restriction mapping, and BAC sequencing (*45*–*47*). All five of CHM13's short arms follow a similar structure consisting of an rDNA array embedded within distal and proximal repeat arrays (Fig. 4). From telomere to centromere, the short arms vary in size from 10.1 Mbp (chromosome 14) to 16.7 Mbp (chromosome 15), with a combined length of 66.1 Mbp.

Compared with other human chromosomes, the short arms of the acrocentrics are unusually similar to one another. Specifically, we find that 5-kbp windows align with a median identity of 98.7% between the short arms, creating many opportunities for interchromosomal exchange (Fig. 4). This high degree of similarity is presumably due to recent non-allelic or ectopic recombination stemming from their colocalization in the nucleolus

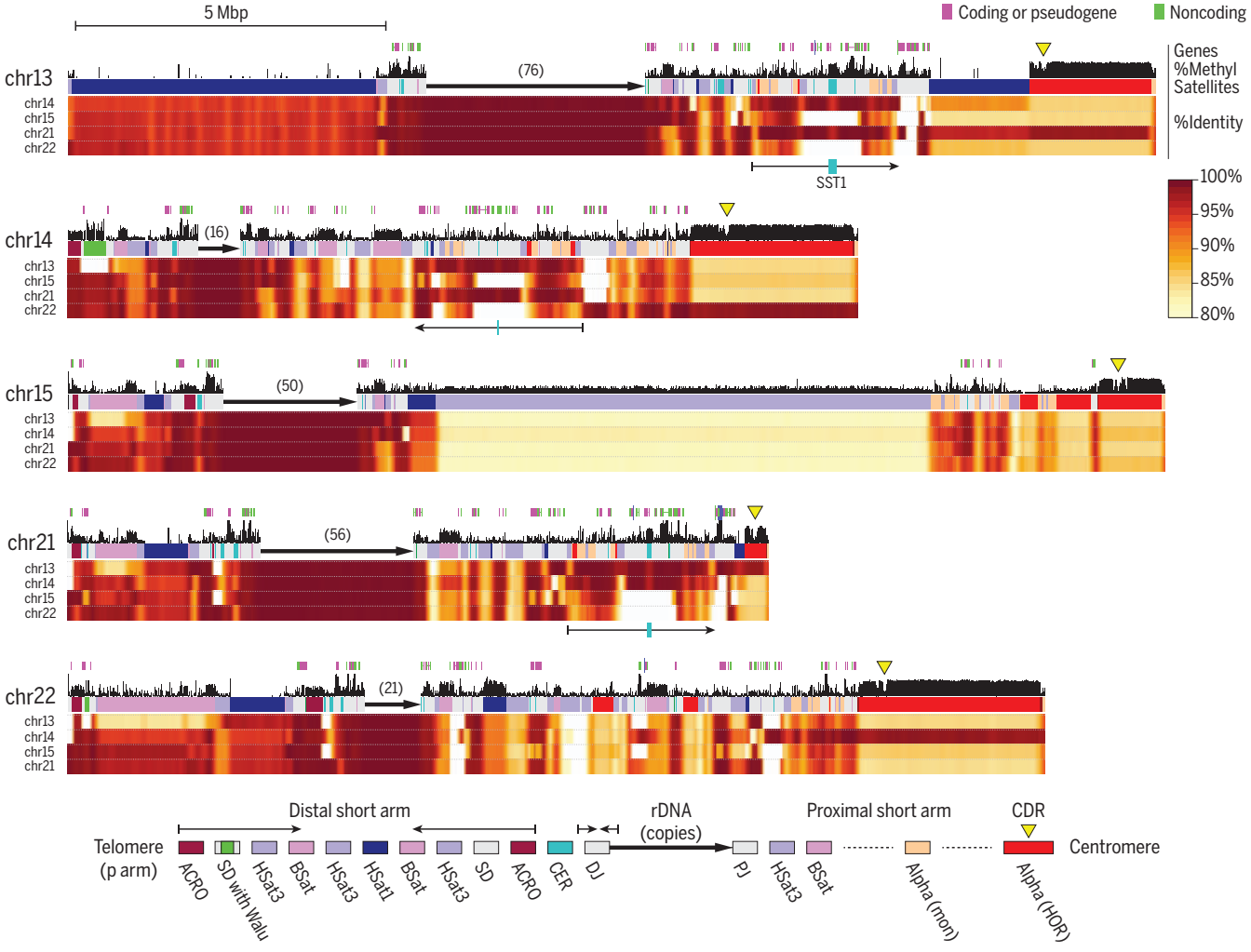


Fig. 4. Short arms of the acrocentric chromosomes. Each short arm is shown along with annotated genes, percentage of methylated CpGs (*29*), and a color-coded satellite repeat annotation (*30*). The rDNA arrays are represented by a directional arrow and copy number because of their high self-similarity, which prohibits ONT mapping. Percent identity heatmaps versus the other four arms were computed in 10-kbp windows and smoothed over 100-kbp intervals. Each position shows the maximum identity of that window to any window in the other chromosome. The distal short arms include conserved satellite structure and inverted repeats (thin arrows), whereas the proximal short arms show a diversity of structures. The proximal short arms of chromosomes 13, 14, and 21 share a segmentally duplicated core, including small alpha satellite HOR arrays and a central, highly methylated SST1 array (thin arrows with teal block). Yellow triangles indicate hypomethylated centromeric dip regions (CDRs), marking the sites of kinetochore assembly (*29*). Numbers in parentheses indicate rDNA copy number. ACRO, acrocentric repeat; CER, centromeric repeat; DJ, distal junction; PJ, proximal junction; SD, segmental duplication.

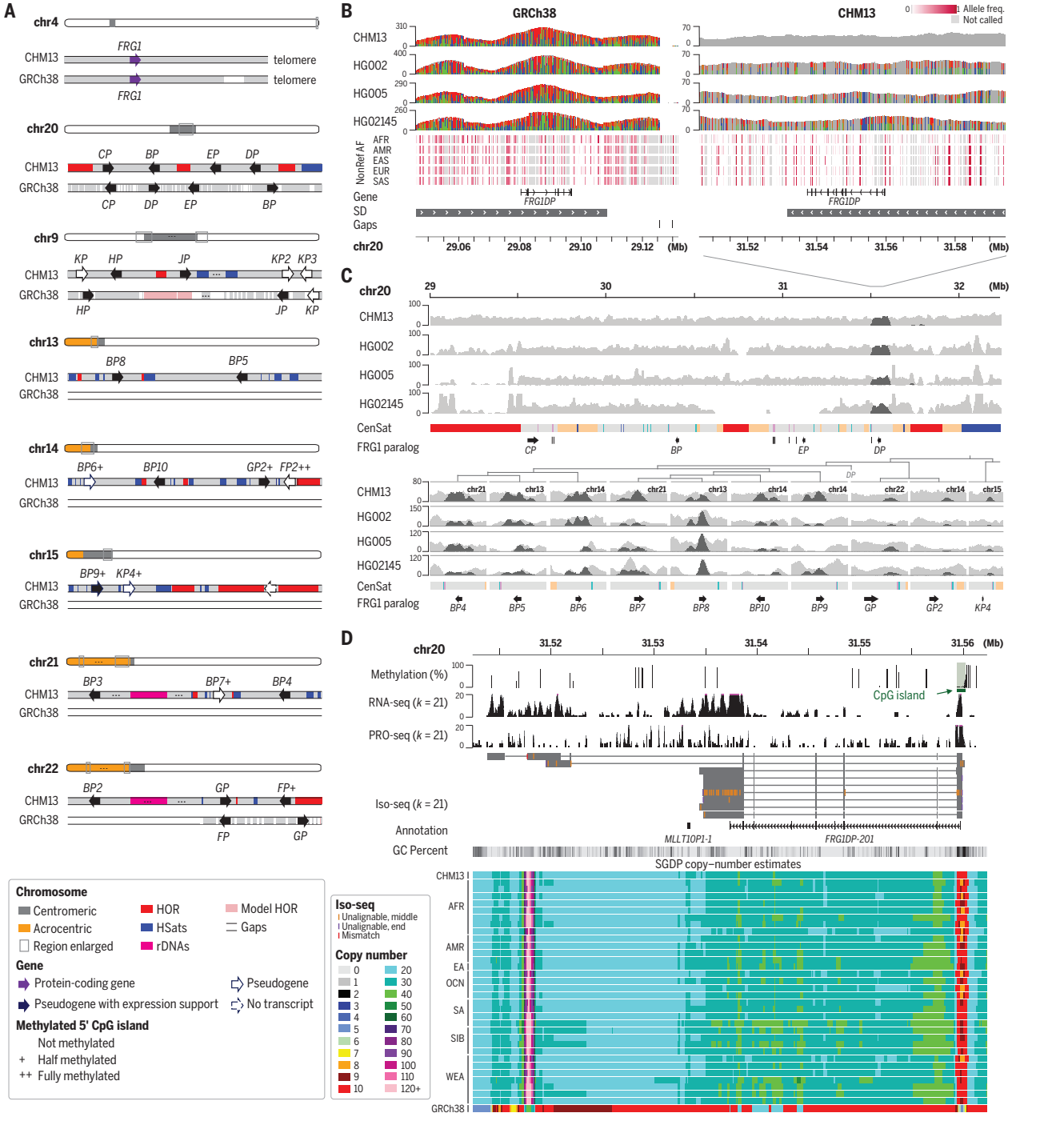


Fig. 5. Resolved FRG1 paralogs. (**A**) Protein-coding gene *FRG1* and its 23 paralogs in CHM13. Only nine are found in GRCh38. Genes are drawn larger than their actual size, and the "*FRG1*" prefix is omitted for brevity. All paralogs are found near satellite arrays. Most copies exhibit evidence of expression, including CpG islands present at the 5′ start site with varying degrees of methylation. (**B**) Reference (gray) and variant (colored) allele coverage is shown for four human HiFi samples mapped to the paralog *FRG1DP*. When mapped to GRCh38, the region shows excessive HiFi coverage and variants, indicating that reads from the missing paralogs are mismapped to *FRG1DP* (variants >80% frequency shown). When mapped to CHM13, HiFi reads show the expected coverage and a typical heterozygous variation pattern for the three non-CHM13 samples (variants >20% frequency shown). These nonreference alleles are also found in other populations from 1KGP ILMN data. NonRef AF, nonreference allele frequency; AFR, African; AMR, ad-mixed American; EAS, East Asian; EUR, European; SAS, South Asian. (**C**) Mapped HiFi read coverage for other *FRG1* paralogs, with an extended context shown for chromosome 20. Coverage of HiFi reads that mapped to *FRG1DP* in GRCh38 is highlighted (dark gray), showing the paralogous copies they originate from (*FRG1BP4* to *FRG1BP10*, *FRG1GP*, *FRG1GP2*, and *FRG1KP4*). Background coverage is variable for some paralogs, suggesting the presence of copy-number polymorphism in the population. (**D**) Methylation and expression profiles suggest transcription of *FRG1DP* in CHM13. In the copy-number display (bottom), 100-bp windows from the CHM13 assembly are highlighted with a color representing the estimated copy number of that sequence in an SGDP sample. The CHM13 and GRCh38 tracks show the estimated copy number of these same sequences in the respective assemblies. CHM13 copy number resembles all samples from the SGDP, whereas GRCh38 underrepresents the true copy number.

(*46*). Additionally, considering an 80% identity threshold, no 5-kbp window on the short arms is unique, and 96% of the non-rDNA sequence can be found elsewhere in the genome, suggesting that the acrocentrics are dynamic sources of segmental duplication.

The rDNA arrays of CHM13 vary in size from 0.7 Mbp (chromosome 14) to 3.6 Mbp (chromosome 13) and are in the expected arrangement, organized as head-to-tail tandem arrays with all 45*S* transcriptional units pointing toward the centromere. No inversions were noted within the arrays, and nearly all rDNA units are full length, in contrast to some prior studies that reported embedded inversions and other noncanonical structures (*47*, *48*). Each array appears highly homogenized, and there is more variation between rDNA units on different chromosomes than within chromosomes (fig. S33), suggesting that intrachromosomal exchange of rDNA units through nonallelic homologous recombination is more common than interchromosomal exchange.

Many 45*S* gene copies on the same chromosome are identical to one another, whereas the identity of the most frequent 45*S* morphs between chromosomes ranges from 99.4 to 99.7%. A chromosome 15 rDNA morph shows the highest identity (98.9%) to the current KY962518.1 rDNA reference sequence, originally derived from a human chromosome 21 BAC clone (*47*). As expected, the 13-kbp 45*S* is more conserved than the intergenic spacer, with all major 45*S* morphs aligning between 99.4 and 99.6% identity to KY962518.1. Certain rDNA variants appear to be chromosome specific, including single-nucleotide variants within the 45*S* and its upstream promoter region (fig. S34). The most evident variants are repeat expansions and contractions within the tandem "R" repeat that immediately follows the 45*S* and the CT-rich "long" repeat located in the middle of the intergenic spacer. The most frequent morph in each array can be specifically distinguished by these two features (fig. S35).

From the telomere to the rDNA array, the structure of all five distal short arms follows a similar pattern that involves a symmetric arrangement of inverted segmental duplications and acrocentric, HSat3, BSat, and HSat1 repeats (Fig. 4); however, the sizes of these repeat arrays vary among chromosomes. Chromosome 13 is missing the distal half of the inverted duplication and has an expanded HSat1 array relative to the others. Despite their variability in size, all satellite arrays share a high degree of similarity (typically >90% identity) both within and between acrocentric chromosomes. Chromosomes 14 and 22 also feature the expansion of a 64-bp Alu-associated satellite repeat ("Walu") within the distal inverted duplication (*49*), the location of which was confirmed by fluorescence in situ hybridization (FISH) (fig. S36). The distal junction immediately before the rDNA array includes centromeric repeats and a highly conserved and actively transcribed 200-kbp palindromic repeat, which agrees with previous characterizations of the rDNA flanking sequences (*46*, *50*).

Extending from the rDNA array to the centromere, the proximal short arms are larger in size and show a higher diversity of structures, including shuffled segmental duplications (*42*), composite transposable element arrays (*49*), satellite arrays (including HSat3, BSat, HSat1, and HSat5), and alpha satellite arrays (both monomeric and HORs) (*30*). Some proximal BSat arrays show a mosaic inversion structure that was also observed in HSat arrays elsewhere in the genome (*30*) (fig. S37). The proximal short arms of chromosomes 13, 14, and 21 appear to share the highest degree of similarity with a large region of segmental duplication, including similar HOR subsets and a central and highly methylated SST1 array (Fig. 4). This coincides with these three chromosomes being most frequently involved in Robertsonian translocations (*51*). Alpha satellite HORs on chromosomes 13 and 21 and chromosomes 14 and 22 also share high similarity within each pair, but not between them (*52*, *53*). Nonsatellite sequences within these segmental duplications often exceed 99% identity and show evidence of transcription (*29*, *42*, *49*). Using the T2T-CHM13 reference as a basis, further study of additional genomes is now needed to understand which of these features are conserved across the human population.

Analyses and resources

A number of companion studies were carried out to characterize the complete sequence of a human genome, including comprehensive analyses of centromeric satellites (*30*), segmental duplications (*42*), transcriptional (*49*) and epigenetic profiles (*29*), mobile elements (*49*), and variant calls (*25*). Up to 99% of the complete CHM13 genome can be confidently mapped with long-read sequencing, opening these regions of the genome to functional and variational analysis (*23*) (fig. S38 and table S14). We have produced a rich collection of annotations and omics datasets for CHM13—including RNA sequencing (RNA-seq) (*30*), Iso-seq (*21*), precision run-on sequencing (PRO-seq) (*49*), cleavage under targets and release using nuclease (CUT&RUN) (*30*), and ONT methylation (*29*) experiments—and have made these datasets available via a centralized University of California, Santa Cruz (UCSC), Assembly Hub genome browser (*54*).

To highlight the utility of these genetic and epigenetic resources mapped to a complete human genome, we provide the example of a segmentally duplicated region of the chromosome 4q subtelomere that is associated with facioscapulohumeral muscular dystrophy (FSHD) (*55*). This region includes FSHD region gene 1 (*FRG1*), FSHD region gene 2 (*FRG2*), and an intervening D4Z4 macrosatellite repeat containing the double homeobox 4 (*DUX4*) gene that has been implicated in the etiology of FSHD (*56*). Numerous duplications of this region throughout the genome have complicated past genetic analyses of FSHD.

The T2T-CHM13 assembly reveals 23 paralogs of *FRG1* spread across all acrocentric chromosomes as well as chromosomes 9 and 20 (Fig. 5A). This gene appears to have undergone recent amplification in the great apes (*57*), and approximate locations of *FRG1* paralogs were previously identified by FISH (*58*). However, only nine *FRG1* paralogs are found in GRCh38, hampering sequence-based analysis.

One of the few *FRG1* paralogs included in GRCh38, *FRG1DP*, is located in the centromeric region of chromosome 20 and shares high identity (97%) with several paralogs (*FRG1BP4* to *FRG1BP10*) (*23*) (fig. S39 and tables S15 and S16). When mapping HiFi reads, the absence of the additional *FRG1* paralogs in GRCh38 causes their reads to incorrectly align to *FRG1DP*, resulting in many false-positive variants (Fig. 5B). Most *FRG1* paralogs appear present in other human genomes (Fig. 5C), and all except *FRG1KP2* and *FRG1KP3* have upstream CpG islands and some degree of expression evidence in CHM13 (Fig. 5D and table S17). Any variants within these paralogs, and others like them, will be overlooked when using GRCh38 as a reference.

Future of the human reference genome

The T2T-CHM13 assembly adds five full chromosome arms and more additional sequence than any genome reference release in the past 20 years (Fig. 1D). This 8% of the genome has not been overlooked because of a lack of importance but rather because of technological limitations. High-accuracy long-read sequencing has finally removed this technological barrier, enabling comprehensive studies of genomic variation across the entire human genome, which we expect to drive future discovery in human genomic health and disease. Such studies will necessarily require a complete and accurate human reference genome.

CHM13 lacks a Y chromosome, and homozygous Y-bearing CHMs are nonviable, so a different sample type will be required to complete this last remaining chromosome. However, given its haploid nature, it should be possible to assemble the Y chromosome from a male sample using the same methods described here and supplement the T2T-CHM13 reference assembly with a Y chromosome as needed.

Extending beyond the human reference genome, large-scale resequencing projects have

revealed genomic variation across human populations. Our reanalyses of the 1KGP (*25*) and SGDP (*42*) datasets have already shown the advantages of T2T-CHM13, even for short-read analyses. However, these studies give only a glimpse of the extensive structural variation that lies within the most repetitive regions of the genome assembled here. Long-read resequencing studies are now needed to comprehensively survey polymorphic variation and reveal any phenotypic associations within these regions.

Although CHM13 represents a complete human haplotype, it does not capture the full diversity of human genetic variation. To address this bias, the Human Pangenome Reference Consortium (*59*) has joined with the T2T Consortium to build a collection of high-quality reference haplotypes from a diverse set of samples. Ideally, all genomes could be assembled at the quality achieved here, but automated T2T assembly of diploid genomes presents a difficult challenge that will require continued development. Until this goal is realized, and any human genome can be completely sequenced without error, the T2T-CHM13 assembly represents a more complete, representative, and accurate reference than GRCh38.

REFERENCES AND NOTES

1. V. A. Schneider *et al*., *Genome Res.* **27**, 849–864 (2017).
2. International Human Genome Sequencing Consortium, *Nature* **409**, 860–921 (2001).
3. J. C. Venter *et al*., *Science* **291**, 1304–1351 (2001).
4. E. W. Myers *et al*., *Science* **287**, 2196–2204 (2000).
5. E. E. Eichler, R. A. Clark, X. She, *Nat. Rev. Genet.* **5**, 345–354 (2004).
6. K. H. Miga *et al*., *Genome Res.* **24**, 697–707 (2014).
7. M. Gupta, A. R. Dhanasekaran, K. J. Gardiner, *Mamm. Genome* **27**, 538–555 (2016).
8. M. J. P. Chaisson *et al*., *Nature* **517**, 608–611 (2015).
9. International Human Genome Sequencing Consortium, *Nature* **431**, 931–945 (2004).
10. J. Eid *et al*., *Science* **323**, 133–138 (2009).
11. K. Berlin *et al*., *Nat. Biotechnol.* **33**, 623–630 (2015).
12. M. Jain *et al*., *Nat. Biotechnol.* **36**, 338–345 (2018).
13. M. Jain *et al*., *Nat. Biotechnol.* **36**, 321–323 (2018).
14. K. H. Miga *et al*., *Nature* **585**, 79–84 (2020).
15. A. M. Wenger *et al*., *Nat. Biotechnol.* **37**, 1155–1162 (2019).
16. S. Nurk *et al*., *Genome Res.* **30**, 1291–1305 (2020).
17. J. Huddleston *et al*., *Genome Res.* **27**, 677–685 (2017).
18. E. E. Eichler, U. Surti, R. Ophoff, Proposal for construction a human haploid BAC library from hydatidiform mole source material (2002); www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf.
19. K. M. Steinberg *et al*., *Genome Res.* **24**, 2066–2076 (2014).
20. M. R. Vollger *et al*., *Ann. Hum. Genet.* **84**, 125–140 (2020).
21. G. A. Logsdon *et al*., *Nature* **593**, 101–107 (2021).
22. J.-B. Fan *et al*., *Genomics* **79**, 58–62 (2002).
23. See supplementary materials.
24. 1000 Genomes Project Consortium, *Nature* **526**, 68–74 (2015).
25. S. Aganezov *et al*., *Science* **376**, eabl3533 (2022).
26. E. W. Myers, *Bioinformatics* **21**, ii79–ii85 (2005).
27. H. Li, *Bioinformatics* **32**, 2103–2110 (2016).
28. M. Rautiainen, T. Marschall, *Genome Biol.* **21**, 253 (2020).
29. A. Gershman *et al*., *Science* **376**, eabj5089 (2022).
30. N. Altemose *et al*., *Science* **376**, eabl4178 (2022).
31. M. M. Parks *et al*., *Sci. Adv.* **4**, eaao0665 (2018).
32. J. O. Nelson, G. J. Watase, N. Warsinger-Pepe, Y. M. Yamashita, *Trends Genet.* **35**, 734–742 (2019).
33. M. Rautiainen, T. Marschall, *Bioinformatics* **37**, 2476–2478 (2021).
34. A. M. McCartney *et al*., *Nat. Methods* https://doi.org/10.1038/s41592-022-01440-3 (2022).
35. J. M. Flynn, M. Long, R. A. Wing, A. G. Clark, *Mol. Biol. Evol.* **37**, 1362–1375 (2020).
36. W. M. Guiblet *et al*., *Genome Res.* **28**, 1767–1778 (2018).
37. A. Mikheenko, A. V. Bzikadze, A. Gurevich, K. H. Miga, P. A. Pevzner, *Bioinformatics* **36**, i75–i83 (2020).
38. M. R. Vollger *et al*., *Nat. Methods* **16**, 88–94 (2019).
39. I. T. Fiddes *et al*., *Genome Res.* **28**, 1029–1038 (2018).
40. A. Shumate, S. L. Salzberg, *Bioinformatics* **37**, 1639–1643 (2021).
41. A. Frankish *et al*., *Nucleic Acids Res.* **47**, D766–D773 (2019).
42. M. R. Vollger *et al*., *Science* **376**, eabj6965 (2022).
43. S. Mallick *et al*., *Nature* **538**, 201–206 (2016).
44. M. S. Lindström *et al*., *Oncogene* **37**, 2351–2366 (2018).
45. R. Lyle *et al*., *Genome Res.* **17**, 1690–1696 (2007).
46. I. Floutsakou *et al*., *Genome Res.* **23**, 2003–2012 (2013).
47. J.-H. Kim *et al*., *Nucleic Acids Res.* **46**, 6712–6725 (2018).
48. S. Caburet *et al*., *Genome Res.* **15**, 1079–1085 (2005).
49. S. J. Hoyt *et al*., *Science* **376**, eabk3112 (2022).
50. M. van Sluis *et al*., *Genes Dev.* **33**, 1688–1701 (2019).
51. B. A. Sullivan, L. S. Jenkins, E. M. Karson, J. Leana-Cox, S. Schwartz, *Am. J. Hum. Genet.* **59**, 167–175 (1996).
52. G. M. Greig, P. E. Warburton, H. F. Willard, *J. Mol. Evol.* **37**, 464–475 (1993).
53. A. L. Jørgensen, S. Kølvraa, C. Jones, A. L. Bak, *Genomics* **3**, 100–109 (1988).
54. W. J. Kent *et al*., *Genome Res.* **12**, 996–1006 (2002).
55. C. Wijmenga *et al*., *Nat. Genet.* **2**, 26–30 (1992).
56. R. J. L. F. Lemmers *et al*., *Science* **329**, 1650–1653 (2010).
57. P. K. Grewal, M. van Geel, R. R. Frants, P. de Jong, J. E. Hewitt, *Gene* **227**, 79–88 (1999).
58. J. C. van Deutekom *et al*., *Hum. Mol. Genet.* **5**, 581–590 (1996).
59. K. H. Miga, T. Wang, *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).
60. R. R. Wick, M. B. Schultz, J. Zobel, K. E. Holt, *Bioinformatics* **31**, 3350–3352 (2015).
61. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, Zenodo (2021); https://doi.org/10.5281/zenodo.5598253.

ACKNOWLEDGMENTS

We thank M. Akeson, A. Carroll, P.-C. Chang, A. Delcher, M. Nattestad, and M. Pop for discussions on sequencing, assembly, and analysis; AnVIL, Amazon Web Services, DNAnexus, the UW Genome Sciences IT Group, and the UConn Computational Biology Core for computational support; and the National Institutes of Health (NIH) Intramural Sequencing Center, the UConn Center for Genome Innovation, and the Stowers Imaging Facility for experimental support. This work used the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). Certain commercial equipment, instruments, or materials are identified to adequately specify experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. **Funding:** Intramural Research Program of the National Human Genome Research Institute, NIH (A.M.P., A.C.Y., A.M.M., M.R., A.R., B.P.W., G.G.B., C.J., J.C.M., N.F.H., S.K., S.N., and S.Y.B.); NIH U01HG010971 (E.E.E., H.L., K.H.M., M.Kr., R.S.F., and T.A.G.-L.); NIH R01HG002385 and R01HG010169 (E.E.E.); NIH R01HG009190 (A.G. and W.T.); NIH R01HG010485 and U01HG010961 (B.P., E.G., and K.S.); NIH U41HG010972 (I.M.H., B.P., E.G., and K.S.); National Science Foundation (NSF) 1627442, 1732253, and 1758800 and NIH U24HG006620, U01CA253481, and R24DK106766 (M.C.S.); NIH U24HG010263 (S.Z. and M.C.S.); Mark Foundation for Cancer Research 19-033-ASP (S.A. and M.C.S.); NIH R01HG006677 (A.Sh. and S.L.S.); NIH U24HG009081 (R.S.F. and T.A.G.-L.); intramural funding at the National Institute of Standards and Technology (J.M., J.M.Z., J.W., and N.D.O.); St. Petersburg State University grant 73023573 (A.M., I.A.A., and T.D.); NIH R01HG002939 (A.F.A.S., I.T.F., and J.M.S.); NIH R01GM124041, R01GM129263, and R21CA238758 (B.A.S.); Intramural Research Program of the National Library of Medicine, NIH (C.X., F.T.-N., and V.A.S.); NIH F31HG011205 (C.J.S.); Damon Runyon Postdoctoral Fellowship and PEW Latin American Fellowship (G.V.C.); Fulbright Fellowship (D.C.S.); Howard Hughes Medical Institute (HHMI) (E.D.J. and G.F.); NIH R01AG054712 (E.I.R.); NIH UM1HG008898 (F.J.S.); NIH R01GM123312 and R21CA240199 and NSF 1613806 and 1643825 (G.A.H., P.G.S.G., S.J.H., and R.J.O.); NIH R21CA240199, NSF 643825, and Connecticut Innovations 20190200 (R.J.O.); NIH F32GM134558 (G.A.L.); NIH R01HG010040 (H.L.); St. Petersburg State University grant 73023573 (I.A.A.); Wellcome WT206194 (J.T., J.M.D.W., K.H., W.C., and Y.S.); Wellcome WT207492 (R.D.); Stowers Institute for Medical Research (J.L.G.); NIH R01HG011274 (K.H.M.); Ministry of Science and Higher Education of the RF 075-10-2020-116/13.1902.21.0023 (L.U.); supported by the Sirius University (L.U.); Russian Science Foundation (RSF) 19-75-30039 Analysis of genomic repeats (I.A.A.); NIH U41HG007234 (M.D.); NIH DP2MH119424 (M.Y.D.); HHMI Hanna H. Gray Fellowship (N.A.); NIH R35GM133747 (R.C.M.); Childcare Foundation, Swiss National Science Foundation, and European Research Council (ERC) 249968 (S.E.A.); German Federal Ministry for Research and Education 031L0184A (T.M.); Chan Zuckerberg Biohub Investigator Award (A.St.); Common Fund, Office of the Director, NIH (V.V.M.); and the Max Planck Society (E.W.M.). E.E.E. and E.D.J. are investigators of the HHMI. **Author contributions:** Analysis teams are listed, with leads indicated by an asterisk. Assembly: S.N.*, S.K.*, M.R.*, M.A., H.C., C.-S.C., R.D., E.G., M.Ki., M.Ko., H.L., T.M., E.W.M., I.S., B.P.W., A.W., A.M.P. Acrocentrics: A.M.P.*, J.L.G.*, M.R., S.E.A., M.B., R.D., L.G.d.L., T.P. Validation: A.R.*, A.V.B.*, A.M.*, M.A.*, A.M.M.*, K.S.*, W.C., L.G.d.L., T.D., G.F., A.F., K.H., C.J., E.D.J., D.P., V.A.S., Y.S., B.A.S., F.T.-N., J.T., J.M.D.W., A.M.P. Segmental duplications: M.R.V.*, E.E.E.*, S.N., S.K., M.D., P.C.D., A.G., G.A.L., D.P., C.J.S., D.C.S., M.Y.D., W.T., K.H.M., A.M.P. Satellite annotation: N.A.*, I.A.A.*, K.H.M.*, A.V.B., L.U., T.D., L.G.d.L., P.A.P., E.I.R., A.St., B.A.S., A.M.P.. Epigenetics: A.G.*, W.T.*, S.K., A.R., M.R.V., N.A., S.J.H., G.A.L., G.V.C., M.C.S., R.J.O., E.E.E., K.H.M., A.M.P. Variants: S.A.*, D.C.S.*, S.M.Y.*, S.Z.*, R.C.M.*, M.Y.D.*, J.M.Z.*, M.C.S.*, N.F.H., M.Ki., J.M., D.E.M., N.D.O., J.A.R., F.J.S., K.S., A.Sh., J.W., C.X., A.M.P. Repeat annotation: S.J.H.*, R.J.O.*, A.G., P.G.S.G., G.A.H., L.G.d.L., A.F.A.S., J.M.S. Gene annotation: M.D.*, M.H.*, A.Sh.*, S.N., S.K., P.C.D., I.T.F., S.L.S., F.T.-N., A.M.P. Browsers: M.D.*, N.-C.C., P.K. Data generation: S.J.H., G.G.B., S.Y.B., G.V.C., R.S.F., T.A.G.-L., I.M.H., M.W.H., M.J., J.K., M.Kr., V.V.M., J.C.M., B.P., P.P., A.C.Y., U.S., M.Y.D., J.L.G., R.J.O., W.T., E.E.E., K.H.M., A.M.P. Computational resources: C.-S.C., A.F., R.J.O., M.C.S., K.H.M., A.M.P. Manuscript draft: A.M.P. Figures: S.K., S.N., A.M.P., A.R. Editing: A.M.P., S.N., S.K., A.R., E.E.E., and K.H.M., with the assistance of all authors. Supplement: S.N. and S.K., with the assistance of the working groups. Supervision: R.C.M., M.Y.D., I.A.A., J.L.G., R.J.O., W.T., J.M.Z., M.C.S., E.E.E., K.H.M., A.M.P. Conceptualization: E.E.E., K.H.M., A.M.P. **Competing interests:** A.F. and C.-S.C. are employees of DNAnexus; I.S., J.K., M.W.H., P.P., and A.W. are employees of Pacific Biosciences; S.A. is an employee and stockholder of Oxford Nanopore Technologies; E.E.E. is a science advisory board member of Variant Bio; K.H.M. is a science advisory board member of Centaura; P.K. owns and receives income from Reservoir Genomics LLC; F.J.S. has received travel funds to speak at events hosted by Pacific Biosciences; S.K., D.E.M., F.J.S., and K.H.M. have received travel funds to speak at events hosted by Oxford Nanopore Technologies; and W.T. has licensed two patents to Oxford Nanopore Technologies (US 8748091 and 8394584). **Data and materials availability:** The T2T-CHM13 and T2T-HG002-ChrX assemblies generated by this study are archived under National Center for Biotechnology Information (NCBI) GenBank accession numbers GCA_009914755 and CP086568, respectively. CHM13hTERT cells were obtained for research use through a material transfer agreement with U. Surti and the University of Pittsburgh. The raw sequencing data were described in prior studies and are summarized in table S1. For convenience, links to the sequence data and genome browsers are also available from https://github.com/marbl/CHM13. Supplementary data for fig. S39 and the string graph construction code are archived at Zenodo (*61*) and also https://github.com/snurk/sg_sandbox.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj6987
Material and Methods
Figs. S1 to S39
Tables S1 to S17
References (*62*–*128*)
MDAR Reproducibility Checklist

26 May 2021; accepted 28 January 2022
10.1126/science.abj6987

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

A complete reference genome improves analysis of human genetic variation

Sergey Aganezov†, Stephanie M. Yan†, Daniela C. Soto†, Melanie Kirsche†, Samantha Zarate†, Pavel Avdeyev, Dylan J. Taylor, Kishwar Shafin, Alaina Shumate, Chunlin Xiao, Justin Wagner, Jennifer McDaniel, Nathan D. Olson, Michael E. G. Sauria, Mitchell R. Vollger, Arang Rhie, Melissa Meredith, Skylar Martin, Joyce Lee, Sergey Koren, Jeffrey A. Rosenfeld, Benedict Paten, Ryan Layer, Chen-Shan Chin, Fritz J. Sedlazeck, Nancy F. Hansen, Danny E. Miller, Adam M. Phillippy, Karen H. Miga, Rajiv C. McCoy*, Megan Y. Dennis*, Justin M. Zook*, Michael C. Schatz*

INTRODUCTION: One of the central applications of the human reference genome has been to serve as a baseline for comparison in nearly all human genomic studies. Unfortunately, many difficult regions of the reference genome have remained unresolved for decades and are affected by collapsed duplications, missing sequences, and other issues. Relative to the current human reference genome, GRCh38, the Telomere-to-Telomere CHM13 (T2T-CHM13) genome closes all remaining gaps, adds nearly 200 million base pairs (Mbp) of sequence, corrects thousands of structural errors, and unlocks the most complex regions of the human genome for scientific inquiry.

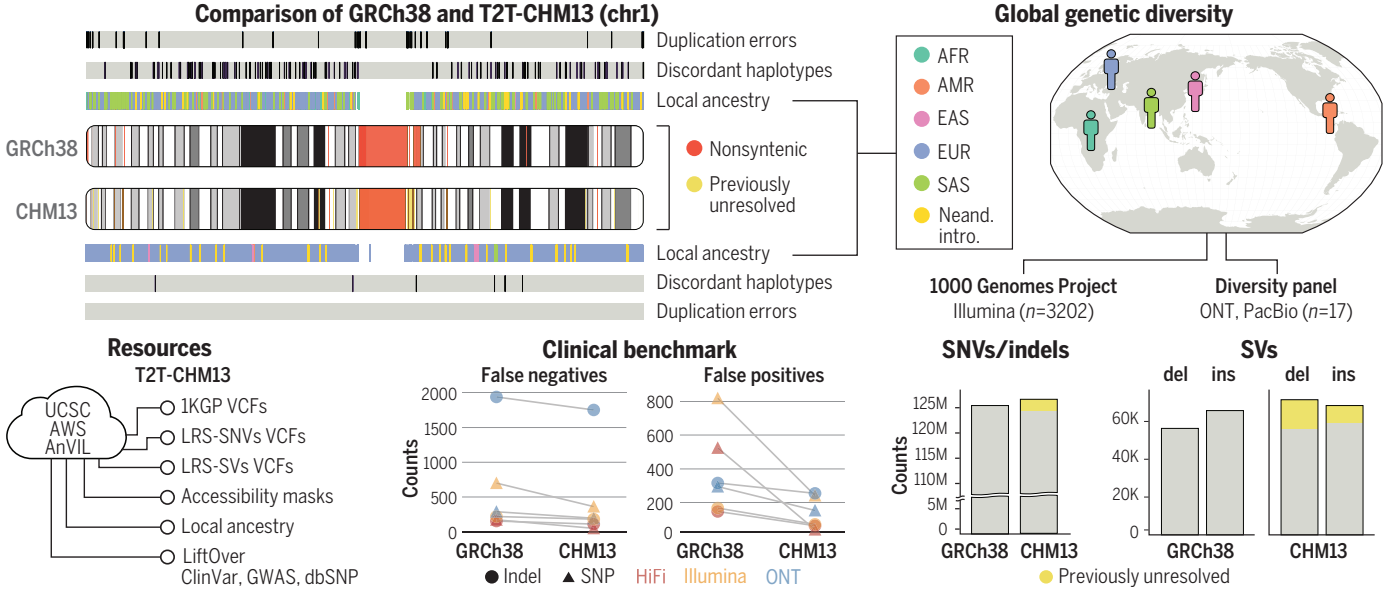
RATIONALE: We demonstrate how the T2T-CHM13 reference genome universally improves read mapping and variant identification in a globally diverse cohort. This cohort includes all 3202 samples from the expanded 1000 Genomes Project (1KGP), sequenced with short reads, as well as 17 globally diverse samples sequenced with long reads. By applying state-of-the-art methods for calling single-nucleotide variants (SNVs) and structural variants (SVs), we document the strengths and limitations of T2T-CHM13 relative to its predecessors and highlight its promise for revealing new biological insights within technically challenging regions of the genome.

RESULTS: Across the 1KGP samples, we found more than 1 million additional high-quality variants genome-wide using T2T-CHM13 than with GRCh38. Within previously unresolved regions of the genome, we identified hundreds of thousands of variants per sample—a promising opportunity for evolutionary and biomedical discovery. T2T-CHM13 improves the Mendelian concordance rate among trios and eliminates tens of thousands of spurious SNVs per sample, including a reduction of false positives in 269 challenging, medically relevant genes by up to a factor of 12. These corrections are in large part due to improvements to 70 protein-coding genes in >9 Mbp of inaccurate sequence caused by falsely collapsed or duplicated regions in GRCh38. Using the T2T-CHM13 genome also yields a more comprehensive view of SVs genome-wide, with a greatly improved balance of insertions and deletions. Finally, by providing numerous resources for T2T-CHM13 (including 1KGP genotypes, accessibility masks, and prominent annotation databases), our work will facilitate the transition to T2T-CHM13 from the current reference genome.

CONCLUSION: The vast improvements in variant discovery across samples of diverse ancestries position T2T-CHM13 to succeed as the next prevailing reference for human genetics. T2T-CHM13 thus offers a model for the construction and study of high-quality reference genomes from globally diverse individuals, such as is now being pursued through collaboration with the Human Pangenome Reference Consortium. As a foundation, our work underscores the benefits of an accurate and complete reference genome for revealing diversity across human populations. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: rajiv.mccoy@jhu.edu (R.C.M.); mydennis@ucdavis.edu (M.Y.D.); justin.zook@nist.gov (J.M.Z.); mschatz@cs.jhu.edu (M.C.S.)
†These authors contributed equally to this work.
Cite this article as S. Aganezov *et al.*, *Science* **376**, eabl3533 (2022). DOI: 10.1126/science.abl3533

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abl3533



Genomic features and resources available for T2T-CHM13. Comparisons to GRCh38 reveal broad improvements in SNVs, indels, and SVs discovered across diverse human populations by means of short-read (1KGP) and long-read sequencing (LRS). These improvements are due to resolution of complex genomic loci (nonsyntenic and previously unresolved), duplication errors, and discordant haplotypes, including those in medically relevant genes.

RESEARCH ARTICLE

HUMAN GENOMICS

A complete reference genome improves analysis of human genetic variation

Sergey Aganezov[1]†, Stephanie M. Yan[2]†, Daniela C. Soto[3]†, Melanie Kirsche[1]†, Samantha Zarate[1]†, Pavel Avdeyev[4], Dylan J. Taylor[2], Kishwar Shafin[5], Alaina Shumate[6], Chunlin Xiao[7], Justin Wagner[8], Jennifer McDaniel[8], Nathan D. Olson[8], Michael E. G. Sauria[2], Mitchell R. Vollger[9], Arang Rhie[4], Melissa Meredith[5], Skylar Martin[10], Joyce Lee[11], Sergey Koren[4], Jeffrey A. Rosenfeld[12], Benedict Paten[5], Ryan Layer[10], Chen-Shan Chin[13], Fritz J. Sedlazeck[14], Nancy F. Hansen[15], Danny E. Miller[9,16], Adam M. Phillippy[4], Karen H. Miga[5], Rajiv C. McCoy[2]*, Megan Y. Dennis[3]*, Justin M. Zook[8]*, Michael C. Schatz[1,2,17]*

Compared to its predecessors, the Telomere-to-Telomere CHM13 genome adds nearly 200 million base pairs of sequence, corrects thousands of structural errors, and unlocks the most complex regions of the human genome for clinical and functional study. We show how this reference universally improves read mapping and variant calling for 3202 and 17 globally diverse samples sequenced with short and long reads, respectively. We identify hundreds of thousands of variants per sample in previously unresolved regions, showcasing the promise of the T2T-CHM13 reference for evolutionary and biomedical discovery. Simultaneously, this reference eliminates tens of thousands of spurious variants per sample, including reduction of false positives in 269 medically relevant genes by up to a factor of 12. Because of these improvements in variant discovery coupled with population and functional genomic resources, T2T-CHM13 is positioned to replace GRCh38 as the prevailing reference for human genetics.

For the past 20 years, the human reference genome (GRCh38) has served as the bedrock of human genetics and genomics (*1*–*3*). One of the central applications of the human reference genome, and of reference genomes in general, has been to serve as a substrate for clinical, comparative, and population genomic analyses. More than 1 million human genomes have been sequenced to study genetic diversity and clinical relationships, and nearly all of them have been analyzed by aligning the sequencing reads from the donors to the reference genome [e.g., (*4*–*6*)]. Even when donor genomes are assembled de novo, independent of any reference, the assembled sequences are almost always compared to a reference genome to characterize variation by leveraging deep catalogs of available annotations (*7*, *8*). Consequently, human genetics and genomics benefit from the availability of a high-quality reference genome, ideally without gaps or errors that may obscure important variation and regulatory relationships.

The current human reference genome, GRCh38, is used for countless applications, with rich resources available to visualize and annotate the sequence across cell types and disease states (*3*, *9*–*12*). However, despite decades of effort to construct and refine its sequence, the human reference genome still suffers from several major limitations that hinder comprehensive analysis. Most immediately, GRCh38 contains more than 100 million nucleotides that either remain entirely unresolved (currently represented as "N" characters), such as the p-arms of the acrocentric chromosomes, or are substituted with artificial models, such as the centromeric satellite arrays (*13*). Furthermore, GRCh38 possesses 11.5 Mbp of unplaced and unlocalized sequences that are represented separately from the primary chromosomes (*3*, *14*). These sequences are difficult to study, and many genomic analyses exclude them to avoid identifying false variants and false regulatory relationships (*6*). Relatedly, artifacts such as an apparent imbalance between insertions and deletions (indels) have been attributed to systematic misassemblies in GRCh38 (*15*–*17*). Overall, these errors and omissions in GRCh38 introduce biases in genomic analyses, particularly in centromeres, satellites, and other complex regions.

Another major concern regards the influence of the reference genome on analysis of variation across large cohorts for population and clinical genomics. Several studies, such as the 1000 Genomes Project (1KGP) (*18*) and gnomAD (*6*), have provided information about the extent of genetic diversity within and between human populations. Many analyses of Mendelian and complex diseases use these catalogs of single-nucleotide variants (SNVs), small indels, and structural variants (SVs) to rank and prioritize potential causal variants on the basis of allele frequencies (AFs) and other evidence (*19*–*21*). When evaluating these resources, the overall quality and representativeness of the human reference genome should be considered. Any gaps or errors in the sequence could obscure variation and its contribution to human phenotypes and disease.

In addition to omissions such as centromeric sequences or acrocentric chromosome arms, the current reference genome possesses other errors and biases, including within genes of known medical relevance (*22*, *23*). Moreover, GRCh38 was assembled from multiple donors with clone-based sequencing, which creates an excess of artificial haplotype structures that can subtly bias analyses (*1*, *24*). Over the years, there have been attempts to replace certain rare alleles with more common alleles, but hundreds of thousands of artificial haplotypes and rare alleles remain to this day (*3*, *25*, *26*). Increasing the continuity, quality, and representativeness of the reference genome is therefore crucial for improving genetic diagnosis, as well as for understanding the complex relationship between genetic and phenotypic variation.

The Telomere-to-Telomere (T2T) CHM13 genome addresses many of the limitations of the current reference (*27*). Specifically, the T2T-CHM13v1.0 assembly adds nearly 200 Mbp of sequence and resolves errors present in GRCh38. Here, we demonstrate the impact of the T2T-CHM13 reference on variant discovery and genotyping in a globally diverse cohort. This includes all 3202 samples from the recently expanded 1KGP sequenced with short reads (*28*) along with 17 samples from diverse populations sequenced with long reads (*8*, *27*, *29*). Our analysis reveals more than 2 million variants within previously unresolved regions of the genome, genome-wide improvements in SV discovery, and enhancement in variant calling accuracy across 622 medically relevant genes. In summary, our work demonstrates universal improvements in read mapping and variant calling, thereby broadening the horizon for future genomic studies.

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Biochemistry and Molecular Medicine, Genome Center, MIND Institute, University of California, Davis, CA, USA. [4]Genome Informatics Section, National Human Genome Research Institute, Bethesda, MD, USA. [5]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [6]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [7]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA. [8]National Institute of Standards and Technology, Gaithersburg, MD, USA. [9]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [10]Department of Computer Science and Biofrontiers Institute, University of Colorado, Boulder, CO, USA. [11]Bionano Genomics, San Diego, CA, USA. [12]Cancer Institute of New Jersey, New Brunswick, NJ, USA. [13]DNAnexus, Mountain View, CA, USA. [14]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [15]Comparative Genomics Analysis Unit, National Human Genome Research Institute, Rockville, MD, USA. [16]Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA. [17]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.
***Corresponding author. Email: rajiv.mccoy@jhu.edu (R.C.M.); mydennis@ucdavis.edu (M.Y.D.); justin.zook@nist.gov (J.M.Z.); mschatz@cs.jhu.edu (M.C.S.)**
†These authors contributed equally to this work.

Structural comparisons of GRCh38 and T2T-CHM13

Introducing the T2T-CHM13 genome

The T2T-CHM13 reference genome was primarily assembled from Pacific Biosciences (PacBio) High Fidelity (HiFi) reads augmented with Oxford Nanopore Technology (ONT) reads to close gaps and resolve complex repeats (*27*). The resulting T2T-CHM13v1.0 assembly was subsequently validated and polished, with a consensus accuracy estimated to be between Phred Q67 and Q73 (*27*, *30*) and with only three minor known structural defects detected (*30*). The assembly is highly contiguous, with only five unresolved regions from the most highly repetitive ribosomal DNA (rDNA) arrays, representing only 9.9 Mbp of sequence out of >3.0 Gbp of fully resolved sequence. The version 1.0 assembly adds or revises 229 Mbp of sequence compared to GRCh38; these are "nonsyntenic" regions of the T2T-CHM13 assembly that do not linearly align to GRCh38 over a 1-Mbp interval. Furthermore, 189 Mbp of sequence are not covered by any primary alignments from GRCh38 and are resolved in the T2T-CHM13 assembly. Figure 1A is a summary diagram of the syntenic/nonsyntenic regions and their associated annotations for chromosomes 1 and 21 (figs. S1 to S4 give details for all chromosomes). Note that the subsequent T2T-CHM13v1.1 assembly (*27*) further resolves the rDNA regions using model sequences for some array elements, although for this study we analyze the v1.0 assembly, which does not contain these representations.

The bulk of the nonsyntenic sequence within T2T-CHM13 comprises centromeric satellites (190 Mbp) (*31*) and copies of segmental duplications (SDs; 218 Mbp) (*32*). These sequences could prove challenging for variant analysis, especially for variants identified using short-read sequencing. However, relative to GRCh38, we report an overall increase in unique sequence, defined as k-length strings (k-mers) found only once in the genome (e.g., 14.9 Mbp of added unique sequence when considering 50-mers, 23.5 Mbp for 100-mers, and 39.5 Mbp for 300-mers). These sequences delineate regions of confident mapping for short paired-end reads or longer reads, including those in previously unrepresented portions of the genome (Fig. 1B and figs. S5 and S6).

More than 106 Mbp of sequence absent from GRCh38 was identified in T2T-CHM13 within highly repetitive regions that require reads of

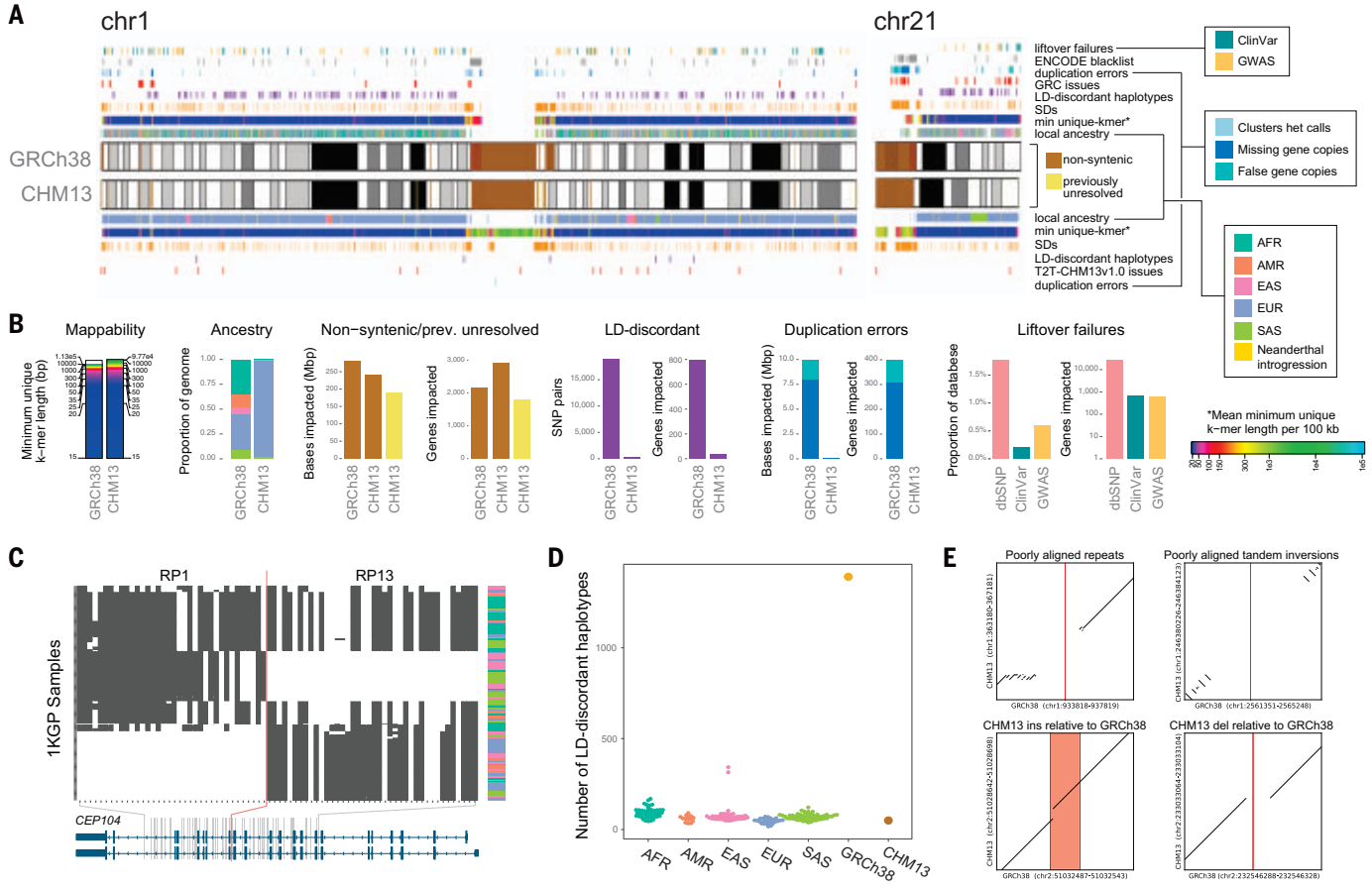


Fig. 1. Genomic comparisons of human assemblies GRCh38 and T2T-CHM13. (**A**) Overview of annotations available for GRCh38 and T2T-CHM13 chromosomes 1 and 21 with colors indicated in legends, which are also used in (B) to (D). Colors for mean minimum (min) unique k-mers are defined in the legend with indicated asterisk. Cytobands are pictured as gray bands with red bands representing centromeric regions within ideograms. Complete annotations of all chromosomes can be found in figs. S1 to S4. Local ancestry is denoted using 1KGP superpopulation abbreviations (AFR, African; AMR, admixed American; EAS, East Asian; EUR, European; SAS, South Asian). (**B**) Summary of the number of bases and/or genes annotated by different features for the assemblies with colors indicated in the legends shown in (A). Note, dbSNP liftover failures (pink) are not annotated in (A). (**C**) Example of a clone boundary (red line) where GRCh38 possesses a combination of alleles that segregate in negative LD within the 1KGP sample (which we term as an "LD-discordant haplotype"). SNPs are depicted in columns; phased 1KGP samples are depicted in rows. White indicates reference allele genotype; black indicates alternative allele genotypes. Superpopulation ancestry of each sample is indicated in the rightmost column with colors indicated in local ancestry legend shown in (A). *CEP104* splice isoforms (blue) are depicted at the bottom. (**D**) Tally of such LD-discordant haplotypes in a selection of 1KGP individuals, colored by population, as well as GRCh38 and T2T-CHM13. (**E**) Examples of variants that cannot be lifted over to T2T-CHM13 because of structural differences between the genomes. The position of the reference allele in GRCh38 is shown in red.

more than 300 bp to uniquely map. Concomitantly, T2T-CHM13 possesses fewer exactly duplicated sequences (≥5 kbp) shared across chromosomes (excluding sequence pairs within centromeres) than GRCh38 (figs. S7 and S8). Specifically, GRCh38 possesses 28 large shared interchromosomal sequences, primarily consisting of pairs of subtelomeric sequences, with an additional 42 pairs involving at least one unplaced contig. All of these identical sequence pairs, except for one between two subtelomeres, are nonidentical in T2T-CHM13, as small but important differences between repetitive elements have now been resolved (*27*, *33*).

T2T-CHM13 accurately represents the haplotype structure of human genomes

The human reference genome serves as the standard to which other genomes are compared, and is typically perceived as a haploid representation of an arbitrary genome from the population (*25*). In contrast with T2T-CHM13, which derives from a single homozygous complete hydatidiform mole, the Human Genome Project constructed the current reference genome via the tiling of sequences obtained from bacterial artificial chromosomes (BACs) and other clones with lengths ranging from ~50 to >250 kbp (*24*), which derived from multiple donor individuals. GRCh38 and its predecessors thus comprise mosaics of many haplotypes, albeit with a single library (RP11) contributing the majority (*24*).

To further characterize this aspect of GRCh38 and its implications for population studies, we performed local ancestry inference for both GRCh38 and T2T-CHM13 through comparison to haplotypes from the 1KGP (*34*) (Fig. 1A and figs. S2 and S9). Continental superpopulation-level ancestry was inferred for 72.9% of GRCh38 clones on the basis of majority votes of nearest-neighbor haplotypes. For the remaining 27.1% of clones, no single superpopulation achieved a majority of nearest neighbors, and ancestry thus remained ambiguous. This ambiguity occurred primarily for short clones with few informative SNPs (fig. S10), but also for some longer clones with potential admixed ancestry.

In accordance with Green *et al.* (*24*), we inferred that library RP11, which constitutes 72.6% of the genome, is derived from an individual of admixed African-American ancestry, with 56.0% and 28.1% of its component clones assigned to African and European local ancestries, respectively. The second most abundant library, CTD (5.5% of the genome), consists of clones of predominantly (86.3%) East Asian local ancestries, and the remaining libraries are derived from individuals of predominantly European ancestries. In contrast, T2T-CHM13 exhibits European ancestries nearly genome-wide (fig. S11). In addition, GRCh38 and T2T-CHM13 harbor 26.7 Mbp and 51.0 Mbp, respectively, of putative Neanderthal-introgressed sequences that originated from ancient interbreeding between the two hominin groups ~60,000 years ago (*24*). The excess of introgressed sequence in T2T-CHM13, even when restricting to the genomic intervals of GRCh38 clones with confident ancestry assignments, is consistent with its greater proportion of non-African ancestry.

We hypothesized that the mosaic nature of GRCh38 would generate abnormal haplotype structures at the boundaries of clones used for its construction, producing combinations of alleles that are rare or absent from the human population. Indeed, some previous patches of the reference genome sought to correct abnormal haplotype structures wherever noticed because of their impacts on genes of clinical importance (e.g., *ABO* and *SLC39A4*) (*3*). Such artificial haplotypes would mimic rare recombinant haplotypes private to any given sample, but at an abundance and genomic scale unrepresentative of any living human. To test this hypothesis, we identified pairs of common [minor allele frequency (MAF) > 10%] SNP alleles always observed on the same haplotype [i.e., segregate in perfect (R^2 = 1) linkage disequilibrium (LD)] in the 2504 unrelated individuals of the 1KGP and queried the allelic states of these SNPs in both GRCh38 and T2T-CHM13 (*34*).

In accordance with our expectations, we identified numerous haplotype transitions in GRCh38 absent from the 1KGP samples, with 18,813 pairs of LD-discordant SNP alleles (i.e., in perfect negative LD) distributed in 1390 narrow nonoverlapping clusters (median length = 3703 bp) throughout the genome (Fig. 1C). Such rare haplotype transitions are comparatively scarce in T2T-CHM13, with only 209 pairs of common high-LD SNPs (50 nonoverlapping clusters) possessing allelic combinations absent from the 1KGP sample (Fig. 1D). Using a leave-one-out analysis, we confirmed that T2T-CHM13 possesses a similar number of LD-discordant haplotypes as phased "haploid" samples from 1KGP, whereas GRCh38 vastly exceeds this range (Fig. 1E). By intersecting the GRCh38 results with the tiling path of BAC clones, we found that 88.9% (16,733 of 18,813) of discordant SNP pairs straddle the documented boundaries of adjacent clones (fig. S12). Of these, 45.9% (7686 of 16,733) of the clone pairs derived from different BAC libraries, whereas the remainder likely largely reflects random sampling of distinct homologous chromosomes from the same donor individual. Thus, our analysis suggests that T2T-CHM13 accurately reflects haplotype patterns observed in contemporary human populations, whereas GRCh38 does not.

T2T-CHM13 corrects genomic collapsed duplications and falsely duplicated regions

Genome assemblies often suffer from errors in complex genomic regions such as SDs. In the case of GRCh38, targeted sequencing of BAC clones has been performed to fix many such loci (*3*, *16*, *35*–*38*), but problems persist. To systematically identify errors in GRCh38 that could produce spurious variant calls, we leveraged the fact that T2T-CHM13 is an effectively haploid cell line that should produce only homozygous variants when its sequence is aligned to GRCh38. Thus, any apparent heterozygous variant can be attributed to mutations accrued in the cell line, sequencing errors, or read mapping errors. In the last case, assembly errors or copy number polymorphism of SDs produce contiguous stretches of heterozygous variants (*39*), which confound the accurate detection of paralog-specific variants (PSVs). By mapping PacBio HiFi reads from the CHM13 cell line (*27*) as well as Illumina-like simulated reads (150 bp) obtained from the T2T-CHM13 reference to GRCh38, we identified 368,574 heterozygous SNVs within the autosomes and chromosome X, of which 56,413 (15.3%) were shared between datasets. This evidence shows that each technology is distinctively informative as a result of differences in mappability (fig. S13 and table S1).

To home in on variants deriving from collapsed duplications, we delineated "clusters" of heterozygous calls (*34*) and identified 908 putative problematic regions (541 supported by both technologies) comprising 20.8 Mbp (Fig. 1 and fig. S13). Many of these loci intersected SD-associated regions (668/908; 73.6%) and centromere-associated regions (542/908; 59.7%) (*31*) as well as known GRCh38 issues (341/908; 37.55%). Variants flagged as excessively heterozygous in the population by gnomAD (*6*) were significantly enriched in these regions (10,000 permutations, empirical $P = 1 \times 10^{-4}$), representing 23.6% (87,005/368,574) of our discovered CHM13 heterozygous variants, which suggests that these spurious variants arise in genome screens and represent false positives (Fig. 1A and figs. S1 and S3).

We next "lifted over" (i.e., converted the coordinates of) 821 of these 908 putative problematic regions to the T2T-CHM13 assembly and used human copy number estimates [n = 268 individuals from the Simons Genome Diversity Project (SGDP)] (*32*, *40*) to conservatively identify 203 loci (8.04 Mbp) evidencing missing copies in GRCh38 (fig. S14). These regions have an impact on 308 gene features, with 14 of the total 48 protein-coding genes fully contained within a problematic region, indicating that complete gene homologs are hidden from GRCh38-based population analyses of variation. Examples include *DUSP22*, a gene involved in immune regulation (*41*), as well as *KMT2C*, a gene implicated in Kleefstra syndrome 2 (*42*) (fig. S15). Additionally, we identified 30 SNPs within problematic regions with known phenotype associations from the GWAS Catalog (*43*). Finally, we evaluated the

status of these regions in the T2T-CHM13 reference by following a similar approach to obtain 9193 heterozygous variants clustered in 11 regions—none of which overlapped GRCh38 problematic regions (table S2). As a result, we are now able to call variants in these 48 previously inaccessible protein-coding genes. We did identify one putative collapsed duplication in T2T-CHM13, based on the presence of a heterozygous variant cluster and reduced copy number in T2T-CHM13, localized to an rDNA array corrected in the most recent version of T2T-CHM13v1.1 (*27*).

Conversely, the T2T-CHM13 reference also corrects regions falsely portrayed as duplicated in GRCh38. Specifically, we identified 12 regions affecting 1.2 Mbp and 74 genes (including 22 protein-coding genes) with duplications private to GRCh38 and not found in T2T-CHM13 or the 268 genomes from SGDP (*40*) (fig. S14 and table S3). In contrast, only five regions affecting 160 kbp have duplications in T2T-CHM13 that are not in GRCh38 or the SGDP, which suggests that genuine rare variation cannot explain the excess of private duplications in GRCh38. Indeed, upon inspecting the CHM13 data, we deemed that these five loci are true duplications with support from mapped HiFi reads (*30*).

The five largest duplications in GRCh38, affecting 15 protein-coding genes on the q-arm of chromosome 21, involve BAC clones with sequence misplaced between gaps on the heterochromatic p-arm of the same chromosome. The Genome Reference Consortium (GRC)—an international team of researchers that has maintained and improved the reference genome and related resources since its initial publication—determined that admixture mapping incorrectly localized these five clones to the acrocentric short arm and therefore should not have been added to GRCh38 (*34*). Of the seven false duplications outside chromosome 21, two occur in short contigs between gaps, two occur adjacent to a gap, two occur on unlocalized "random" contigs, and one occurs as a tandem duplication (table S4). We provide an exhaustive list of falsely duplicated gene pairs corrected in T2T-CHM13 (table S5). Thus, T2T-CHM13 authoritatively corrects many false duplications, improving variant calling for short- and long-read technologies, including in medically relevant genes.

Liftover of clinically relevant and trait-associated variation from GRCh38 to T2T-CHM13

In transitioning to a different reference genome, it is imperative to document the locations of known genetic variation of biological and clinical relevance with respect to the updated coordinate system. To this end, we sought to lift over 802,674 unique variants in the ClinVar database and 736,178,420 variants from the NCBI dbSNP database (including 151,876 NHGRI-EBI GWAS Catalog variants) from the GRCh38 reference to the T2T-CHM13 reference. Liftover was successful for 800,942 (99.8%) ClinVar variants, 723,117,125 (98.2%) NCBI dbSNP variants, and 150,962 (99.4%) GWAS Catalog SNPs (table S6). We provide these lifted-over datasets as a resource for the scientific community within the UCSC Genome Browser and the NHGRI AnVIL, along with lists of all variants that failed liftover and the associated reasons (Fig. 1, A and B, and figs. S1 and S4). Critically, this resource includes 138,319 of 138,927 (99.6%) ClinVar variants annotated as "pathogenic" or "likely pathogenic."

Of the 1732 ClinVar variants that failed to lift over, 1186 overlap documented insertions or deletions that distinguish the GRCh38 and T2T-CHM13 assemblies. The remaining 546 variants (<0.1% of all variants) lie within regions of poor alignment between the GRCh38 and T2T-CHM13 assemblies (Fig. 1E). The modes of liftover failure for variants in dbSNP and the GWAS Catalog follow similar distributions (table S6). In all, these annotated variants offer a resource to enable researchers to interpret genetic results using the T2T-CHM13 assembly.

T2T-CHM13 improves analysis of global genetic diversity based on 3202 short-read samples from the 1KGP dataset

T2T-CHM13 improves short-read mapping across populations

To investigate how the T2T-CHM13 assembly affects short-read variant calling, we realigned and reprocessed all 3202 samples from the 1KGP cohort (*28*) using the NHGRI AnVIL Platform (*44*) (figs. S16 and S17). In this collection, each sample is sequenced to at least 30× coverage with paired-end Illumina sequencing, with samples from 26 diverse populations across five major continental superpopulations (fig. S18). Although most samples are unrelated, the expanded collection includes 602 complete trios that we used to estimate the rate of false variants below on the basis of discordance with Mendelian expectations. We matched the analysis pipeline for GRCh38 (*28*) as closely as possible so that any major differences would be attributable to the reference genome rather than technical differences in the analysis software (*34*).

On average, BWA-MEM (*45*) maps an additional 7.4×10^6 (0.97%) of properly paired reads to T2T-CHM13 compared to GRCh38, even when considering the alternative (ALT) and decoy sequences used in the original analysis (fig. S19). Interestingly, even though more reads align to T2T-CHM13, the subsequent per-read mismatch rate is 20 to 25% lower across all continental populations. African samples continue to present the highest mismatch rate (Fig. 2A), as the observed mismatch rate includes both genuine sequencing errors, which are largely consistent across all samples, and any true biological differences between the read and the reference genome, which vary substantially according to the ancestry of the sample. Relatedly, T2T-CHM13 improved other mapping characteristics, including reducing the number of misoriented read pairs (Fig. 2A). Finally, by considering the alignment coverage across 500-bp bins across the respective genomes, we observed improvement in coverage uniformity within every sample's genome when using T2T-CHM13 rather than GRCh38. For example, within gene regions, we noted a factor of 4 decrease in the standard deviation of the coverage (Fig. 2A) and similar improvements in other types of genomic regions among all population groups (fig. S20). Overall, these improvements in error rates, mapping characteristics, and coverage uniformity demonstrate the superiority of T2T-CHM13 as a reference genome for short-read alignment across all populations.

T2T-CHM13 improves variant calling across populations

From these alignments, we next generated SNV and small indel variant calls with the GATK Haplotype Caller, which uses a joint genotyping approach to optimize accuracy across large populations (*46*). Again, we matched the pipeline used in the prior 1KGP study, albeit with updated versions of some analysis tools, to minimize software discrepancies and attribute differences to changes in the reference genome. Across all samples, we identified 126,591,489 high-quality ("PASS") variants relative to T2T-CHM13 (per-sample mean, 4,717,525; median, 4,419,802) compared to 125,484,020 variants relative to GRCh38 (per-sample mean, 5,101,897; median, 4,867,871), additionally noting a decrease in the number of called variants per individual genome (Fig. 2B and fig. S21). We performed all subsequent analyses using these high-quality variants, as the PASS filter successfully removed spurious variants (fig. S22), particularly in complex regions (fig. S23).

As with the improvement to the per-read mismatch rate, we attribute the reduction in the number of per-sample variant calls to improvements in the number of rare alleles, consensus errors, and structural errors in T2T-CHM13. This conclusion is supported by the observation that the number of heterozygous variants per sample is more similar (Fig. 2C and fig. S24) across reference genomes in contrast to homozygous variants (Fig. 2D and fig. S24). This discrepancy is especially pronounced in non-African samples, which have on average 200,000 to 300,000 more homozygous variants relative to GRCh38 than to T2T-CHM13, likely because ~70% of the GRCh38 sequence comes from an individual with African-American ancestry, and African populations are enriched for rare and private variants (*18*).

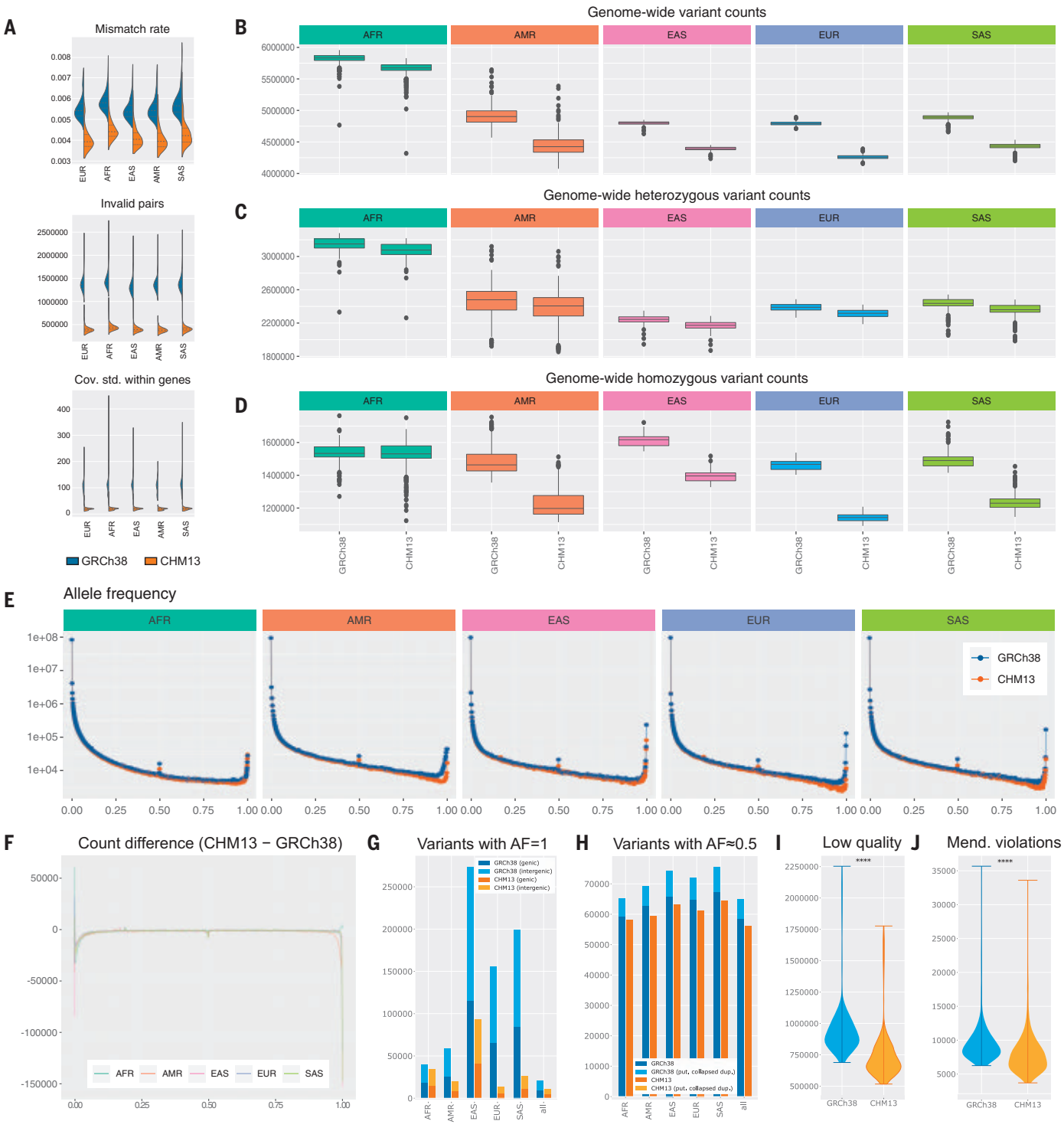


Fig. 2. Improvements to short-read mapping and variant calling. (**A**) Summary of alignment characteristics aligning to T2T-CHM13 instead of GRCh38. (**B**) Boxplot of overall number of variants found in each person across superpopulations, with colors indicated in Fig. 1A legend. (**C**) Boxplot of the number of heterozygous variants found in each person across superpopulations. (**D**) Boxplot of the number of homozygous variants found in each person across superpopulations. (**E**) AF distribution of each superpopulation relative to T2T-CHM13 and GRCh38. (**F**) Change in AF distribution. (**G**) Number of variants with AF equal to 100%, both within protein-coding genes and without. (**H**) Number of variants with AF equal to 50%, both within putative collapsed duplications and without. (**I**) Violin plot of the number of low-quality variants found when aligning to GRCh38 and T2T-CHM13. (**J**) Violin plot of the number of Mendelian violations found when aligning to GRCh38 and T2T-CHM13.

Further investigating this relationship, we computed the AFs of variants from unrelated samples from each of the five continental superpopulations (Fig. 2E). Although the distributions were nearly equivalent over most of the AF spectrum, we observed substantial differences for rare alleles (AF < 0.05); intermediate-frequency alleles, including errors where nearly all individuals are heterozygous (AF ≈ 0.5); and fixed or nearly fixed alleles (AF > 0.95). The most prominent difference in AF distributions affected fixed or nearly fixed alleles in each assembly, for which all non-African superpopulations showed an excess of ~150,000 variants in GRCh38, whereas the African superpopulation showed an excess of 2364 variants in T2T-CHM13 (Fig. 2F). This observation was driven by a decrease in the number of completely fixed variants (100% AF) relative to GRCh38 (Fig. 2G). Such variants represent positions where the reference genome itself is the only sample observed to possess the corresponding allele. These alleles arise either because of genuine private variants in one of the GRCh38 donors, or from errors in the reference genome itself, and result in 100% of other individuals possessing two copies of the alternative allele. As a result, these "variants" will not be reported at all if the same reads are mapped to a different genome that does not have these private alleles. Interestingly, the number of such private "singleton" variants in T2T-CHM13 lies squarely within the observed range of singleton counts among 1KGP samples, adjusting for the difference in ploidy (fig. S25). In addition to the lower rate of private variation relative to GRCh38, T2T-CHM13 possesses fewer ultra-rare variants, effectively reducing the number of "nearly fixed" alleles in population data such as the 1KGP.

Finally, the reduction in AF ≈ 0.5 variants is largely explained by the corrections to collapsed SDs (table S1), as these regions are highly enriched for heterozygous PSVs in nearly all individuals. Such erroneous heterozygous variants are caused by the false pileup of reads from the duplicated regions to a single location (Fig. 2H). Collectively, the decrease in variants with AF = 1 and AF ≈ 0.5 largely explains the decrease in the overall number of variants observed per sample and across the entire population for T2T-CHM13.

Informed by these results, we considered the feasibility of calling variants using the T2T-CHM13 reference and then lifting over the results to GRCh38 for further analyses. Using a liftover tool to transform a variant call set for a single sample into a call set with respect to GRCh38 requires special handling to account for variants for which the two references have different alleles. Specifically, if one of the reference alleles is not present in the sample, it will be necessary to genotype the site against the T2T-CHM13 reference. Although this issue is less of a concern for large datasets such as the 1KGP, even these large samples will contain a small number of variants that become invisible when switching reference genomes (fig. S26). In addition, differences in variant representation, especially in regions of low complexity, may cause lifted variant sets to differ from those called against the target reference.

Reduction of Mendelian-discordant variants

As further quality control for the variant calls, we performed a Mendelian concordance analysis using the 602 trios represented in the 1KGP cohort. We observed a statistically significant decrease in both the number of low-quality variants [median, 890,701 (GRCh38) versus 682,609 (T2T-CHM13); $P = 4.943 \times 10^{-96}$, Wilcoxon signed-rank test] (Fig. 2I) and the number of Mendelian-discordant variants (i.e., variants found in children but not their parents, or homozygous parental variants not observed in their children) [median, 8879 (GRCh38) versus 7484 (T2T-CHM13); $P = 7.346 \times 10^{-96}$, Wilcoxon signed-rank test] (Fig. 2J) when aligned to T2T-CHM13 as compared to GRCh38. In addition to providing an estimate of the error rate for variant calls in this call set, this improvement has broad implications for clinical genetics analyses of de novo or somatic mutations, which have been implicated as causes of autism spectrum disorders (*47*) and many forms of cancer (*48*).

T2T-CHM13 improves SV analysis of 17 diverse long-read samples

T2T-CHM13 improves long-read mapping across populations

Next, we investigated the effects of using T2T-CHM13 as a reference genome for alignment and large SV calling from both PacBio HiFi and ONT long reads. To this end, we aligned reads and called SVs in 17 samples of diverse ancestries from the Human Pangenome Reference Consortium (HPRC+) (*27*) and the Genome in a Bottle Consortium (GIAB) (*29*), including two trios (table S7). All of these samples had HiFi data available, and 14 had also been sequenced with ONT (Fig. 3A), with mean read lengths of 18.1 kbp and 21.9 kbp and read N50 values of 18.3 kbp and 44.9 kbp, respectively (fig. S27).

In line with our short-read results, aligning long reads to T2T-CHM13 versus GRCh38 did not substantially change the number of reads mapped with either Winnowmap (*49*) or minimap2 (*50*) because most of the previously unresolved sequence in T2T-CHM13 represents additional copies of SDs or satellite repeats already partially represented in GRCh38 (fig. S28). However, aligning to T2T-CHM13 reduced the observed mismatch rate per mapped read by 5% to 40% across the four combinations of sequencing technologies and aligners because GRCh38 has more rare alleles. T2T-CHM13 also corrects structural errors in GRCh38 and is a complete assembly of the genome, which facilitates accurate alignment, similar to what we observed for short reads (Fig. 3B). Relatedly, we found that previously reported African-specific (*51*) and Icelandic-specific (*52*) sequences at least 1 kbp in length align with substantially greater identity and completeness to T2T-CHM13 than to GRCh38 (*34*) (fig. S29 and S30).

To study coverage uniformity, we next measured the average coverage across each 500-bp bin on a per-sample basis and computed the standard deviation of the coverage. Across all aligners and technologies, the median standard deviation of the per-bin coverage was reduced by more than a factor of 3, indicating more stable mapping to T2T-CHM13 (Fig. 3B). This difference in coverage uniformity was pronounced in satellite repeats and other regions of GRCh38 that are nonsyntenic with T2T-CHM13 (fig. S31 and S32). This coverage uniformity will broadly improve variant calling and other long read–based analyses.

T2T-CHM13 improves the balance of apparent indels

We next used our optimized SV-calling pipeline, including Sniffles (*53*), Iris, and Jasmine (*54*), to call SVs in all 17 samples (figs. S33 and S34) and consolidate them into a cohort-level call set in each reference from HiFi data. From these results, we observed a reduction from 5147 to 2260 SVs that are homozygous in all 17 individuals when calling variants relative to T2T-CHM13 instead of GRCh38 (Fig. 3C). Previous studies (*16*, *17*) have noted the excess of such SV calls when using GRCh38 as a reference and attributed them to structural errors. Here, we found that using a complete and accurate reference genome naturally reduces the number of such variants. In addition, the number of indels was more balanced when calling against T2T-CHM13, whereas GRCh38 exhibited a bias toward insertions caused by missing or incomplete sequence (Fig. 3D), such as incorrectly collapsed tandem repeats (*16*). With respect to T2T-CHM13, we observed a small bias toward deletions, which likely results from the challenges in calling insertions with mapping-based methods and in representing SVs within repeats, as this difference is especially prominent in highly repetitive regions such as centromeres and satellite repeats (fig. S35). The variants we observed relative to T2T-CHM13 are enriched in the centromeres and subtelomeric sequences, likely because of a combination of repetitive sequence and greater recombination rates (*17*). We observed similar trends among SVs unique to single samples (fig. S36).

We also observed similar improvements in the insertion/deletion balance for large SVs

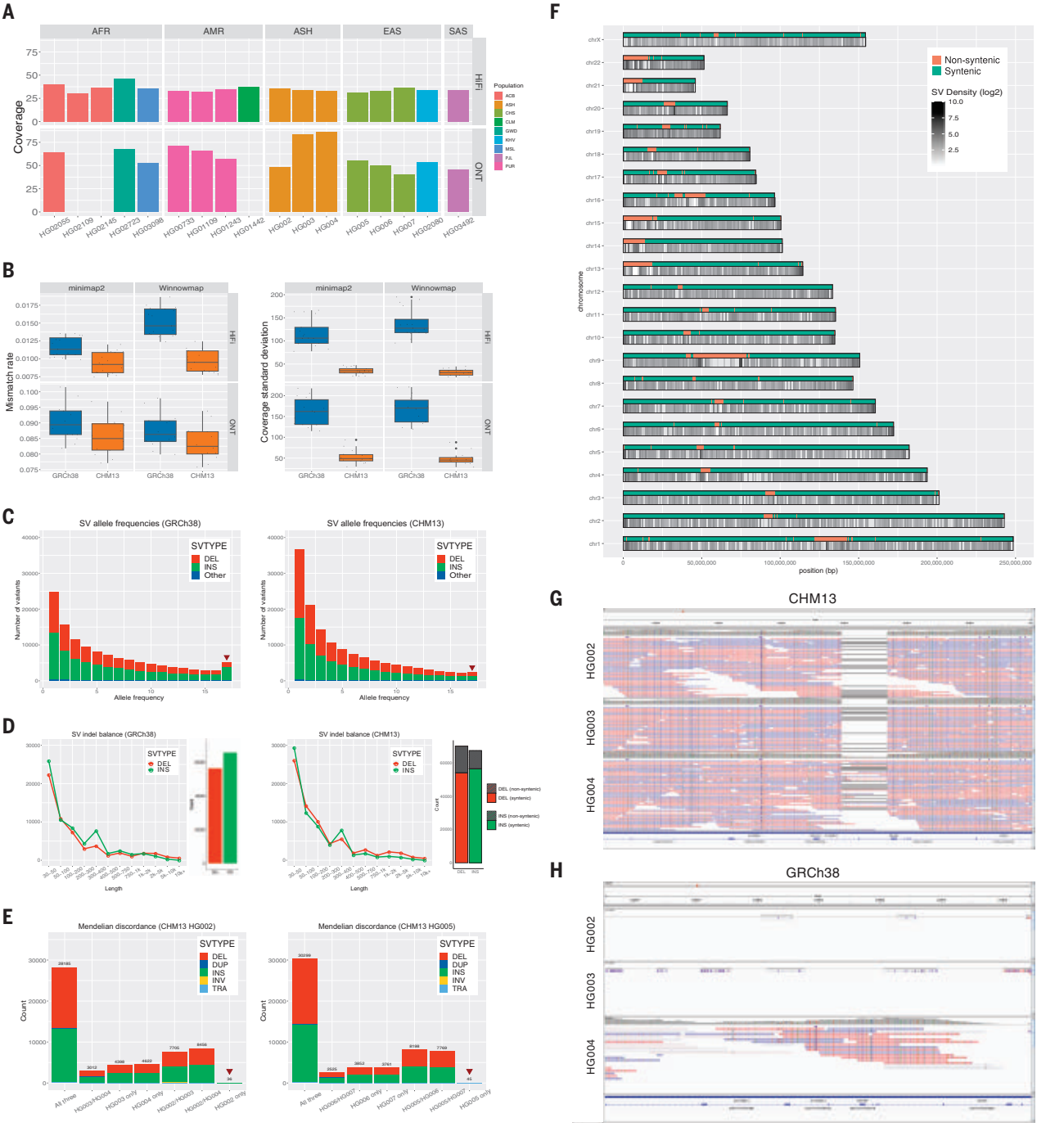


Fig. 3. Improvements to long-read alignment and SV calling in CHM13. (**A**) The coverage, ancestry, and sequencing platforms available for the 17 samples sequenced with long reads (headers: AFR, African; AMR, Admixed American; ASH, Ashkenazi; EAS, East Asian; SAS, South Asian; populations: ACB, African Caribbean in Barbados; ASH, Ashkenazi; CHS, Southern Han Chinese; CLM, Colombian in Medellin, Colombia; GWD, Gambian in Western Division, The Gambia; KHV, Kinh in Ho Chi Minh City, Vietnam; MSL, Mende in Sierra Leone; PJL, Punjabi in Lahore, Pakistan; PUR, Puerto Rican in Puerto Rico). (**B**) The genome-wide mapping error rate and the standard deviation of the coverage for T2T-CHM13 (orange) and GRCh38 (blue). The standard deviation was computed across each 500-bp bin of the genome. (**C**) The allele frequency of SVs derived from HiFi data in T2T-CHM13 and GRCh38 among the 17-sample cohort. The red arrows indicate fixed (100% frequency) variants. (**D**) The balance of insertions (INS) vs. deletion (DEL) calls in the 17-sample cohort in T2T-CHM13 and GRCh38. Variants in T2T-CHM13 are stratified by whether or not they intersect regions which are nonsyntenic with GRCh38. (**E**) The SV calls in T2T-CHM13 for two trios: a trio of Ashkenazi ancestry [child HG002, and parents HG003 (46XY), and HG004 (46XX)], and a trio of Han Chinese ancestry [child HG005, and parents HG006 (46XY) and HG007 (46XX)]. The red arrows indicate child-only, or candidate de novo, variants (DEL, Deletion; DUP, Duplication; INS, Insertion; INV, Inversion; TRA, Translocation). (**F**) The density of SVs called from HiFi data in the 17-sample cohort across T2T-CHM13. (**G**) Alignments of HiFi reads in the HG002 trio to T2T-CHM13 showing a deletion spanning an exon of the transcript AC134980.2 viewed using the Integrative Genomic Viewer (IGV). Pink horizontal rectangles indicate reads aligned to the forward strand; blue horizontal rectangles indicate reads aligned to the reverse strand. Thin black lines indicate split-read alignments. Small vertical rectangles indicate SNVs (**H**) Alignments of HiFi reads in the HG002 trio to the same region of GRCh38 as shown in (G), showing much poorer mapping to GRCh38 than to T2T-CHM13, viewed using IGV with colors same as (G).

(>500 bp) detected by Bionano optical mapping data in HG002 against the T2T-CHM13 reference, with an increase in deletions (1199 versus 1379) and a decrease in insertions (2771 versus 1431) with GRCh38 and T2T-CHM13, respectively (fig. S37). Using the T2T-CHM13 reference for Bionano optical mapping also improved SV calling around gaps in GRCh38 that are closed in T2T-CHM13 (fig. S38), which suggests that T2T-CHM13 offers improved indel balance relative to GRCh38 across multiple SV-calling methods.

T2T-CHM13 facilitates the discovery of de novo SVs

To investigate the impacts of the T2T-CHM13 reference on our ability to accurately detect de novo variants, we called SVs in both of our trio datasets using a combination of HiFi and ONT data and identified SVs only present in the child of the trio and supported by both technologies—approximately 40 variants per trio (Fig. 3E). Manual inspection revealed a few variants in each trio that were strongly supported with consistent coverage and alignment breakpoints, whereas the other candidates exhibited less reliable alignments, as noted in previous reports (*54*). In HG002, we detected six strongly supported candidate de novo SVs that had been previously reported (*29*, *54*). In HG005, we detected a 1571-bp deletion at chr17:49401990 in T2T-CHM13 that was supported as a candidate de novo SV relative to both T2T-CHM13 and GRCh38 (fig. S39). This demonstrates the ability of T2T-CHM13 to be used as a reference genome for de novo SV analysis.

T2T-CHM13 enables the discovery of additional SVs within previously unresolved sequences

The improved accuracy and completeness of the T2T-CHM13 genome make it easier to resolve complex genomic regions. Within nonsyntenic regions, we identified a total of 27,055 SVs (Fig. 3D), the majority of which were deletions (15,998) and insertions (10,912). Of these SVs, 22,362 (82.7%; 8903 insertions, 13,334 deletions) overlap previously unresolved sequences in T2T-CHM13, whereas the remaining SVs are now accessible because of the accuracy of the T2T-CHM13 reference. The AF and size distributions for these variants mirror the characteristics of the syntenic regions, with rare variants (fig. S40) and smaller (30 to 50 bp) indels (fig. S41) being the most abundant. However, we also note some nonsyntenic regions with few or zero SVs identified. Many of these regions lie at the interiors of p-arms of acrocentric centromeres, which are gaps in T2T-CHM13v1.0 that have been resolved in later versions of the assembly; however, we also noticed depletions of SVs in a few other highly repetitive regions, such as the resolved human satellite array on chromosome 9 (Fig. 3F). We largely attribute the reduction in variant density to the low mappability of these complex and repetitive regions. Future improvements in read lengths and alignment algorithms are needed to further resolve such loci.

Within syntenic regions, we also noted improvements to alignment and variant calling accuracy, including the identification of variant calls not previously observed within homologous regions of GRCh38. For example, in T2T-CHM13, we observed a deletion in all of the samples of the HG002 trio in an exon of the olfactory receptor gene *AC134980.2* (Fig. 3G), whereas the reads from those samples largely failed to align to the corresponding region of GRCh38 (Fig. 3H). Meanwhile, reads from African samples (fig. S42) aligned to both references at this locus. The difference in alignment among different samples is likely because the region is highly polymorphic for copy number variation; GRCh38 contains a reasonable representation of that region for the tested African samples, whereas the homologous region in T2T-CHM13 more closely resembles European samples (fig. S43). This highlights the need for T2T reference genomes for as many diverse individuals as possible to account for common haplotype diversity.

Variation within previously unresolved regions of the genome

T2T-CHM13 enables variant calling in previously unresolved and corrected regions of the genome

The T2T-CHM13 genome contains 229 Mbp of sequence that is nonsyntenic to GRCh38, which intersects 207 protein-coding genes (Fig. 4A and Table 1). Within these regions, we report 3,692,439 PASS variants across all 1KGP samples from short reads (Fig. 4B and Table 1). Comparing variants called in a subset of 14 HPRC+ samples with Illumina, HiFi, and ONT data, we found that 73 to 78% of the Illumina-discovered SNVs are concordant with variants identified with PacBio HiFi long-read data using the PEPPER-Margin-DeepVariant algorithm (51,306 to 74,122 matching SNVs and genotypes per sample) (*55*). Long reads discovered more than 10 times as many SNVs per sample as short reads in these regions, with 447,742 to 615,085 (41 to 43%) SNVs matching between HiFi and ONT with PEPPER-Margin-DeepVariant. In nonsyntenic regions, 97% of the SNVs called by HiFi fell in centromeric regions of CHM13, so we stratified concordance by type of satellite repeat within the centromere. We found that nonsatellites in centric transition regions and monomeric satellites had higher concordance between HiFi and ONT, with >99% concordance in a few regions, but some as low as 50%. Human satellite (HSAT) regions, which pose some of the greatest challenges for read mapping and harbor abundant structural variation, exhibit the lowest rates of concordance between the platforms.

We further defined conservative high-confidence regions by excluding regions with abnormal coverage in any long-read sample (i.e., coverage outside of 1.5× the interquartile range). This effectively excludes difficult-to-map regions with excessively repetitive alignments as well as copy number–variable regions. After excluding abnormal coverage from nonsyntenic regions, 14 Mbp remained, and SNVs from HiFi and ONT long reads were 91 to 95% concordant (21,835 to 28,237 variants). We found 95 to 96% (14,575 to 18,949) of short-read SNVs in HiFi long-read calls, although 37 to 40% of HiFi SNVs were still missing from the short-read calls as a result of poorer mappability of the short reads (table S8). Although many nonsyntenic regions will require further method development [e.g., pangenome references (*56*)] to achieve accurate variant calls, the concordance of long- and short-read calls for tens of thousands of variants highlights previously unresolved sequences that are immediately accessible to both technologies.

Because these broadly defined nonsyntenic regions include inversions and other structural changes between GRCh38 and T2T-CHM13 that do not necessarily alter many of the variants contained within, we also considered a narrower class of "previously unresolved" sequences, representing segments of the T2T-CHM13 genome that do not align to GRCh38 with Winnowmap (*49*). Within these previously unresolved sequences, which span a total of 189 Mbp (Fig. 4A, Table 1, and fig. S44), we report a total of 2,370,384 PASS variants in 1KGP samples based on short reads, intersecting 207 protein-coding genes (Fig. 4B, Table 1, and fig. S45). We note that this set of 207 genes is distinct from the 207 genes that intersected with the nonsyntenic regions, and these two sets together comprise 329 unique genes. Because these previously unresolved sequences are enriched for highly repetitive sequences, concordance is slightly lower, such that 64 to 69% of the SNVs in each sample match variants found in PacBio HiFi long-read data from the same samples (24,371 to 36,501 matching SNVs and genotypes per sample), and 339,783 to 473,074 (38 to 40%) of SNVs match between HiFi and ONT. When removing difficult-to-map and copy number–variable regions as above, 3 Mbp of high-confidence regions remained. Within high-confidence regions, 84 to 88% of short-read SNVs in each sample matched variants found in each sample's PacBio HiFi long-read data (2938 to 3811 matching SNVs and genotypes per sample), and 5544 to 8298 (81 to 90%) of SNVs matched between HiFi and ONT (table S8). Although these previously unresolved regions are more challenging than nonsyntenic regions, thousands of variants can still be called concordantly with short and long reads.

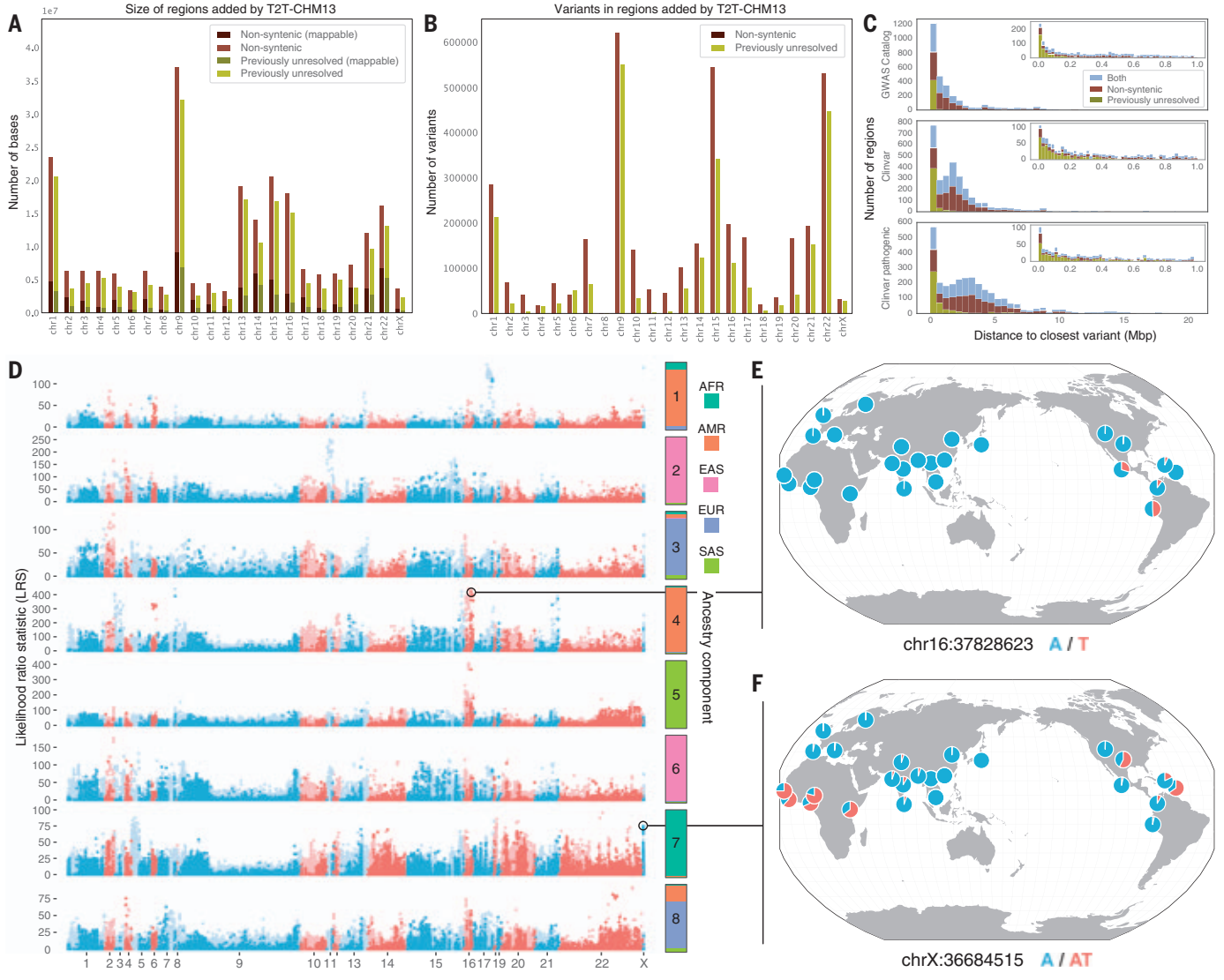


Fig. 4. Characterization of variants within regions of the genome resolved by T2T-CHM13. (**A**) Number of bases added in nonsyntenic and previously unresolved regions by chromosome, along with how many variants for each respective region are mappable (have contiguous unique 100mers). (**B**) Number of variants in nonsyntenic and previously unresolved regions by chromosome. (**C**) Distance from each previously unresolved–only, nonsyntenic-only, or overlapping region to the closest Clinvar or GWAS Catalog variant. Insets are zoomed to 1 Mbp. (**D**) Scan for variants in nonsyntenic (light blue and red) and previously unresolved (dark blue and red) regions that exhibit extreme patterns of allele frequency differentiation. Allele frequency outliers were identified for each of eight ancestry components, colored by the superpopulation membership of the corresponding 1KGP samples. Large values of the likelihood ratio statistic (LRS) denote variants for which AF differences in the corresponding ancestry component exceeds that of a null model based on genome-wide covariances in allele frequencies. (**E** and **F**) Population-specific allele frequencies of two highly differentiated variants in previously unresolved regions.

We noted homology between GRCh38 collapsed duplications and many T2T-CHM13 nonsyntenic and/or previously unresolved regions (137 regions comprising 6.8 Mbp), indicating that the T2T-CHM13 assembly corrects these sequences through the deconvolution of nearly identical repeats. Comparing total variants identified in the 1KGP dataset, we observed a significant decrease in variant densities of 41 protein-coding genes intersecting with GRCh38 collapsed duplications in T2T-CHM13 (mean, 27 variants per kbp) compared with GRCh38 (mean, 46 variants per kbp; $P = 6.906 \times 10^{-8}$, Wilcoxon signed-rank test) (fig. S46). Besides differences in local ancestries between the references, these higher variant densities in GRCh38 in part represent PSVs or misassigned alleles from missing paralogs (*57*). Conversely, 1KGP variants were significantly increased in 32 protein-coding genes contained within GRCh38 false duplications using the T2T-CHM13 reference genome (mean values of 48 variants per kbp in T2T-CHM13 versus 12 variants per kbp in GRCh38; $P = 4.657 \times 10^{-10}$, Wilcoxon signed-rank test).

To assess whether these corrected complex regions in T2T-CHM13 accurately reveal variation, we evaluated the concordance of variants generated from short-read Illumina and PacBio HiFi sequencing datasets of two trios from the GIAB consortium and the Personal Genome Project (*58*) and observed similar recall for Illumina data in T2T-CHM13 (20.1 to 28.3%) and GRCh38 (21.5 to 25.4%), but with improved precision in the variants identified (98.1 to 99.7% in T2T-CHM13 versus 64.3 to 67.3% in GRCh38) in a subset of the GRCh38 collapsed duplications (copy number < 10; ~910 kbp) (table S9). Corrected false duplications (1.2 Mbp) exhibited improved recall for Illumina data by a factor of 50 relative to HiFi

Table 1. Overview of nonsyntenic and previously unresolved regions and their respective variant counts.

	Nonsyntenic	Previously unresolved
Summary		
Total span (bp) (excluding Ns)	240,044,315 (228,569,315)	189,036,735 (177,561,735)
Unique span (100mers)	65,471,195	40,205,401
Protein-coding genes	207	207
Entire region		
1KGP SNVs + indels (within genes)	3,692,439 (138,829)	2,370,384 (52,567)
Short-read SNVs per sample	65,931 to 101,161	35,506 to 56,489
Long-read SNVs per sample	1,178,371 to 1,467,243	957,629 to 1,197,463
Short-read SNVs confirmed by long reads	73 to 78%	64 to 69%
Long-read SNVs identified in short reads	4 to 5%	3%
SNVs concordant between long reads	41 to 43%	38 to 40%
High-confidence regions (excluding coverage abnormalities)		
High-confidence region bases	13,683,528	2,987,935
Short-read SNVs confirmed by long reads in high-confidence regions	95 to 96%	84 to 88%
Long-read SNVs identified in short reads in high-confidence regions	60 to 63%	39 to 46%
SNVs concordant between long reads in high-confidence regions	91 to 95%	81 to 90%

in T2T-CHM13 (57.4 to 68.3%) versus GRCh38 (1.1 to 1.8%), as well as improved precision in T2T-CHM13 (98.5 to 99.3%) versus GRCh38 (76.5 to 95.8%) (table S9). These improvements show that variants can be discovered and genotyped in regions corrected by the T2T-CHM13 assembly.

Phenotypic associations and evolutionary signatures within nonsyntenic T2T-CHM13 regions

Sequences in the T2T-CHM13 assembly that are nonsyntenic with GRCh38 offer opportunities for future genetic studies. Several such loci lie in close proximity to variation that has been implicated in complex phenotypes or disease, supporting their potential biomedical importance. These include eight loci occurring within 10 kbp of GWAS hits and 19 loci within 10 kbp of ClinVar pathogenic variants (Fig. 4C). In addition, 113 of 22,474 GWAS hits (representing 0.5% of all variants in the studies we tested) segregated in LD ($R^2 \geq 0.5$) with variants in nonsyntenic regions, thereby expanding the catalog of potential causal variants for these GWAS phenotypes (*43*) (fig. S47 and table S10).

Using short read–based genotypes generated from the 1KGP cohort, we also searched for variants within nonsyntenic regions that exhibit large differences in AF between populations—a signature that can reflect historical positive selection or demographic forces shaping these previously inaccessible regions of the genome. To study these signatures, we applied Ohana (*59*), a method that models individuals as possessing ancestry from k components and tests for ancestry component–specific frequency outliers. Focusing on continental-scale patterns (k = 8; fig. S48), we identified 5154 unique SNVs and indels across all ancestry components that exhibited strong deviation from genome-wide patterns of AF (99.9th percentile of distribution for each ancestry component; Fig. 4D). These included 814 variants overlapping with annotated genes and 195 variants that intersected annotated exons.

We first focused on the 3038 highly differentiated nonsyntenic variants that lift over from T2T-CHM13 to GRCh38. These successful liftovers allowed us to make direct comparisons to selection results, generated with identical methods, using 1KGP phase 3 data aligned to GRCh38 (fig. S49) (*34*, *60*). For 41.3% of the lifted-over variants, we found GRCh38 variants within a 2-kbp window that possessed similar or higher likelihood ratio statistics for the same ancestry component, indicating that these loci were possible to identify in scans of GRCh38 (fig. S50). The remaining 58.7% of lifted-over variants may represent regions of the genome where differences in the T2T-CHM13 and 1KGP phase 3 variant calling or filtering procedures lead to discrepancies in AFs between these two datasets. They may also indicate regions whose more accurate representation in T2T-CHM13 improves variant calling enough to resolve previously unknown signatures of AF differentiation (fig. S51). We then investigated the 943 variants that could not be lifted over from T2T-CHM13 to GRCh38 and were located in both previously unresolved sequences and regions deemed mappable from unique 100-mer analysis. Some of these variants overlap with genes, including several annotated with RNA transcripts in regions not present in the GRCh38 assembly (Fig. 4D and table S11).

We highlight two loci that exhibit some of the strongest allele frequency differentiation observed across ancestry components. The first locus, located in a centromeric alpha satellite on chromosome 16, contains variants that reach intermediate allele frequency in the ancestry component corresponding to the Peruvian in Lima, Peru (PEL) and other admixed American populations of 1KGP [AFs, 0.49 in PEL; 0.20 in CLM (Colombian in Medellin, Colombia) and MXL (Mexican ancestry in Los Angeles, California); absent or nearly absent elsewhere; Fig. 4E and figs. S52 and S53]. Variants at the second locus, located in a previously unresolved T2T-CHM13 sequence on the X chromosome that contains a multi-kbp imperfect AT tandem repeat, exhibit high AFs in the ancestry component corresponding to African populations of 1KGP and low AFs in other populations (AFs, 0.67 in African populations and 0.014 in European populations; Fig. 4F and figs. S54 and S55). The variant at this locus with the strongest signature of frequency differentiation also lies within 10 kbp of two pseudogenes, *MOB1AP2-201* (MOB kinase activator 1A pseudogene 2) and *BX842568.1-201* (ferritin, heavy polypeptide-like 17 pseudogene).

We note that as a consequence of the repetitive nature of the sequences in which they reside, many of the loci that we highlight here remain challenging to genotype with short reads, and individual variant calls remain uncertain. Nonetheless, patterns of AF differentiation across populations are relatively robust to such challenges and can still serve as proxies for more complex SVs whose sequences cannot be resolved by short reads alone. The presence of population-specific signatures at these loci highlights the potential for T2T-CHM13 to reveal evolutionary signals in previously unresolved regions of the genome.

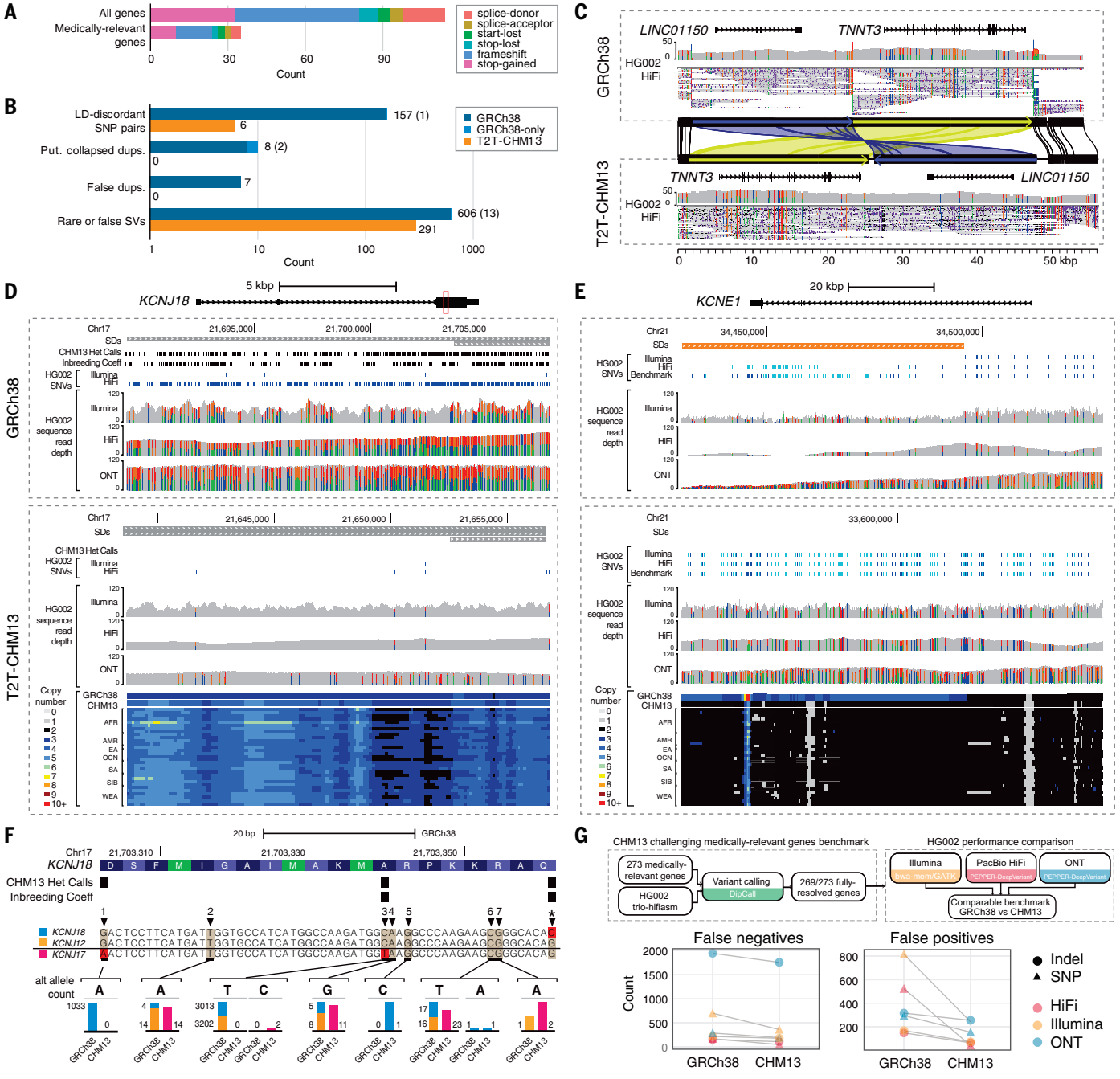


Fig. 5. T2T-CHM13 improves clinical genomics variant calling. (**A**) Numbers of potential loss-of-function mutations in the T2T-CHM13 reference. (**B**) The counts of medically relevant genes affected by genomic features and variation in GRCh38 (blue) and T2T-CHM13 (orange) are depicted as bar plots on logarithmic scale. Light blue indicates genes affected in GRCh38 where homologous genes were not identified in T2T-CHM13 due to inability to lift over, with counts included in parentheses. (**C**) An example erroneous GRCh38 complex SV corrected in T2T-CHM13 affecting *TNNT3* and *LINC01150*, displayed by sequence comparison using miropeats (*88*) with homologous regions colored in green and blue, respectively. HG002 PacBio HiFi data are displayed showing read coverages and mappings from IGV, with allele fractions of variant sites colored (red, T; green, A; blue, C; black, G) within histograms of read depth (0 to 50). (**D** and **E**) Snapshots of regions using IGV and UCSC Genome Browser representing a collapsed duplication in GRCh38 corrected in T2T-CHM13 affecting *KCNJ18* (D) and a false duplication in GRCh38 affecting most of *KCNE1* (E). SDs depicted on top are colored by sequence similarity to paralog (gray, 90 to 98%; orange, >99%). Read mappings and variants from HG002 Illumina, PacBio HiFi, and ONT (mappings only), with homozygous (light blue) and heterozygous (dark blue) variants depicted as dashes. Colors within histograms of read depth (0–120) are the same as described in (C). Copy number estimates, displayed as colors indicated in legends, across k-merized versions of the GRCh38 and T2T-CHM13 references as well as representative examples of the SGDP individuals. (**F**) An example CDS region of *KCNJ18* (highlighted as a red box in D), with amino acids colored in alternating shades of blue and potential start codons (methionines) labeled in green using the UCSC Genome Browser codon-coloring scheme. Alignments of *KCNJ18* (blue), *KCNJ12* (orange), and *KCNJ17* (pink) along with allele counts of variants in each gene identified on GRCh38 and T2T-CHM13 are shown as bar plots (to approximate scale per variant), with examples 1 to 7 described in table S14. (**G**) Schematic depicts a benchmark for 269 challenging medically relevant genes for HG002. The number of variant-calling errors from three sequencing technologies on each reference is plotted.

Impact of T2T-CHM13 on clinical genomics

Variants of potential clinical relevance in T2T-CHM13

A deleterious variant in a reference genome can mislead the interpretation of a clinical variant identified in a patient because it may not be flagged as such using standard analysis tools. The GRCh38 reference genome is known to contain such variants that likely affect gene expression, protein structure, or protein function (*25*), although systematic efforts have sought to identify and remove these alleles (*3*). To determine the existence and location of loss-of-function variants in T2T-CHM13, we aligned the assembly to GRCh38 using dipcall (*61*) to identify and functionally annotate nucleotide differences (*62*) (Fig. 5A). This analysis identified 210 putative loss-of-function variants (defined as variants that affect protein-coding regions and predicted splice sites) affecting 189 genes, 31 of which are clinically relevant (*23*). These results are in line with work showing that the average diploid human genome contains ~450 putative loss-of-function variants affecting ~200 to 300 genes when low-coverage Illumina sequencing is applied (before stringent filtering) (*63*).

Of these 210 variants, 158 have been identified in at least one individual from the 1KGP, with most variants relatively common in human populations (median AF of 0.47), suggesting that they are functionally tolerated. The remaining variants not found in 1KGP individuals comprise larger indels, which are more difficult to identify with 1KGP Illumina data, as well as alleles that are rare or unique to CHM13. We curated the 10 variants affecting medically relevant genes and found seven that likely derived from duplicate paralogs: a 100-bp insertion also found in long reads of HG002, a stop gain in a final exon in one gnomAD sample, and an insertion in a homopolymer in a variable-number tandem repeat in *CEL*, which may be an error in the assembly. Understanding that the T2T-CHM13 assembly represents a human genome harboring potentially functional or rare variants that in turn would affect the ability to call variants at those sites, we have made available the full list of putative loss-of-function variants to aid in the interpretation of sequencing results (table S12).

T2T-CHM13 improves variant calling for medically relevant genes

We sought to understand how the transition from GRCh38 to the T2T-CHM13 reference might have an impact on variants identified in a previously compiled (*23*) set of 4964 medically relevant genes residing on human autosomes and chromosome X (representing 4924 genes in T2T-CHM13 via liftover; table S13). Of these genes, 28 mapped to previously unresolved and/or nonsyntenic regions of T2T-CHM13. We found more than twice as many medically relevant genes affected by rare or erroneous structural alleles on GRCh38 (n = 756 including 14 with no T2T-CHM13 liftover) compared to T2T-CHM13 (n = 306) (Fig. 5B), of which 622 genes appear corrected in T2T-CHM13. This includes 116 genes falling in regions previously flagged as erroneous in GRCh38 by the GRC. The majority (82%) of affected clinically relevant genes in GRCh38 overlap SVs that exist in all 13 HiFi-sequenced individuals, likely representing rare alleles or errors in the reference (see above), including 13 of the 14 genes with no T2T-CHM13 liftover.

One example of a resolved gene structure involves *TNNT3*, which encodes Troponin T3, fast skeletal type, and is implicated in forms of arthrogryposis (*64*). When calling SVs with respect to GRCh38, *TNNT3* was previously postulated to be affected by a complex structural rearrangement in all individuals, consisting of a 24-kbp inversion and 22-kbp upstream deletion, which also ablates *LINC01150* (Fig. 5C). The GRC determined that a problem existed with the GRCh38 reference in this region (GRC issue HG-28). Analysis of this region in T2T-CHM13 instead shows a complex rearrangement with the 22-kbp region upstream of *TNNT3* inversely transposed in the T2T-CHM13 assembly to the proximal side of the gene. Besides potentially affecting interpretations of gene regulation, this structural correction of the reference places *TNNT3* >20 kbp closer to its genetically linked partner *TNNI2* (*65*). Other genes have variable number tandem repeats (VNTRs) that are collapsed in GRCh38, such as one expanded by 17 kbp in most individuals in the medically relevant gene *GPI*. *MUC3A* was also flagged with a whole-gene amplification in all individuals, which we identified as residing within a falsely collapsed SD in GRCh38, further evidencing that finding (Fig. 1A).

Seventeen medically relevant genes reside within erroneous duplicated and putative collapsed regions in GRCh38 (tables S1 and S3), including *KCNE1* (false duplication) and *KCNJ18* (collapsed duplication) (Fig. 5, D and E). For these genes, we show that a significant skew in total variant density occurs in GRCh38 (58 variants per kbp for eight genes in collapsed duplications and 21 variants per kbp for seven genes in false duplications; $P = 5.684 \times 10^{-3}$ and 6.195×10^{-4}, respectively, Mann-Whitney U test) versus the rest of the 4909 medically relevant gene set (40 variants per kbp) that largely disappears in T2T-CHM13 (40 variants per kbp in collapsed duplications and 47 variants per kbp in false duplications versus 41 variants per kbp for the remaining gene set; P = 0.8778 and 0.0219, respectively) (fig. S45). Examining *KCNE1*, we found that coverage is much lower than normal on GRCh38 for short and long reads and that most variants are missed because many reads incorrectly map to a likely false duplication (*KCNE1B* on the p-arm of chromosome 21). The k-mer–based copy number of this region in all 266 SGDP genomes supports the T2T-CHM13 copy number as well as its lack of duplication in GRCh37 (*23*). As for *KCNJ18*, which resides within a GRCh38 collapsed duplication at chromosome 17p11.2 (*66*), we found increased coverage and variants within HG002 using short- and long-read sequences in GRCh38 relative to T2T-CHM13.

To verify whether the additional variants identified using GRCh38 are false heterozygous calls from PSVs derived from missing duplicate paralogs, we compared the distributions of MAFs across the 49-kbp SD. We observed a shift in SNV proportions, with a relative decrease in intermediate-frequency alleles and a relative increase in rare alleles for *KCNJ18* and *KCNJ12* (another collapsed duplication residing distally at chromosome 17p11.2) in T2T-CHM13 compared with GRCh38 ($P = 8.885 \times 10^{-2}$ and 3.102×10^{-2}, respectively; Mann-Whitney U test) (fig. S56). We matched the homologous positions of discovered alternative alleles in GRCh38 and T2T-CHM13 across the three paralogs—including the previously missing paralog located in a centromere-associated region on chromosome 17p *KCNJ17*, denoted *KCNJ18-1* in T2T-CHM13—and observed that even true variants (i.e., non-PSVs) had discordant allele counts in *KCNJ18* and *KCNJ12* between the two references (Fig. 5F and table S14). Considering that rare variants of *KCNJ18* contribute to muscle channelopathy-thyrotoxic periodic paralysis (*66*), including nine "pathogenic" or "likely pathogenic" variants in ClinVar, increased sensitivity to discover variants in patients using T2T-CHM13 would have a substantial clinical impact. In summary, the improved representation of this gene and other collapsed duplications in T2T-CHM13 not only eliminates false positives but also improves detection and genotyping of true variants.

Clinical gene benchmark demonstrates that T2T-CHM13 reduces errors across technologies

Finally, to determine how the T2T-CHM13 genome improved the ability to assay variation broadly, we used a curated diploid assembly to develop a benchmark for 269 challenging medically relevant genes in GIAB Ashkenazi son HG002 (*23*), with comparable benchmark regions on GRCh38 and T2T-CHM13. We tested three short- and long-read variant call sets against this benchmark: Illumina-BWAMEM-GATK, HiFi-PEPPER-DeepVariant, and ONT-PEPPER-DeepVariant. Counts of both false positives and false negatives substantially decreased for all three call sets when using T2T-CHM13 as a reference instead of GRCh38 (Fig. 5G and table S15). The number of false positives for HiFi decreased by a factor of 12 in these genes, primarily because of the addition of missing sequences similar to *KMT2C* (fig. S15)

and removal of false duplications of *CBS*, *CRYAA*, *H19*, and *KCNE1* (Fig. 5G). As demonstrated above, T2T-CHM13 better represents these genes and others for a diverse set of individuals, so performance should be higher across diverse ancestries. Furthermore, the number of true positives decreased by a much smaller fraction than the errors (~14%); this is due to a reduction of true homozygous variants caused by T2T-CHM13 possessing fewer ultra-rare and private alleles (Fig. 2G). This benchmarking demonstrates concrete performance gains in specific medically relevant genes resulting from the highly accurate assembly of a single genome.

Discussion

Difficult regions of the human reference genome, ranging from collapsed duplications to missing sequences, have remained unresolved for decades. The assumptions that most genomic analyses make about the correctness of the reference genome have contributed to spurious clinical findings and mistaken disease associations (*67*–*70*). Here, we identify variation in difficult-to-resolve regions and show that the T2T-CHM13 reference genome universally improves genomic analyses for all populations by correcting major structural defects and adding sequences that were absent from GRCh38. In particular, we show that the T2T-CHM13 assembly (i) revealed millions of additional variants and the existence of additional copies of medically relevant genes (e.g., *KCNJ17*) within the 240 Mbp and 189 Mbp of nonsyntenic and previously unresolved sequence, respectively; (ii) eliminated tens of thousands of spurious variants and incorrect genotypes per sample, including within medically relevant genes (e.g., *KCNJ18*) by expanding 203 loci (8.04 Mbp) that were collapsed in GRCh38; (iii) improved genotyping by eliminating 12 loci (1.2 Mbp) that were duplicated in GRCh38; and (iv) yielded more comprehensive SV calling genome-wide, with an improved insertion/deletion balance, by correcting collapsed tandem repeats. Overall, the T2T-CHM13 assembly reduced false positive and false negative SNVs from short and long reads by as much as a factor of 12 in challenging, medically relevant genes. The T2T-CHM13 reference also accurately represents the haplotype structure of human genomes, eliminating 1390 artificial recombinant haplotypes in GRCh38 that occurred as artifacts of BAC clone boundaries. These improvements will broadly enable future discoveries and refine analyses across all of human genetics and genomics.

Given these advances, we advocate for a rapid transition to the T2T-CHM13 genome as a reference. Although we appreciate that transitioning institutional databases, pipelines, and clinical knowledge from GRCh38 to T2T-CHM13 will require substantial bioinformatics and clinical effort, we provide several resources to advance this goal. On a practical level, improvements to large genomic regions, such as entire p-arms of the acrocentric chromosomes, and the discovery of clinically relevant genes and disease-causing variants justify the labor and cost required to incorporate T2T-CHM13 into basic science and clinical genomic studies. On a technical level, T2T-CHM13 simplifies genome analysis and interpretation because it consists of 23 complete linear sequences and is free of "patch," unplaced, or unlocalized sequences. Many of the corrections introduced by T2T-CHM13 were previously noted and addressed by the GRC as "fix patches," but few studies use these existing resources. The reduced contig set of T2T-CHM13 also facilitates interpretation and is directly compatible with the most commonly used analysis tools. To promote this transition, we provide variant calls and several other annotations for the T2T-CHM13 genome within the UCSC Genome Browser and the NHGRI AnVIL as a resource for the human genomics and medical communities.

Finally, our work underscores the need for additional T2T genomes. Most urgently, the CHM13 genome lacks a Y chromosome, so our analysis relied on the incomplete representation of chromosome Y from GRCh38. A T2T representation of the Y chromosome should further improve mapping and variant analysis, especially with respect to variants on the Y chromosome itself. Furthermore, many of the previously unresolved regions in T2T-CHM13 are present in all human genomes and enable variant calling with traditional methods from short and/or long reads. However, many previously unresolved regions identified in the T2T-CHM13 genome exhibit substantial variation within and between populations, including satellite DNA (*31*) and SDs that are polymorphic in copy number and structure (*32*). Relatedly, the T2T-CHM13 reference provides a basis for calling millions of variants that were previously hidden, but many of these variants are challenging to resolve accurately with current sequencing technologies and analysis algorithms. Robust variant calling in these regions will require many hundreds or thousands of diverse haplotype-resolved T2T assemblies to construct a pangenome reference, such as the effort now underway by the Human Pangenome Reference Consortium (*56*). These assemblies will then motivate further development of methods for discovering, representing, comparing, and interpreting complex variation, as well as benchmarks to evaluate their respective performances (*71*, *72*).

Through our detailed assessment of variant calling across global population samples, our study showcases T2T-CHM13 as a preeminent reference for human genetics. The annotation resources provided herein will help facilitate this transition, expanding knowledge of human genetic diversity by revealing hidden functional variation.

Methods summary

Haplotype structure

We examined the impact of the fact that GRCh38 comprises a mosaic of clones derived from multiple donor individuals on its haplotype structure. To this end, we searched for "LD-discordant" SNP pairs, defined as common (>10% MAF) SNPs that segregate in perfect LD (R^2 = 1) in the 1KGP sample, but for which GRCh38 possesses a pair of alleles that are never observed together on a single phased haplotype among 1KGP samples (i.e., alleles in perfect negative LD). We then compared these results to the same analysis applied to each 1KGP sample using a leave-one-out strategy.

Duplication errors

We flagged putatively collapsed duplications as regions >5 kbp containing clusters of heterozygous variants identified from two CHM13 datasets [simulated Illumina-like reads from T2T-CHM13 reference v1.0 including the GRCh38 Y chromosome and PacBio HiFi reads (*73*)] mapping against GRCh38 and T2T-CHM13 references. False duplications were identified as regions, converted to T2T-CHM13 coordinates, with median read-depth copy numbers (*32*) lower in kmerized GRCh38 compared to kmerized T2T-CHM13 and 88% of SGDP individuals. Alternatively, false duplications were identified as regions >3 kbp with copy numbers greater in kmerized GRCh38 compared to kmerized T2T-CHM13 and 99% of SGDP individuals using a genome-wide sliding-window approach.

Liftover of resources from GRCh38 to T2T-CHM13

Using the GATK release 4.1.9 (*74*) LiftoverVcf (Picard) tool, we lifted dbSNP build 154 (*75*), the March 8, 2021 release of Clinvar (*76*), and GWAS Catalog v1.0 (*43*) from the GRCh38 assembly to the T2T-CHM13 assembly. Initial liftover was done with default LiftoverVcf parameters. A secondary round of liftover was performed to recover variants with swapped reference and alternative alleles between GRCh38 and T2T-CHM13. We cataloged variants that failed to lift over because they overlap an indel that distinguishes T2T-CHM13 and GRCh38 based on results from dipcall.

Short-read variant calling

To evaluate short-read small-variant calling between GRCh38 and T2T-CHM13, we used the NHGRI AnVIL (*44*) to align all 3,202 1KGP samples to CHM13 with BWA-MEM (*45*) and performed variant calling with GATK HaplotypeCaller (*77*) using a workflow modeled on the one developed by the New York Genome Center (NYGC) for 1KGP analysis performed on GRCh38 (*28*). As in the NYGC analysis, we recalibrated the variant calls with GATK VariantRecalibrator. We analyzed coverage statistics using samtools and AF using bedtools. To

identify Mendelian-discordant variants, we used GATK VariantEval.

Long-read variant calling

To compare long-read mapping and large SV calling between T2T-CHM13 and GRCh38, we aligned HiFi and ONT data from 17 samples of diverse ancestry to each reference with both Winnowmap (*49*) and minimap2 (*50*) and called SVs with Sniffles (*53*). Variant calls were refined with Iris, and HiFi-derived calls from both aligners were merged with Jasmine (*54*); the resulting sets of 124,566 SVs in GRCh38 and 141,193 SVs in T2T-CHM13 to compute AFs and other cohort-level statistics. In addition, we constructed trio-level callsets for two trios—the HG002 and HG005 trios from the GIAB Consortium—to compare Mendelian discordance rates between the two references.

Concordance of variants analysis across sequencing type

To evaluate the variant calls in nonsyntenic regions, we derived concordance between variant calls generated with HiFi, ONT, and Illumina reads. For each sample, we used bcftools to filter the non-PASS variants, indels, and nonautosomal variants from each callset. We then used hap.py (*78*) to derive the precision, recall, and F1-score between each variant callset to determine how many variants are common between each pair of sets.

AF differentiation of nonsyntenic variants

Using short read–based variant calls within T2T-CHM13 nonsyntenic regions, we searched for variants with signatures of extreme AF differentiation across human populations. We performed this analysis with Ohana (*59*), a method that infers admixture components for each sample and quantifies frequency variation among the components. For outlier nonsyntenic variants with extreme patterns of AF differentiation, we used liftover to compare our results to previous results generated with 1KGP phase 3 data aligned to GRCh38 (*60*).

T2T-CHM13 dipcall and Variant Effect Predictor (VEP)

VEP (*62*) (version 102.0) was used to annotate variants generated by dipcall (*61*) when aligning the T2T-CHM13 reference genome (chm13_v1.0_plus38Y.fa) to the GRCh38 reference genome (hg38.no_alt.fa). VCF files were annotated without the –filter_common and –canonical flags. CADD (*79*) v1.6 and raw SpliceAI (*80*) scores were added using both the CADD and SpliceAI plugins. Variants were filtered based on predicted HIGH functional impact.

HG002 medically relevant genes benchmark

To evaluate variant call accuracy when using T2T-CHM13 vs. GRCh38 as a reference, we developed equivalent small variant benchmarks for GIAB sample HG002 in 269 challenging, medically relevant genes. Methods were adapted from a companion manuscript that describes a curated benchmark for these genes created by using variants generated by dipcall (*61*) when aligning a trio-based hifiasm assembly to GRCh37 and GRCh38 (*23*).

REFERENCES AND NOTES

1. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). doi: 10.1038/35057062; pmid: 11237011
2. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). doi: 10.1038/nature03001; pmid: 15496913
3. V. A. Schneider *et al.*, Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017). doi: 10.1101/gr.213611.116; pmid: 28396521
4. Z. D. Stephens *et al.*, Big Data: Astronomical or Genomical? *PLOS Biol.* **13**, e1002195 (2015). doi: 10.1371/journal.pbio.1002195; pmid: 26151137
5. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). doi: 10.1038/nature15394; pmid: 26432246
6. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi: 10.1038/s41586-020-2308-7; pmid: 32461654
7. J.-S. Seo *et al.*, De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016). doi: 10.1038/nature20098; pmid: 27706134
8. K. Shafin *et al.*, Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020). doi: 10.1038/s41587-020-0503-6; pmid: 32686750
9. J. Navarro Gonzalez *et al.*, The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021). doi: 10.1093/nar/gkaa1070; pmid: 33221922
10. J. E. Moore *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). doi: 10.1038/s41586-020-2493-4; pmid: 32728249
11. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). doi: 10.1126/science.aaz1776; pmid: 32913098
12. D. Taliun *et al.*, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021). doi: 10.1038/s41586-021-03205-y; pmid: 33568819
13. K. H. Miga *et al.*, Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014). doi: 10.1101/gr.159624.113; pmid: 24501022
14. D. M. Church *et al.*, Extending reference assembly models. *Genome Biol.* **16**, 13 (2015). doi: 10.1186/s13059-015-0587-3; pmid: 25651527
15. M. J. P. Chaisson *et al.*, Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019). doi: 10.1038/s41467-018-08148-z; pmid: 30992455
16. M. J. P. Chaisson *et al.*, Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). doi: 10.1038/nature13907; pmid: 25383537
17. P. A. Audano *et al.*, Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019). doi: 10.1016/j.cell.2018.12.019; pmid: 30661756
18. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
19. M. Yandell *et al.*, A probabilistic disease-gene finder for personal genomes. *Genome Res.* **21**, 1529–1542 (2011). doi: 10.1101/gr.123158.111; pmid: 21700766
20. M. Kircher *et al.*, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014). doi: 10.1038/ng.2892; pmid: 24487276
21. B. Gulko, M. J. Hubisz, I. Gronau, A. Siepel, A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015). doi: 10.1038/ng.3196; pmid: 25599402
22. C. A. Miller *et al.*, Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. *J. Mol. Diagn.* **24**, 219–223 (2022). doi: 10.1016/j.jmoldx.2021.10.013
23. J. Wagner *et al.*, Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* 1–9, (2022). doi: 10.1038/s41587-021-01158-1
24. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010). doi: 10.1126/science.1188021; pmid: 20448178
25. S. Ballouz, A. Dobin, J. A. Gillis, Is it time to change the reference genome? *Genome Biol.* **20**, 159 (2019). doi: 10.1186/s13059-019-1774-4; pmid: 31399121
26. D. R. Zerbino, A. Frankish, P. Flicek, Progress, Challenges, and Surprises in Annotating the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **21**, 55–79 (2020). doi: 10.1146/annurev-genom-121119-083418; pmid: 32421357
27. S. Nurk *et al.*, The complete sequence of a human genome. *Science* **376**, 44–53 (2022). doi: 10.1126/science.abj6987
28. M. Byrska-Bishop *et al.*, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv 430068 (2021). doi: 10.1101/2021.02.06.430068
29. J. M. Zook *et al.*, A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020). doi: 10.1038/s41587-020-0538-8; pmid: 32541955
30. A. M. Mc Cartney *et al.*, Chasing perfection: Validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* 10.1038/s41592-022-01440-3 (2022). doi: 10.1038/s41592-022-01440-3
31. N. Altemose *et al.*, Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022). doi: 10.1126/science.abl4178
32. M. R. Vollger *et al.*, Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022). doi: 10.1126/science.abj6965
33. S. J. Hoyt *et al.*, From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022). doi: 10.1126/science.abk3112
34. See supplementary materials.
35. K. M. Steinberg *et al.*, Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* **24**, 2066–2076 (2014). doi: 10.1101/gr.180893.114; pmid: 25373144
36. J. Huddleston *et al.*, Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688–696 (2014). doi: 10.1101/gr.168450.113; pmid: 24418700
37. M. Y. Dennis *et al.*, The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* **1**, 69 (2017). doi: 10.1038/s41559-016-0069; pmid: 28580430
38. M. O'Bleness *et al.*, Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics* **15**, 387 (2014). doi: 10.1186/1471-2164-15-387; pmid: 24885025
39. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021). doi: 10.1038/s41592-020-01056-5; pmid: 33526886
40. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi: 10.1038/nature18964; pmid: 27654912
41. J.-P. Li *et al.*, The phosphatase JKAP/DUSP22 inhibits T-cell receptor signalling and autoimmunity by inactivating Lck. *Nat. Commun.* **5**, 3618 (2014). doi: 10.1038/ncomms4618; pmid: 24714587
42. OMIM Entry #617768, KLEEFSTRA SYNDROME 2; KLEFS2; www.omim.org/entry/617768.
43. A. Buniello *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019). doi: 10.1093/nar/gky1120; pmid: 30445434
44. M. C. Schatz *et al.*, Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genomics* **2**, 100085 (2022). doi: 10.1016/j.xgen.2021.100085
45. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997 (2013).
46. R. Poplin *et al.*, Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 201178 (2021). doi: 10.1101/201178
47. I. Iossifov *et al.*, The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014). doi: 10.1038/nature13908; pmid: 25363768
48. L. B. Alexandrov *et al.*, Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013). doi: 10.1038/nature12477; pmid: 23945592

49. C. Jain *et al*., Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36** (suppl. 1), i111–i118 (2020). doi: 10.1093/bioinformatics/btaa435; pmid: 32657365
50. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi: 10.1093/bioinformatics/bty191; pmid: 29750242
51. R. M. Sherman *et al*., Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019). doi: 10.1038/s41588-018-0273-y; pmid: 30455414
52. D. Beyter *et al*., Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021). doi: 10.1038/s41588-021-00865-4; pmid: 33972781
53. F. J. Sedlazeck *et al*., Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018). doi: 10.1038/s41592-018-0001-7; pmid: 29713083
54. M. Kirsche *et al*., Jasmine: Population-scale structural variant comparison and analysis. bioRxiv 445886 (2021). doi: 10.1101/2021.05.27.445886
55. K. Shafin *et al*., Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021). doi: 10.1038/s41592-021-01299-w
56. K. H. Miga, T. Wang, The Need for a Human Pangenome Reference Sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021). doi: 10.1146/annurev-genom-120120-081921; pmid: 33929893
57. D. A. Hartasánchez, M. Brasó-Vives, J. M. Heredia-Genestar, M. Pybus, A. Navarro, Effect of Collapsed Duplications on Diversity Estimates: What to Expect. *Genome Biol. Evol.* **10**, 2899–2905 (2018). doi: 10.1093/gbe/evy223; pmid: 30364947
58. M. P. Ball *et al*., A public resource facilitating clinical use of genomes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 11920–11927 (2012). doi: 10.1073/pnas.1201904109; pmid: 22797899
59. J. Y. Cheng, A. J. Stern, F. Racimo, R. Nielsen, Detecting selection in multiple populations by modeling ancestral admixture components. *Mol. Biol. Evol.* **39**, msab294 (2022). doi: 10.1093/molbev/msab294; pmid: 34626111
60. S. M. Yan *et al*., Local adaptation and archaic introgression shape global diversity at human structural variant loci. *eLife* **10**, e67615 (2021). doi: 10.7554/eLife.67615; pmid: 34528508
61. H. Li *et al*., A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018). doi: 10.1038/s41592-018-0054-7; pmid: 30013044
62. W. McLaren *et al*., The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016). doi: 10.1186/s13059-016-0974-4; pmid: 27268795
63. D. G. MacArthur *et al*., A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012). doi: 10.1126/science.1215040; pmid: 22344438
64. S. S. Sung *et al*., Mutations in TNNT3 cause multiple congenital contractures: A second locus for distal arthrogryposis type 2B. *Am. J. Hum. Genet.* **73**, 212–214 (2003). doi: 10.1086/376418; pmid: 12865991
65. J.-J. Sheng, J.-P. Jin, TNNI1, TNNI2 and TNNI3: Evolution, regulation, and protein structure-function relationships. *Gene* **576**, 385–394 (2016). doi: 10.1016/j.gene.2015.10.052; pmid: 26526134
66. D. P. Ryan *et al*., Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell* **140**, 88–98 (2010). doi: 10.1016/j.cell.2009.12.024; pmid: 20074522
67. N. Khalilipour *et al*., Familial Esophageal Squamous Cell Carcinoma with damaging rare/germline mutations in KCNJ12/KCNJ18 and GPRIN2 genes. *Cancer Genet.* **221**, 46–52 (2018). doi: 10.1016/j.cancergen.2017.11.011; pmid: 29405996
68. S. Munchel *et al*., Targeted or whole genome sequencing of formalin fixed tissue samples: Potential applications in cancer genomics. *Oncotarget* **6**, 25943–25961 (2015). doi: 10.18632/oncotarget.4671; pmid: 26305677
69. K. Gürünlüoğlu *et al*., Whole exome sequencing analysis for mutations in isolated type III biliary atresia patients. *Clin. Exp. Hepatol.* **6**, 347–353 (2020). doi: 10.5114/ceh.2020.102156; pmid: 33511283
70. F. Lalrohlui, J. Zohmingthanga, V. Hruaii, A. Vanlallawma, N. Senthil Kumar, Whole exome sequencing identifies the novel putative gene variants related with type 2 diabetes in Mizo population, northeast India. *Gene* **769**, 145229 (2021). doi: 10.1016/j.gene.2020.145229; pmid: 33059026
71. J. M. Eizenga *et al*., Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020). doi: 10.1146/annurev-genom-120219-080406; pmid: 32453966
72. J. Pritt, N.-C. Chen, B. Langmead, FORGe: Prioritizing variants for graph genomes. *Genome Biol.* **19**, 220 (2018). doi: 10.1186/s13059-018-1595-x; pmid: 30558649
73. M. R. Vollger *et al*., Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* **84**, 125–140 (2020). doi: 10.1111/ahg.12364; pmid: 31711268
74. G. A. Van der Auwera, B. D. O'Connor, *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media Inc., 2020); https://play.google.com/store/books/details?id=vsXaDwAAQBAJ.
75. S. T. Sherry, M. Ward, K. Sirotkin, dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999). doi: 10.1101/gr.9.8.677; pmid: 10447503
76. M. J. Landrum *et al*., ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018). doi: 10.1093/nar/gkx1153; pmid: 29165669
77. G. A. Van der Auwera *et al*., From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).
78. P. Krusche *et al*., Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019). doi: 10.1038/s41587-019-0054-x; pmid: 30858580
79. P. Rentzsch, M. Schubach, J. Shendure, M. Kircher, CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021). doi: 10.1186/s13073-021-00835-9; pmid: 33618777
80. K. Jaganathan *et al*., Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019). doi: 10.1016/j.cell.2018.12.015; pmid: 30661751
81. M. Sauria, msauria/T2T_MUK_Analysis: T2T_resubmission (2021); https://zenodo.org/record/5596590.
82. R. McCoy, D. Taylor, S. Yan, mccoy-lab/t2t-variants: First release (2021); https://zenodo.org/record/5591054.
83. D. C. Soto, mydennislab/t2t-variants: T2T-variants (2021); https://zenodo.org/record/5595398.
84. M. Schatz, S. Zarate, S. Aganezov, schatzlab/t2t-variants: T2TVariants1.0 (2021); https://zenodo.org/record/5598342.
85. M. Kirsche, Jasmine: Population-scale structural variant merging (2021); https://zenodo.org/record/5586905.
86. M. Kirsche, Iris: Structural variant breakpoint and sequence refinement (2021); https://zenodo.org/record/5586965.
87. J. Wagner, N. D. Olson, J. McDaniel, J. M. Zook, Challenging medically-relevant genes benchmark set (NIST Public Data Repository, 2021). doi: 10.18434/MDS2-2475
88. J. D. Parsons, Miropeats: Graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995). pmid: 8808577

ACKNOWLEDGMENTS

We thank M. Zody, B. Grüning, H. Li, S. Langley, C. Langley, S.-H. Kim, G. Van der Auwera, V. Schneider, S. Salzberg, B. Langmead, A. Battle, and several of their lab members for helpful discussions. Certain commercial equipment, instruments, or materials are identified to specify adequate experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov) and the Maryland Advanced Research Computing Center (www.marcc.jhu.edu/). **Funding:** Supported by NIH grants R35GM133747 (R.C.M.), DP2MH119424 (M.Y.D.), R01HG006677 (A.S.), R01HG010485 (B.P.), U41HG010972 (B.P.), U01HG010961 (B.P., U24HG011853 (B.P.), OT2OD026682 (B.P.), UM1HG008898 (F.J.S.), 1R01HG011274-01 (K.H.M.), 1R21HG010548-01 (K.H.M.), 1U01HG010971 (K.H.M.), U24HG010263 (M.C.S.), U24HG006620 (M.C.S.), U01CA253481 (M.C.S.), R24DK106766-01A1 (M.C.S.), and R00HG009532 (R.M.L.); NSF grants DBI-1627442, IOS-1732253, and IOS-1758800 (M.C.S.); the Intramural Research Program of the National Human Genome Research Institute (P.A., A.R., S.K., A.M.P.); the Intramural Program of the National Institute of Standards and Technology (N.D.O., J.M., J.W., J.M.Z.); the Intramural Research Program of the National Library of Medicine (C.X.); the Mark Foundation for Cancer Research (19-033-ASP) (M.C.S.); and a Fulbright fellowship (D.C.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. **Author contributions:** Annotation and comparison of GRCh38 and T2T-CHM13, A.S., A.R., D.C.S., D.E.M., D.J.T., S.M.Y., M.E.G.S., M.R.V., N.F.H., M.Y.D., J.M.Z., R.C.M.; short-read analysis, S.Z., S.A., S.M.Y., R.C.M., M.C.S.; long-read analysis, K.S., M.K., S.A., J.Z., M.C.S.; population genetics, S.A., D.J.T., S.M.Y., M.C.S., R.C.M.; clinical variant analysis, D.C.S., D.E.M., M.Y.D., J.M.Z.; variant analysis, A.M.P., B.P., C.S.C., C.X., D.C.S., F.J.S., J.L., J.M., J.M.Z., J.W., M.M., M.C.S., M.Y.D., N.D.O., P.A., R.C.M., R.L., S.A., S.K., S.M.; project design, A.M.P., B.P., F.J.S., J.A.R., J.M.Z., K.H.M., M.Y.D., R.C.M., M.C.S., R.L.; writing, R.C.M., M.Y.D., J.M.Z., and M.C.S. with input from all of the authors. **Competing interests:** C.S.C. is an employee of DNAnexus. J.L. is a former employee and shareholder of Bionano Genomics. S.A. is an employee of Oxford Nanopore Technologies. F.J.S. has received travel funds and spoken at PacBio and Oxford Nanopore Technologies events. K.H.M., S.K., and D.E.M. have received travel funds to speak at symposia organized by Oxford Nanopore Technologies. K.H.M. is a SAB member of Centaura Inc. **Data and materials availability:** Sequencing data: 1KGP short-read datasets available within the ENA under projects PRJEB31736 and PRJEB36890. Long-read data accessions presented in table S7. Variant calling workflows and downstream analysis scripts: Mappability analysis, https://github.com/msauria/T2T_MUK_Analysis and published in Zenodo (*81*). Local ancestry and genome masks, https://github.com/mccoy-lab/t2t-variants and published on Zenodo (*82*). Collapsed duplications and clinical genomics, https://github.com/mydennislab/t2t-variants and published on Zenodo (*83*). Short-read variant calling, https://github.com/schatzlab/t2t-variants and published on Zenodo (*84*). Jasmine v1.1.0 for long-read SV comparison, https://github.com/mkirsche/Jasmine and published on Zenodo (*85*). Iris v1.0.4 for insertion SV polishing, https://github.com/mkirsche/Iris and published on Zenodo (*86*). Aligned reads, variant calls, and other summarizations available within the NHGRI AnVIL Platform, along with notebooks and workflows for computing the analysis, https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T. UCSC assembly hub browser with variant calls and genome annotations, http://genome.ucsc.edu/cgi-bin/hgTracks?genome=t2t-chm13-v1.0&hubUrl=http://t2t.gi.ucsc.edu/chm13/hub/hub.txt. Mirror of short- and long-read variant calls, https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/. Genome in a Bottle Challenging Medically-Relevant Gene Benchmarks, https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/ and published at NCBI (*87*). Stratification bed files for collapsed duplications, false duplications, and local ancestry in GRCh38 are available for stratifying GIAB/GA4GH benchmarking results at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl3533
Materials and Methods
Figs. S1 to S56
Tables S1 to S16
References (*89*–*115*)
MDAR Reproducibility Checklist

7 July 2021; accepted 3 February 2022
10.1126/science.abl3533



Pushing the Boundaries of Knowledge

As AAAS's first multidisciplinary, open access journal, *Science Advances* publishes research that reflects the selectivity of high impact, innovative research you expect from the *Science* family of journals, published in an open access format to serve a vast and growing global audience. Check out the latest findings or learn how to submit your research: **ScienceAdvances.org**

Science Advances
AAAS

GOLD OPEN ACCESS, DIGITAL, AND FREE TO ALL READERS

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

Segmental duplications and their variation in a complete human genome

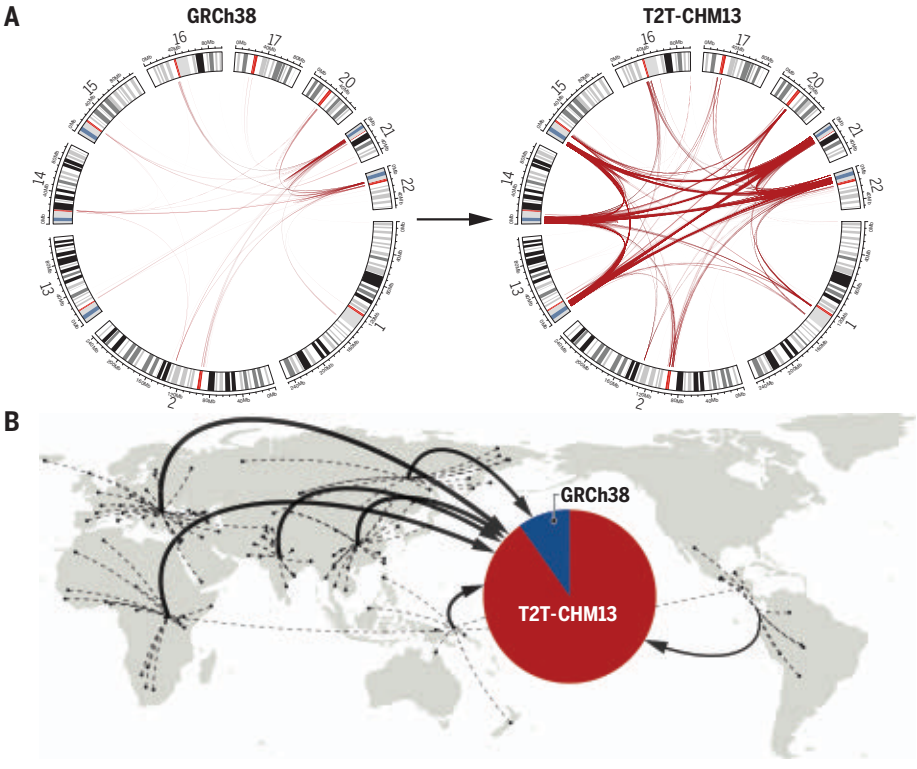
Mitchell R. Vollger, Xavi Guitart, Philip C. Dishuck, Ludovica Mercuri, William T. Harvey, Ariel Gershman, Mark Diekhans, Arvis Sulovari, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, David Porubsky, Ruiyang Li, Sergey Nurk, Sergey Koren, Karen H. Miga, Adam M. Phillippy, Winston Timp, Mario Ventura, Evan E. Eichler*

INTRODUCTION: Large, high-identity duplicated sequences—termed segmental duplications (SDs)—are frequently the last regions of genomes to be sequenced and assembled. While the human reference genome provided a roadmap of the SD landscape, >50% of the remaining gaps correspond to regions of complex SDs.

RATIONALE: SDs are major sources of evolutionary gene innovations and contribute disproportionately to genetic variation within and between ape species. With the complete human genome (T2T-CHM13), researchers have the potential to identify genes and uncover patterns of human genetic variation.

RESULTS: We identified 51 million base pairs (Mbp) of additional human SD in T2T-CHM13 and now estimate that 7% of the human genome consists of SDs [(218 Mbp of 3.1 billion base pairs (Gbp)]. SDs make up two-thirds (45.1 of 68.1 Mbp) of acrocentric short arms, and these SDs are the largest in the human genome (see the figure, panel A). Additionally, 54% of acrocentric SDs are copy number variable or map to different chromosomes among the six individuals examined. A detailed comparison between the current reference genome (GRCh38) and T2T-CHM13 for SD content identifies 81 Mbp of previously unresolved or structurally variable SDs. Short-read whole-genome sequence data from a diversity panel of 268 humans show that human copy number is nine times (59.26 versus 6.55 Mbp) more likely to match T2T-CHM13 rather than GRCh38, including 119 protein-coding genes (see the figure, panel B). Using long-read–sequencing data from 25 human haplotypes, we investigated patterns of human genetic variation identifying significant increases in structural and single-nucleotide diversity. We identified gene-rich regions (e.g., *TBC1D3*) that vary by hundreds of kilo–base pairs and gene copy number between individuals showing some of the highest genome-wide structural heterozygosity (85 to 90%). Our analysis identified 182 candidate protein-coding genes as well as the complete sequence for structurally variable gene models that were previously unresolved. Among these is the complete gene structure of lipoprotein A (*LPA*), including the expanded kringle IV repeat domain. Reduced copies of this domain are among the strongest genetic associations with cardiovascular disease, especially among African Americans, and sequencing of multiple human haplotypes identified not only copy number variation but also other forms of rare coding variation potentially relevant to disease risk. Finally, we compared global methylation and expression patterns between duplicated and unique genes. Transcriptionally inactive duplicate genes are more likely to map to hypomethylated genomic regions; however, specifically over the transcription start site we observe an increase in methylation, suggesting that as many as two-thirds of duplicated genes are epigenetically silenced. Additionally, SD genes show a high degree of concordance between methylation profiles and transcription levels, allowing us to define the actively transcribed members of high-identity gene families that are otherwise indistinguishable by coding sequence.

CONCLUSION: A complete human genome provides a more comprehensive understanding of the organization, expression, and regulation of duplicated genes. Our analysis reveals underappreciated patterns of human genetic diversity and suggests characteristic features of methylation and gene regulation. This resource will serve as a critical baseline for improved gene annotation, genotyping, and previously unknown associations for some of the most dynamic regions of our genome. ■



More-complete segmental duplication content improves genotyping. (**A**) Increase (by a factor of 10) in the number of large (>10 kilo–base pairs) acrocentric segmental duplications (red) in T2T-CHM13 (right) compared with GRCh38 (left). (**B**) Copy number genotyping based on read-depth from 268 diverse human genomes across the globe shows that 90% of new SDs in T2T-CHM13 (red) are more likely to reflect human copy number when compared to GRCh38 (blue) irrespective of human population group considered.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: eee@gs.washington.edu
Cite this article as M. R. Vollger *et al.*, *Science* **376**, eabj6965 (2022). DOI: 10.1126/science.abj6965

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abj6965

RESEARCH ARTICLE

HUMAN GENOMICS

Segmental duplications and their variation in a complete human genome

Mitchell R. Vollger[1], Xavi Guitart[1], Philip C. Dishuck[1], Ludovica Mercuri[2], William T. Harvey[1], Ariel Gershman[3], Mark Diekhans[4], Arvis Sulovari[1], Katherine M. Munson[1], Alexandra P. Lewis[1], Kendra Hoekzema[1], David Porubsky[1], Ruiyang Li[1], Sergey Nurk[5], Sergey Koren[5], Karen H. Miga[4], Adam M. Phillippy[5], Winston Timp[3], Mario Ventura[2], Evan E. Eichler[1,6]*

Despite their importance in disease and evolution, highly identical segmental duplications (SDs) are among the last regions of the human reference genome (GRCh38) to be fully sequenced. Using a complete telomere-to-telomere human genome (T2T-CHM13), we present a comprehensive view of human SD organization. SDs account for nearly one-third of the additional sequence, increasing the genome-wide estimate from 5.4 to 7.0% [218 million base pairs (Mbp)]. An analysis of 268 human genomes shows that 91% of the previously unresolved T2T-CHM13 SD sequence (68.3 Mbp) better represents human copy number variation. Comparing long-read assemblies from human (n = 12) and nonhuman primate (n = 5) genomes, we systematically reconstruct the evolution and structural haplotype diversity of biomedically relevant and duplicated genes. This analysis reveals patterns of structural heterozygosity and evolutionary differences in SD organization between humans and other primates.

Genomic duplications have long been recognized as important sources of structural change and gene innovation (*1*, *2*). In humans, the most recent and highly identical sequences [>90% and >1 kilo–base pairs (kbp)]—referred to as segmental duplications (SDs) (*3*)—promote meiotic unequal crossover events that contribute to recurrent rearrangements associated with ~5% of developmental delay and autism (*4*). These same SDs are reservoirs for human-specific genes that have been important in increasing synaptic density and the expansion of the frontal cortex since humans diverged from other ape lineages (*5*–*8*). SDs are also enriched ~10-fold for normal copy number variation, although most of this genetic diversity has yet to be fully characterized or associated with human phenotypes (*9*, *10*). SD length (frequently >100 kbp), sequence identity, and extensive structural diversity among human haplotypes have hampered our ability to characterize these regions at a genomic level. This is because sequence reads have been insufficiently long and human haplotypes too structurally diverse to resolve duplicate copies or distinguish allelic variants.

One of the first human whole–genome sequence (WGS) assembly drafts created with Sanger sequencing technology was almost devoid of SDs and their underlying genes (*11*, *12*). Similarly, bacterial artificial chromosome (BAC)-based approaches to assembling the human genome from different haplotypes led to many misjoins, creating de facto gaps that took years to resolve (*13*). Although combining WGS- and BAC-based data for early sequencing of human genomes provided a roadmap of the SD landscape (*14*), more than 50% of the gaps within the human reference genome have corresponded to regions of complex SDs.

The development of genomic resources (*15*–*17*), including BAC libraries and long-read sequence data from complete hydatidiform moles (CHM, which represent a single human haplotype), was motivated in large part by efforts to resolve the organization of these regions and concomitantly complete the human reference genome. The CHM13 cell line has the advantage of originating from a single haplotype and predominantly a single ancestral group (European) (*18*) in contrast to the GRCh38 reference, which is a composite representation of multiple human haplotypes and ancestries (*19*). These resources, combined with advances in long-read technologies, have produced the gapless human genome assembly T2T-CHM13 (*20*). We use this genome assembly to present a complete view of SDs in a human genome and highlight their importance in advancing our understanding of genetic diversity, evolution, and disease in humans.

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Department of Biology, University of Bari, Aldo Moro, Bari 70125, Italy. [3]Department of Molecular Biology and Genetics, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [5]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.
*Corresponding author. Email: eee@gs.washington.edu

SD content and organization

We characterized the SD content of the T2T-CHM13 v1.0 assembly by sequence read-depth and pairwise sequence alignments (>90% and >1 kbp) (*21*). Our analysis of the assembly identifies 208 Mbp of nonredundant segmentally duplicated sequences within chromosome-level scaffolds (including 15.6 Mbp of SD located on chrY, which is included from GRCh38), compared with just 167 Mbp in the current reference (GRCh38) (Table 1 and Fig. 1). This raises the percent estimate of the human genome that is segmentally duplicated from 5.4 to 6.7%. However, five SD-related gaps remained in the initial assembly of the female CHM13 genome (T2T-CHM13 v1.0). Each corresponded to a cluster of tandemly repeated ribosomal DNA (rDNA) genes on each acrocentric chromosome where we confirmed long-read sequence pileups consistent with unresolved SDs. The estimated amount of missing rDNA sequence was calculated by Nurk *et al.* by using both digital droplet PCR (*22*) and a whole-genome Illumina coverage analysis (*20*). Assuming a canonical repeat length of 45 kbp for the rDNA molecule (*23*, *24*), the total amount of missing sequences was approximated at ~10 Mbp and ~200 copies of unresolved rDNA sequence (*20*). These findings are consistent with the subsequent specialized assembly of the rDNA released as part of the T2T-CHM13 v1.1 assembly. Including this estimate, the overall SD content of the human genome is now 7.0% (6.7% not including rDNA; see Table 1 for statistics breakdown by SD type) and is likely to increase as more complete genomes of diverse origins are sequenced and assembled.

One-third (81.3 Mbp) (*25*) of SD sequence in T2T-CHM13 is wholly uncharacterized in GRCh38 (16.5 Mbp) or differs in copy number and structure (64.8 Mbp) (*25*). Most of these involve large, high-identity SDs. For example, there is a 70% increase (41,285 versus 24,280) in the number of SD pairs and a doubling of the number of bases in pairwise alignments with >95% identity (Fig. 1C). Among these previously unresolved or variable SDs, 13,258 (35.0 Mbp) map to the acrocentric short arms of chromosomes 13, 14, 15, 21, and 22 (Fig. 1B and Table 1), which are unassembled in the GRCh38. These SDs do not correspond to rDNA duplications but represent other segments predominantly shared among acrocentric (n = 5332 alignments) and nonacrocentric chromosomes (n = 5500 alignments; table S1). In particular, the pericentromeric regions of chromosomes 1, 3, 4, 7, 9, 16, and 20 show the most extensive SD homology with acrocentric DNA (Fig. 1B). Non-rDNA acrocentric SDs are 1.74 times as long as all other SDs (N50: 74,704 versus 42,842) and significantly longer

Table 1. Summary statistics of segmental duplications in T2T-CHM13 and GRCh38. Mbp, number of nonredundant Mbp of SD; peri, within 5 Mbp of the heterochromatin surrounding the centromere; telo, within 500 kbp of the telomere; acro, within the short arms of the acrocentric chromosomes. Difference: SD content difference between T2T-CHM13 v1.0 and GRCh38. Previously unresolved or structurally variable: Sequence in T2T-CHM13 that does not have 1 Mbp of synteny with GRCh38. GRCh38 contains 149,690,719 bp of gap sequence included in the reported number of Gbp.

Assembly	Gbp	% SD	SD (Mbp)	# SDs	inter (Mbp)	# inter	intra (Mbp)	# intra	acro (Mbp)	# acro	peri (Mbp)	# peri	telo (Mbp)	# telo
T2T-CHM13 v1.0*	3.114	6.665	207.563	41289	121.113	30484	142.958	10805	35.106	13264	88.606	24985	10.975	4998
GRCh38	3.114	5.372	167.297	24280	83.556	16348	120.710	7932	6.624	1407	53.944	10606	8.926	1529
Difference	0.000	1.293	40.266	17009	37.556	14136	22.248	2873	28.482	11857	34.662	14379	2.049	3469
Previously unresolved or structurally variable	0.240	33.885	81.338	25161	61.873	20579	54.932	4582	35.039	13258	54.037	19607	5.616	4005
T2T-CHM13 v1.0* + rDNA estimate	3.114	6.987	217.598	66042	131.148	49213	152.993	16829	45.141	38017	98.641	49738	10.975	4998

*The version of T2T-CHM13 that was used (v1.0) included chrY from GRCh38.

(P value < 0.01, one-sided Wilcoxon rank-sum test) than any other defined SD category in the human genome (intrachromosomal, interchromosomal, pericentromeric, and telomeric; fig. S1).

We annotated all T2T-CHM13 SDs with DupMasker (*26*), which defines ancestral evolutionary units of duplication on the basis of mammalian outgroups and a repeat graph (*27*). Focusing on duplicons that carry genes or duplicated portions of genes, we identified 30 duplicons that show the greatest copy number change between T2T-CHM13 and GRCh38. These 30 genic SDs represent regions where gene annotation is most likely to change; all predicted differences favor an increase in copy number for the T2T-CHM13 assembly (Fig. 1D and table S2).

We also compared the number of SDs more directly by defining syntenic regions (5 Mbp) between GRCh38 and T2T-CHM13 (*25*). Of the 15 windows with the largest increase, nine mapped to the acrocentric short arms and six were in pericentromeric regions (fig. S1 and table S3). In particular, the intervals between the centromeric satellite and secondary constrictions (qh regions) on chromosomes 1, 9, and 16 show a 4.6-fold increase in the number of SDs (5254 versus 1141) and the most differences in organization compared with GRCh38. SDs in these regions are almost exclusively interchromosomal and depleted for intrachromosomal duplications (figs. S2 and S3).

Validation and heteromorphic variation

Because the acrocentric short arms as well as the qh regions on chromosomes 1, 9, and 16 were either previously unresolved or showed the most considerable differences in terms of SD content, we focused first on validating their organization. We mapped available end-sequence data from a human fosmid genome library (*28*) to the T2T-CHM13 assembly and selected nine distinct clones as probes (Fig. 2A) to confirm the patterns of high-identity (>95%) SDs (*25*). All 30 of the distinct duplication predictions from the T2T-CHM13 SDs were corroborated by fluorescence in situ hybridization (FISH) against chromosomal metaphases of the CHM13 cell line (Fig. 2, B and C, and table S4).

FISH also revealed nine additional signals not originally predicted by our SD analysis (fig. S4). However, we were able to identify lower identity duplications that confirmed seven of these sites, leading to an overall concordance of 95% (37 of 39) between FISH and the T2T-CHM13 SD assembly content. We extended this analysis to five additional human cell lines of diploid origin because both pericentromeric and acrocentric portions of chromosomes have been shown to be cytogenetically heteromorphic (*29*–*31*). In total, we identified 61 distinct cytogenetic locations of which 28 (46%) were fixed whereas 33 (54%) were variable in their presence or absence on specific homologs (both acrocentric and pericentromeric regions of the human genome) (fig. S4). Of the 61 FISH signals, all but six were observed in more than one of the six human cell lines indicating that such heteromorphic variation is common and prevalent.

We found a correlation (Pearson's correlation coefficient, r = 0.96) between genome-wide copy number variation from the assembly and Illumina read-depth data generated from the same CHM13 source (*25*). Because SDs frequently map to the breakpoints of inversion polymorphisms (*28*, *32*, *33*), we validated 65 inversions relative to GRCh38 with single-cell DNA template strand sequencing (Strand-seq analysis) of the T2T-CHM13 assembly (figs. S5 and S6) (*25*). Although 32 of these represent known human polymorphisms, 33 have not been observed in six previously analyzed human genomes (*32*). However, by analysis of Strand-seq data from one additional human haplotype (CHM1), we further confirmed 30 of these inversions (i.e., present in CHM1 and CHM13), suggesting that at least 95.4% (62 of 65) represent true large-scale human inversion polymorphisms (fig. S5). Consistent with previous literature (*34*), inversions associated with SDs (n = 30) are significantly longer than those not associated with SDs (P value < 0.01, one-sided Wilcoxon rank-sum test) and are polymorphic among humans (fig. S6). One notable example is an inversion polymorphism mapping to chromosome 1q21. It is a complex event consisting of two inversions (262.3 kbp and 2.26 Mbp) originally predicted by Sanders and colleagues (*33*) but our sequence analysis shows a relocation of 767.6 kbp of genic sequences (Fig. 2D). The large inversion (chr1:146,350,000 to 148,610,000) is flanked by the core duplicon—the *NBPF* gene family—and in combination with the other rearrangements changes the order of human-specific genes *NOTCH2NLA*, *-B, and -C*, which have been implicated in expansion of the frontal cortex (*8*, *35*). As a final test, we resolved this region in eight additional human haplotypes (*25*), all of which support the T2T-CHM13 configuration with one exception (CHM1), which is consistent with the GRCh38 configuration (fig. S7).

Single-nucleotide and copy number variation within SDs

The high quality and single haplotype nature of both the T2T-CHM13 and GRCh38 reference genomes provides an opportunity to compare the genome-wide pattern of single-nucleotide

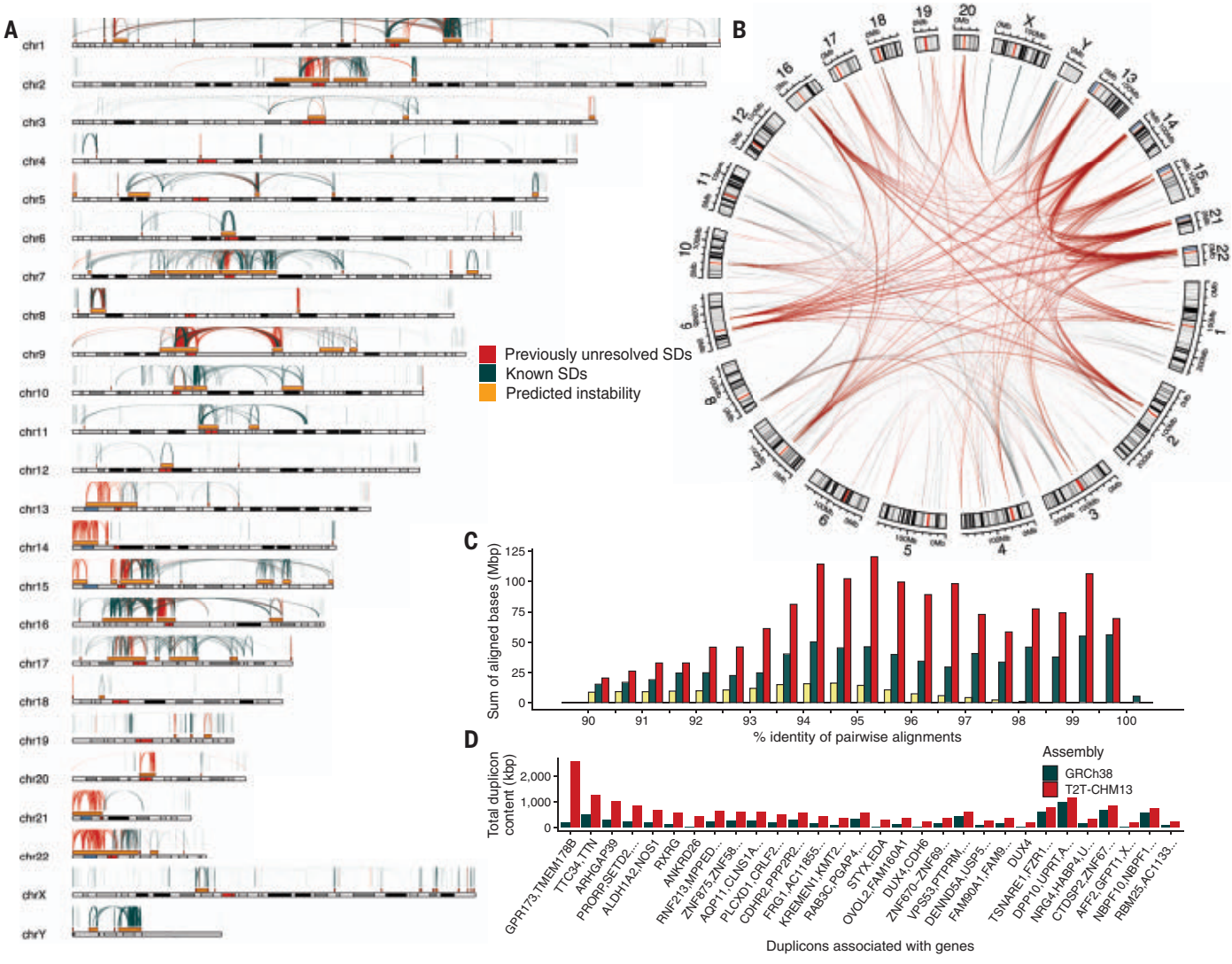


Fig. 1. Segmental duplication (SD) content of the T2T-CHM13 genome. (**A**) The pattern of previously unresolved or structurally variant intrachromosomal duplications in T2T-CHM13 (red) compared with known duplications in GRCh38 (blue-gray). These predict hotspots of genomic instability (gold) flanked by large (>10 kbp), high-identity (>95%), interspersed (>50 kbp) SDs. (**B**) Circos plot highlighting previously unresolved interchromosomal SDs (red) and showing the preponderance of previously unresolved SDs mapping to pericentromeric and acrocentric regions. (**C**) A histogram comparing SD content in different human reference genomes, including the sum of bases in pairwise SD alignments stratified by their percent identity for the celera (yellow, Sanger-based); GRCh38 (blue-gray, BAC-based); and T2T-CHM13 (red, long-read) assemblies. (**D**) The 30 genic duplicons (ancestral repeat units) with the greatest copy number difference between GRCh38 and T2T-CHM13 as determined by DupMasker (table S2). The 30 largest differences all exhibit increase in T2T-CHM13.

variation in regions that have typically been excluded from most analyses because of their repetitive nature. We aligned GRCh38 to T2T-CHM13 and retained only regions deemed to be "syntenic" on the basis of an unambiguous one-to-one correspondence between both reference genomes and at least 1 Mbp of aligned sequence (*25*).

Most unique regions of the genome (2693 Mbp) could be compared, whereas only 60% (124 Mbp) of the SDs within T2T-CHM13 exhibited a clear orthologous relationship between the two human reference genomes. As expected, the X chromosome and the region corresponding to the major histocompatibility complex (MHC) are the least and most divergent, respectively (Fig. 3A), as a result of the slower rate of evolution for the female X and the deep coalescence of MHC.

Notably, SD sequences are significantly more diverged than unique sequences (P value < 0.001, one-sided Mann-Whitney U test) (fig. S8). Comparing only syntenic regions (*25*) between GRCh38 and T2T-CHM13, we estimate the single-nucleotide variant (SNV) density to be 0.95 SNVs/kbp for unique regions of the genome when compared with SD regions where density rises to 1.47 SNVs per kbp (table S5). This 50% increase could be a result of an increased mutation rate of SDs (e.g., a result of the action of interlocus gene conversion), or a deeper average coalescence of duplicated sequences. Another possible explanation for this observation is erroneous alignment of paralogous instead of allelic sequences; however, we believe this is unlikely given the requirement of at least 1 Mbp of continuous, one-to-one, best alignment between GRCh38 and T2T-CHM13 (*25*).

As part of this analysis, we also identified regions that structurally differ or are absent from GRCh38 when compared with the T2T-CHM13 assembly. Using 1-Mbp LASTZ alignments (*25*), we identified 126 nonsyntenic regions for a total of 240 Mbp (N50 length of 12.7 Mbp; fig. S9). Of these, 33.9% (81.34 of 240 Mbp) overlapped SD regions. Using sequence read depth (*25*) from 268 human genomes [(Simons Genome Diversity Project (SGDP)],

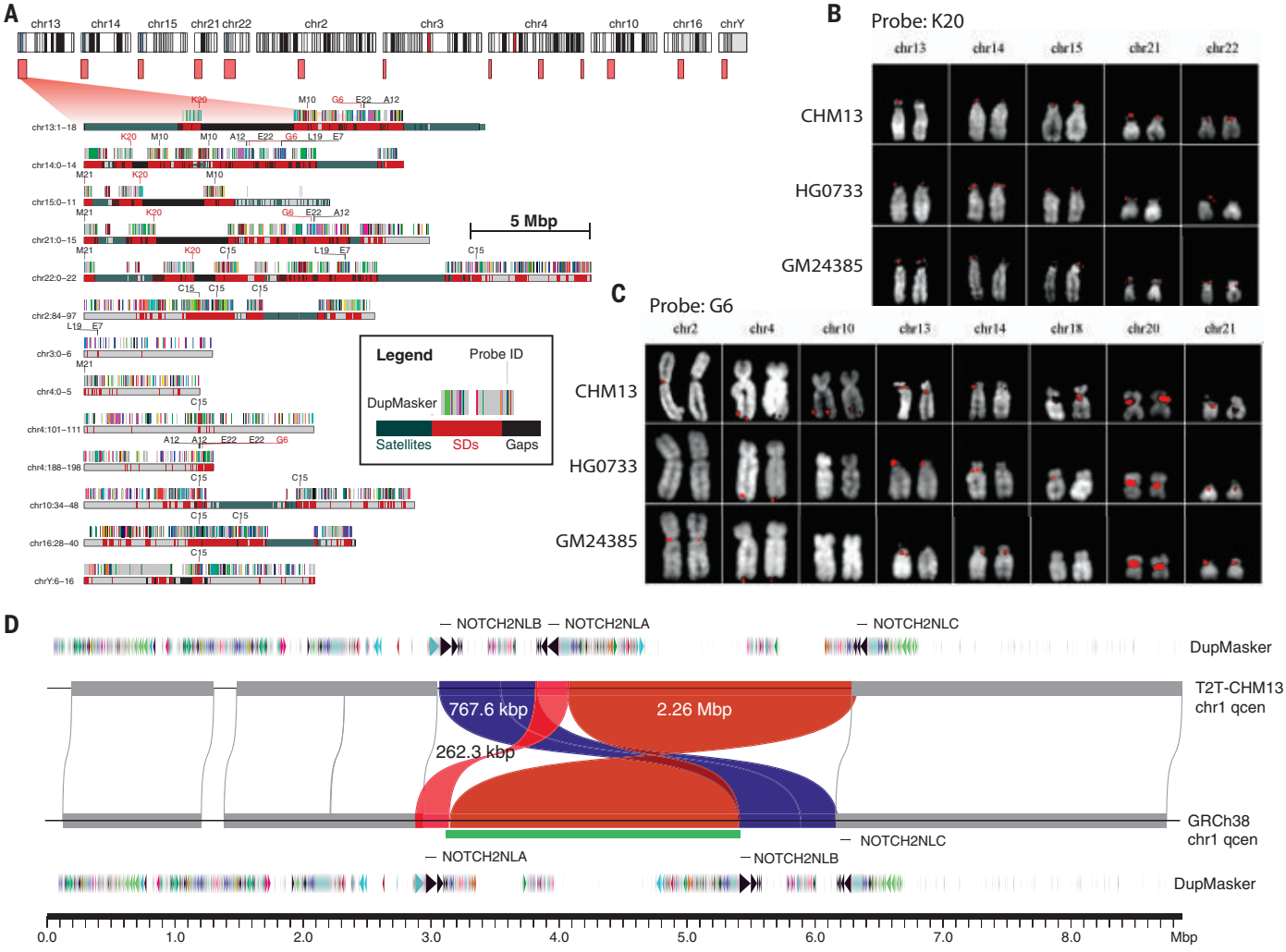


Fig. 2. Validation of previously unresolved SDs in T2T-CHM13 and heteromorphic variation. (**A**) Ideogram (top) depicts large SD regions (light red boxes) present in T2T-CHM13 but absent from the current reference human genome (GRCh38). An expanded view of the duplication (red) and satellite organization (blue) are depicted below showing the location of fosmid FISH probes (e.g., C15) and SD organization compared with ancestral duplicon segments (inset, multicolored bars). (**B** and **C**) FISH signals (red) shown on extracted metaphase for two probes and three human cell lines. Probe K20 shows a fixed signal (except for one heterozygous signal), and G6 is heteromorphic among humans (see table S4 and fig. S4 for complete descriptions of all nine probes). (**D**) Inversion polymorphism (green bar) between T2T-CHM13 and GRCh38 in the pericentromeric chromosome 1q region. The inversion (green bar) identified by Strand-seq (*32*) is confirmed in the assembly; however, the sequence-resolved assembly shows a more complex structure including two inversions (red) and one reordered segment (blue) mapping near the *NOTCH2NL* human-specific duplications.

we compared the copy number of both T2T-CHM13 and GRCh38 (*36*), successfully genotyping 1292 distinct copy number variable regions (74.85 Mbp). We found that T2T-CHM13 approximated the median (±2 SD) human copy number from SGDP for 94% of bases (70.6 Mbp) in contrast to GRCh38 in which 57% of bases (42.8 Mbp) meet this metric (fig. S10). In particular, the copy number for humans is nine times (59.26 versus 6.55 Mbp) more likely to more closely match the T2T-CHM13 copy number than that of GRCh38 in nonsyntenic SD regions (Fig. 3B). Thus, T2T-CHM13 is a better predictor (AUC 0.91) than GRCh38 (AUC 0.77) of human copy number variation and better approximates an in silico human reference constructed with the median copy number of the SGDP samples at every site (AUC 0.96; Fig. 3C). In nonsyntenic SD regions, GRCh38 tends to underestimate normal human copy number by an average of 9.2 copies or a median of 3.0 copies.

We identify 119 protein-encoding genes where the T2T-CHM13 copy number better represents the true human copy number state. By contrast, only 65 genes are better represented by GRCh38 (table S6). These include both biomedically important genes relevant to disease risk (*LPA* and *MUC3A*) (*37*–*43*) as well as gene families implicated in the expansion of the human brain during human evolution (*TBC1D3*, *NPIP*, and *NBPF*) (Fig. 3D and table S6) (*7*, *44*–*46*). In T2T-CHM13 for example, there are additional copies of *NPIP*, *NBPF*, and *GOLGA* that are absent from GRCh38, and each of these has been described as a core duplicon responsible for the expansion of interspersed duplications in the human genome (*27*) as well as the emergence of human-specific gene families.

Notably, African genomes tend to have an overall higher copy number status when compared with non-African genomes. In particular, *TBC1D3* shows about seven fewer copies in non-Africans when compared with Africans (P value < 1×10^{-12}). These findings suggest that higher copy number is likely ancestral (table S7) and that T2T-CHM13 once again better captures that diversity. Despite its primarily European origin, our results show that the more complete genome assembly serves as a better

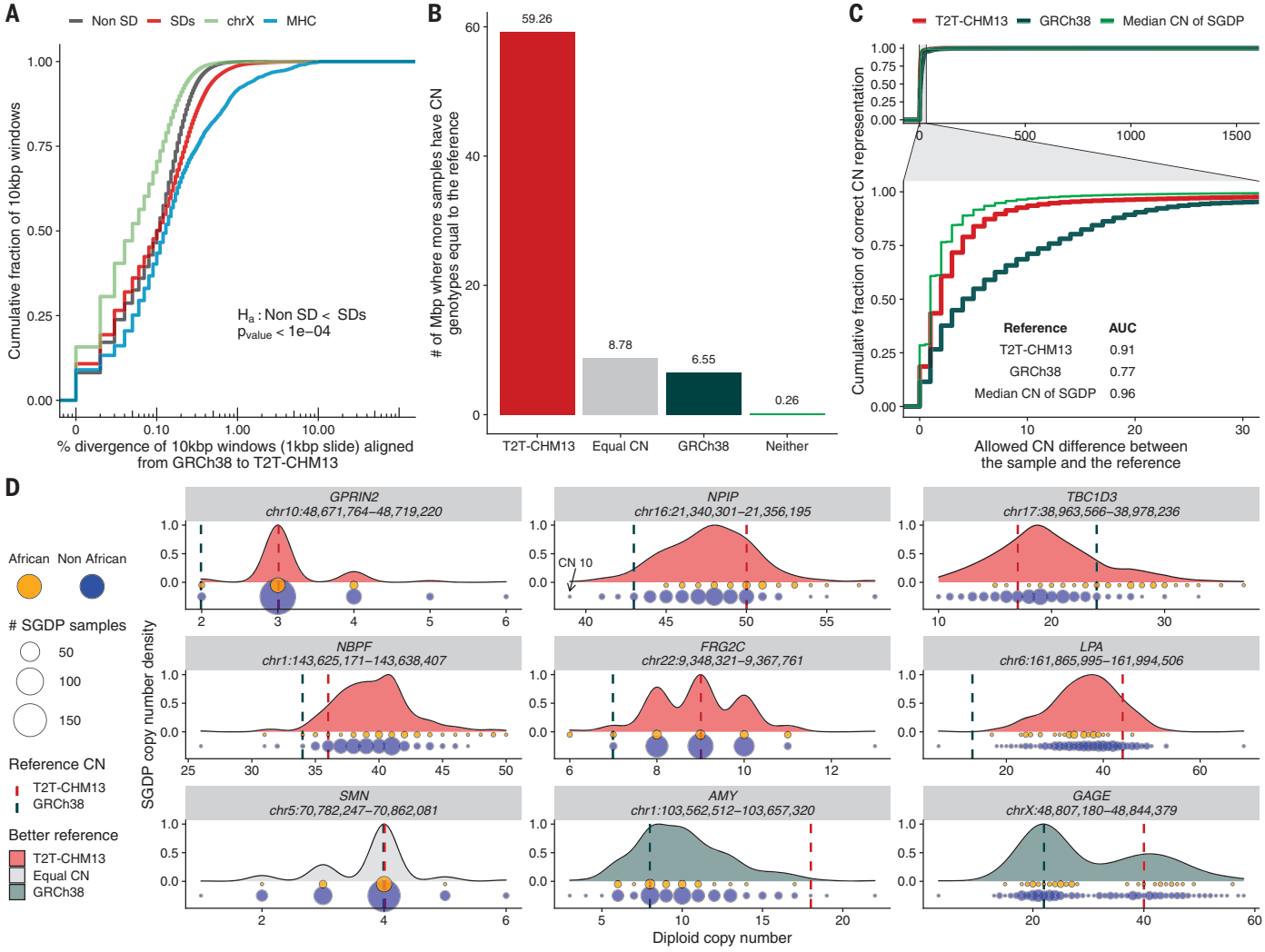


Fig. 3. SD single-nucleotide and copy number variation. (**A**) Sequence divergence (percent in 10-kbp bins) on the basis of syntenic alignments between GRCh38 and T2T-CHM13 for SDs (red), and unique genomic regions (black). SD regions show considerably more divergence when compared with unique sequences (black) and chromosome X (blue) but less than that of the MHC regions (green). (**B**) Copy number of SD regions that were previously unresolved or structurally different in T2T-CHM13 compared with GRCh38 on the basis of 268 human genomes from the Simons Genome Diversity Project (SGDP). The histogram shows the number of Mbp, in which more samples support the copy number of the given assembly [T2T-CHM13 (red), GRCh38 (blue), neither (green), or both equally (equal copy number, gray)]. (**C**) Empirical cumulative distribution showing how many samples genotype correctly with either GRCh38 or T2T-CHM13 as a function of the allowed difference between sample and reference copy number. The inset shows the area under the curve (AUC) calculation for both references allowing a maximum copy number difference of 30. The green curve shows an in silico reference made using the median copy number of the SGDP samples at each site. (**D**) Genic copy number variation. Copy number variation in nine gene families are shown (generated with SGDP) and distribution is colored according to which reference better reflects the median copy number; GRCh38 generally underestimates copy number (vertical lines) and Africans (orange) tend to show higher copy number than non-Africans (blue); circle size indicates number of samples.

reference for copy number variation irrespective of population group (fig. S11).

Structural variation and massive evolutionary changes in the human lineage

Advances in long-read genome assembly (*47*, *48*) enable sequence resolution of complex structural variation associated with SDs at the haplotype level (*49*). We generated or used existing high-fidelity (HiFi) sequence data from 12 human and five nonhuman primate genomes to understand both the structural diversity and evolution of specific SD regions. To guide the selection of candidate regions for analysis, we constructed a hifiasm assembly of a chimpanzee genome (note that the cell line used to make the sequencing data for this genome assembly was originally made from a now deceased chimpanzee individual, Clint), compared it with the T2T-CHM13 assembly, and searched for regions of substantial structural difference between the lineages. We focused first on the largest regions of insertion on the human lineage before subselecting those regions that contain genes of biomedical or evolutionary importance (tables S8 and S9). We restricted the analysis to insertions >50 kbp in length and selected 10 loci for a more detailed analysis, including genes associated with the expansion of the human frontal cortex (tables S8 and S9 and fig. S12). Assemblies of additional haplotypes recapitulated the structural organization of T2T-CHM13 for eight of the 10 loci, whereas evidence for the structural organization of GRCh38 was only found in five of the 10 loci (*25*). Overall, 73% of the human haplotype assemblies were successfully reconstructed (table S8); however, the fraction of human haplotypes resolved at each

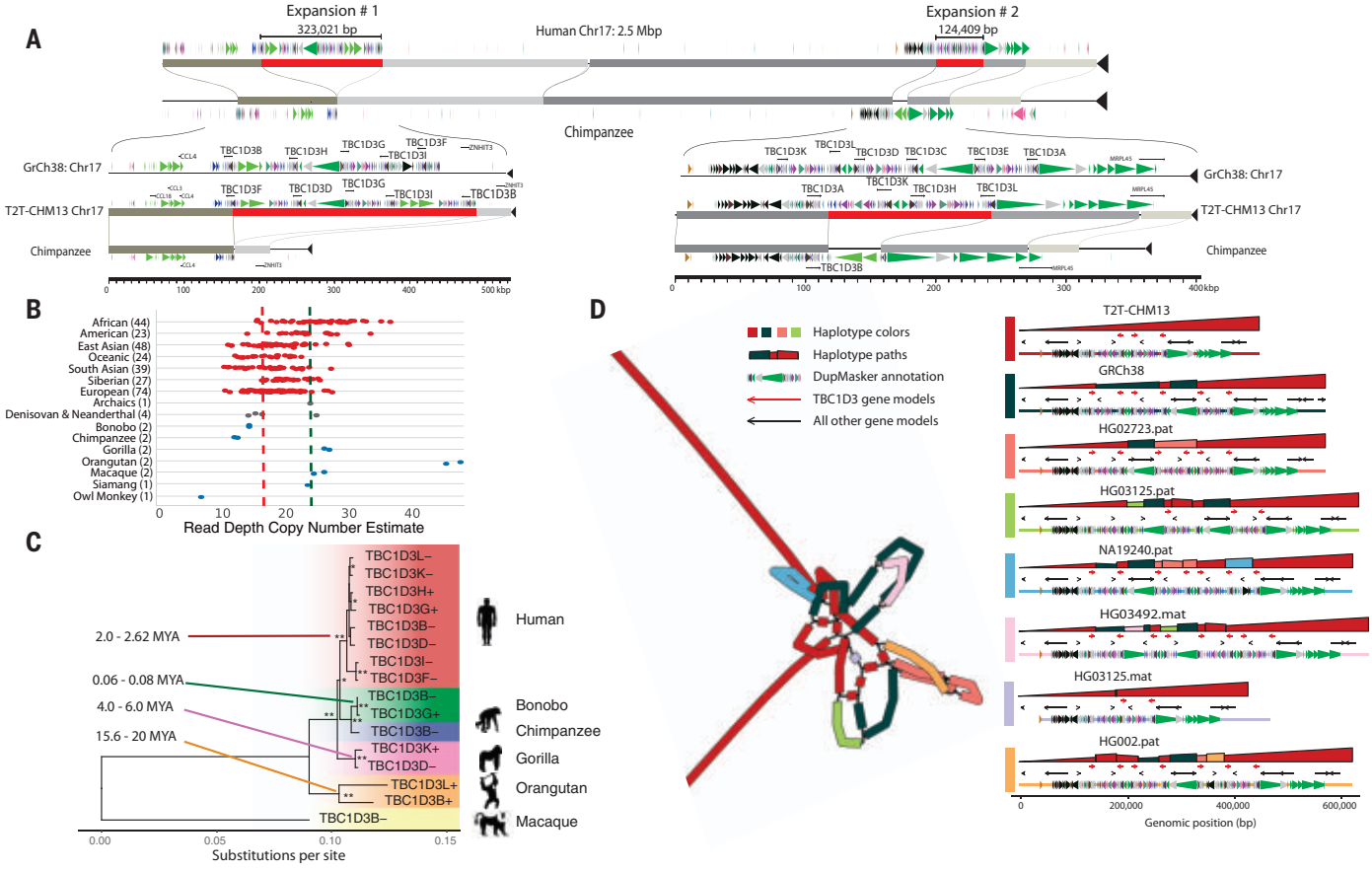


Fig. 4. Human-specific expansion of *TBC1D3* compared with nonhuman primates. (**A**) Regions of homology between human T2T-CHM13's chromosome 17 (top) and a HiFi assembly of the chimpanzee genome (bottom). Red blocks represent regions of human-specific expansion, including *TBC1D3* duplications. Colored arrows above and below the homologous sequence represent unique ancestral units (duplicons) identified by DupMasker. Inset plots for both expansion sites are included below with the gene models identified by Liftoff (*94*). (**B**) Copy number (diploid) estimates from an Illumina read-depth analysis of SGDP, ancient hominids, and nonhuman primates for a *TBC1D3* paralog (table S14). Copy number estimates include five pseudogenes not included in the phylogeny, explaining the higher counts observed. The T2T-CHM13 copy number and GRCh38 copy number are represented by red and blue lines, respectively. (**C**) Phylogeny of *TBC1D3* copies at these two expansion sites as well as nonhuman primate copies. Single asterisks at nodes indicate bootstrap values ≥70% whereas double asterisks indicate 100%. The data illustrate a human-specific expansion as well as several independent expansions in macaques, gorillas, and orangutans. Using a macaque sequence as an outgroup, we estimate the human-specific expansion to be ~2.3 million years ago (MYA). (**D**) Variation in human haplotypes across the first *TBC1D3* expansion site: a graph representation (rGFA, left) of the locus where colors indicate the source genome for the sequence, and on the right the path for each haplotype-resolved assembly through the graph. The top row for each haplotype composed of large polygons represents an alignment comparing the haplotype-resolved sequence (horizontal) against the graph (vertical), and color represents the source haplotype for the vertical sequence. For example, a single large red triangle indicates there is a one-to-one alignment between T2T-CHM13 and the haplotype. Structural variants can be identified from discontinuities in height (deletion), changes between colors (insertion), or changes in the direction of the polygon (inversion). Below is shown the gene of interest (red arrow) and other genic content in the region (black arrow). Colored bars show ancestral duplication segments (duplicons) that compose the larger duplication blocks.

locus varied considerably depending on the size and complexity of the region (fig. S13). For example, in the case of the 8.9-Mbp region corresponding to *NOTCH2NL* and *SRGAP2B/2D*, we recovered only 37.5% of human haplotypes (table S8 and fig. S7). Similarly, we resolved only six haplotypes (from a potential of 24 haplotypes) for the 3.4-Mbp region harboring the *SMN1* and *SMN2* loci (fig. S14).

Among the haplotypes that could be resolved, we found a high degree of structural heterozygosity among human genomes (*25*) with 249 kbp differing on average when compared with T2T-CHM13 (table S10). In some cases the structural changes are simple, such as ~12 kbp insertion or deletion of *CYP2D6*, which contributes to differential drug metabolism activity in addition to other human disease susceptibilities (*50*–*56*) (fig. S15). In other cases the patterns of structural variation are complex, involving hundreds of kilobase pairs of inserted or deleted gene-rich sequences along with large-scale inversion events that alter gene order for specific human haplotypes (see *ARHGAP11A/B*; fig. S16, and *NOTCH2NLA/B*; fig. S7). Furthermore, the spinal muscular atrophy (SMA) locus containing *SMN1* and *SMN2*—one of the most difficult regions to finish as part of the Human Genome Project on chromosome 5 (*57*)—shows a distinct structure for all seven assembled haplotypes including GRCh38. Some haplotypes not only show increases in *SMN2* copy number (fig. S14)—a genetic modifier of SMA (*58*)—but also potential functional differences in the organization and composition of *SMN2*. Because *SMN2* serves as a target for small-molecule drug therapy that improves splice-site efficiency compensating for the loss of *SMN1* in SMA patients (*59*), this level of

sequence resolution suggests practical utility for disease risk assessment and treatment of patients.

Of particular interest is the *TBC1D3* gene family (*44*) (Fig. 4 and figs. S17 and S18), the protein products of which modulate epidermal growth factor receptor signaling and trafficking (*60*); further, their duplication in humans has been associated with expansion of the human prefrontal cortex as evidenced by mouse transgenic experiments (*7*). A comparison with chimpanzees (Fig. 4A) shows two massive genomic expansions in the human lineage (323.0 and 124.4 kbp). Both the high sequence identity (99.6%) and sequence read-depth comparisons of *TBC1D3* copy number are consistent with expansion occurring in the human lineage after divergence from chimpanzees (Fig. 4B).

We extended this analysis to other nonhuman primates by generating HiFi assemblies for bonobos, gorillas, orangutans, and macaques. We identified *TBC1D3* homologs in each species and constructed a maximum likelihood phylogeny by using intronic or noncoding sequences flanking the gene (Fig. 4C). The analysis reveals recurrent and independent expansions of *TBC1D3* in orangutan, gorilla, and macaque species at different time points during primate evolution, with the most recent expansions occurring 2 million and 2.6 million years ago. However, these estimates assume that there has not been substantial interlocus gene conversion, which may not be the case.

Complete sequencing of human *TBC1D3* haplotypes reveals notable structural diversity (Fig. 4D) with *TBC1D3* copy number ranging from three to 14 *TBC1D3* copies at expansion site 1 and two to nine copies at expansion site 2. In total, approximately one-third of human expansion site 2 shows large-scale structural variation, and we identify >1.8 Mbp of duplicated sequence and >650 kbp of inverted sequence across the 18 haplotypes (including GRCh38). We estimate the structural heterozygosity of this locus to be 90.1% with 14 of 18 haplotypes showing structurally distinct duplication configurations (fig. S18). Similarly, *TBC1D3* expansion site 1 is 87.6% heterozygous with 14 of 22 haplotypes displaying unique structures corresponding to copy number differences in the *TBC1D3* gene family (fig. S17). Using orthogonal Oxford Nanopore Technologies (ONT) ultra-long-read sequencing, we validated these complex patterns of structural variation in a subset of the samples investigated here (*25*) (figs. S19 and S20). To better represent the structural genetic variation at this locus, we used a graph-based representation (*61*), which identified two *TBC1D3* genes as common among all human haplotypes examined thus far (*TBC1D3B* at site 1 and *TBC1D3A* at site 2).

Additional gene models and variable duplicate genes

We identified 182 candidate genes that were previously unresolved or nonsyntenic (*25*) in the T2T-CHM13 genome assembly (compared with GRCh38) with open reading frames and multiple exons (table S11). Of these, 91% (166) corresponded to SD gene families (Fig. 5A). Many of these represent expanded tandem duplications (e.g., *GAGE* gene family members on the X chromosome) or large interspersed duplications (e.g., beta-defensin locus) adding additional copies of nearly identical genes to the human genome (Fig. 5A).

We searched for evidence that these copy number polymorphic or structurally variant regions were transcribed by aligning long-read transcript sequencing data and searching for perfect matches (*25*). We constructed a database of 44.2 million full-length cDNA transcripts derived from 31 human tissue samples and compared them with both the GRCh38 and T2T-CHM13 human genome references. For those 182 previously unresolved protein-coding genes where an unambiguous assignment could be made, 36% (65 of 182, >20 Iso-Seq reads) were confirmed as expressed, and 23 of them showed that more reads mapped better to T2T-CHM13 when compared with GRCh38 (Fig. 5B).

Overall, across the entire genome 12% of full-length transcripts exhibited at least 0.2% higher alignment identity when mapped against T2T-CHM13, whereas 8% aligned better to GRCh38. These results are consistent with the notion that T2T-CHM13 is more complete, but that in some cases both assemblies capture different structurally variant haplotypes associated with genes. In addition to entirely new genes, we identified several gene models that were previously incomplete—many of which encode proteins with large tandem repeat domains (ZNF, LPA, Mucin; Fig. 5C). Among these is the complete gene structure of the kringle IV domain of the lipoprotein A gene. Reduced copies of this domain have some of the strongest genetic associations with cardiovascular disease, especially among African Americans (*37*–*40*, *62*). Sequencing of multiple human haplotypes not only identified length variation but also other forms of rare coding variants potentially relevant for disease risk (Fig. 5D).

SD methylation and transcription

Because methylation is an important consideration in regulating gene transcription, we took advantage of the signal inherent in ultra-long ONT data (*63*–*65*) to investigate the CpG methylation status of SD genes within the CHM13 genome (*25*). Using hierarchical clustering, we found that SD blocks are generally either methylated or unmethylated as an entire block; (fig. S21 and Fig. 6A). Specifically, we found that 452 SD blocks (127.7 Mbp) flanked by unique sequences are hypermethylated in contrast to 222 hypomethylated SD blocks (52.1 Mbp). Methylation status does not appear to be driven by genomic location, e.g., proximity to centromeres, acrocentric short arms, or telomeres (Fig. 6A).

Using full-length transcript data from CHM13, we compared methylation and transcription status of duplicated genes (*25*). If we stratify genes by their number of full-length transcripts, we observe distinct methylation patterns for transcribed and nontranscribed SD genes (Fig. 6B). For highly transcribed SD genes (genes without at least one exon overlapping with SD sequence) and unique genes, the gene body and flanking sequence are generally hypermethylated with a pronounced dip near the transcription start site (TSS) and promoter (*66*). By contrast, nontranscribed genes show moderate to low methylation across the gene body and flanking sequence.

Restricting the analysis to genes mapping within SDs, we find that transcriptionally silenced duplicate genes are more likely (10,000 permutations, $P = 0.0018$) to map to hypomethylated regions of SD sequences (Fig. 6A) when compared with transcribed duplicate genes. Additionally, in untranscribed SD genes, we observe a statistically significant (P value < 0.001, one-sided Mann-Whitney U test) increase in TSS methylation (6.6% increase) when compared with unique genes where the TSS is more likely to be depleted for methylation (8.2% decrease).

One important consideration in this analysis is the presence of a CpG island within 1500 bp of the promoter (*67*). In our analysis of CHM13, for example, unexpressed unique genes have a low CpG count, consistent with a lack of CpG islands (fig. S22). If we repeat the same analysis on SD genes, we find that the unexpressed SD genes exist with and without CpG islands (fig. S22). In total, these observations suggest a process of epigenetic silencing for a subset of duplicate genes through general demethylation of the gene body but hypermethylation of promoter regions. On the basis of these observed signatures, we investigated whether these epigenetic features coincided with actively transcribed members of duplicate gene families.

We investigated a recently duplicated hominid gene family (*NPIPA*) (*68*) where sufficient paralogous sequence differences exist to unambiguously assign full-length transcripts to specific loci. Although promoter and TSS signatures are less evident at the individual gene level, the gene body methylation signal appears diagnostic (Fig. 6C). *NPIPA1* and *NPIPA9*, for example, are the most transcriptionally active and show demonstrably distinct methylation patterns providing an epigenetic signature to distinguish transcriptionally active loci associated with high-identity gene families that are otherwise largely indistinguishable. We

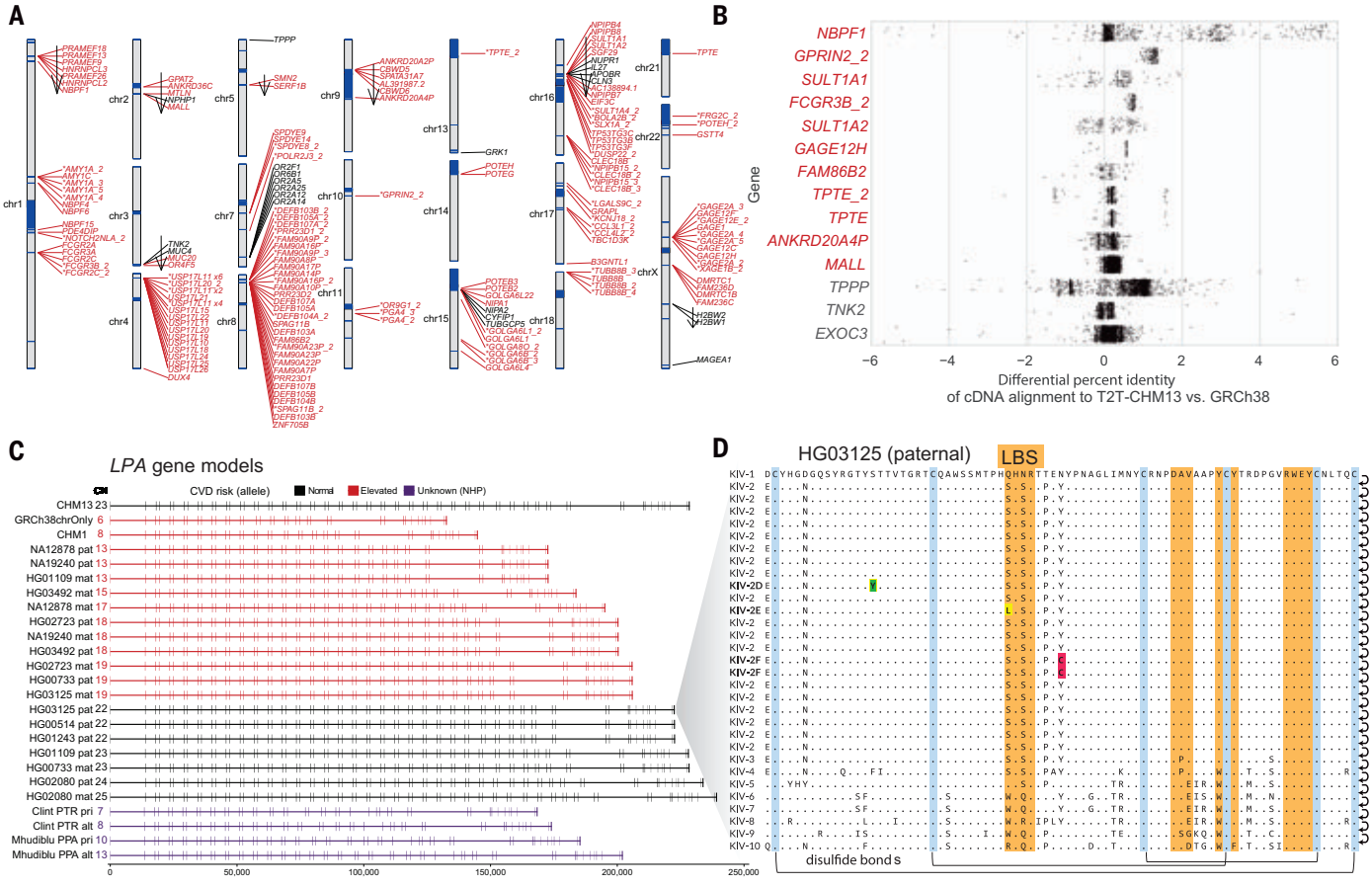


Fig. 5. Genic variation in previously unresolved SD regions of T2T-CHM13. (**A**) Ideogram showing the previously unresolved or nonsyntenic gene models [open reading frames (ORFs) with >200 bp of coding sequence and multiple exons] in the T2T-CHM13 assembly as predicted by Liftoff. Previously unresolved genes mapping to SDs (red) are indicated with an asterisk if predicted to be an expansion in the gene family relative to GRCh38 (*25*). Arrows indicate inverted regions. Most unique genes mapping to nonsyntenic regions (black) are the result of an inversion (arrow). (**B**) Percent improvement in mapping of CHM13 Iso-Seq reads in candidate duplicated genes (red) mapping to nonsyntenic regions of the T2T-CHM13 assembly. Positive values identify Iso-Seq reads aligning better to T2T-CHM13 than GRCh38. (**C**) Gene models of *LPA* with ORF generated from haplotype-resolved HiFi assemblies. The double-exon repeat in these gene models encode for the kringle IV subtype 2 domain of the LPA protein. Highlighted in red are haplotypes with reduced kringle IV subtype 2 repeats predicted to increase risk of cardiovascular disease. (**D**) Amino acid variation in the kringle IV subtype 2 repeat in the paternal haplotype of HG01325 identifies a previously unknown set of amino acid substitutions including rare variants: Ser42Leu in the active site, Ser24Tyr, and Tyr49Cys.

show that this trend also holds for other gene families with high copy numbers (fig. S23).

Discussion

This work provides a comprehensive view of the organization of SDs in the human genome. The T2T-CHM13 reference adds a chromosome's worth (81 Mbp) of SDs, increasing the human genome average from 5.4 to 7.0% and nearly doubling the number of SD pairwise relationships (24 thousand versus 41 thousand) and thus predicts regions of genomic instability as a result of their potential to drive unequal crossing-over events during meiosis.

By every metric, T2T-CHM13 improves our representation of the structure of the human genome. This includes sequence-based organization of the short arms of chromosomes 13, 14, 15, 20, and 21, where we find that SDs account for more sequence (34.6 Mbp) than either heterochromatic satellite (26.7 Mbp) or rDNA (10 Mbp). Acrocentric SDs are almost twice as large when compared with non-acrocentric regions likely because of ectopic exchange events occurring among the short arms, which associate more frequently during the formation of the nucleolus (*69*).

Notably, nearly half of the acrocentric SDs involve duplications with nonacrocentric pericentromeric regions of chromosomes 1, 3, 4, 7, 9, 16, and 20. These duplicated islands of euchromatic-like sequences within acrocentric DNA are much more extensive than previously thought but have been shown to be transcriptionally active (*70*). Although the underlying mechanism for their formation is unknown, it is noteworthy that three of the nonacrocentric regions have large secondary constriction sites (chromosomes 1q, 9q, and 16q) composed almost entirely of heterochromatic satellites (HSAT2 and HSAT3) (fig. S2). These particular SD blocks thus are bracketed by large tracts of heterochromatic satellites, and such configurations may make them particularly prone to double-strand breakage events (*71*) promoting interchromosomal duplications (fig. S3) between acrocentric and nonacrocentric chromosomes.

The T2T-CHM13 reference, along with resources from other human genomes, provides a baseline for investigating more complex forms of human genetic variation. For example, this complete reference sequence facilitates the design of sequence-anchored probes to systematically discover and characterize SD heteromorphic variation where chromosome organization differs among individuals (Fig. 2). Such chromosomal heteromorphisms have been traditionally investigated cytogenetically and are thought to be clinically benign (*29*–*31*). However,

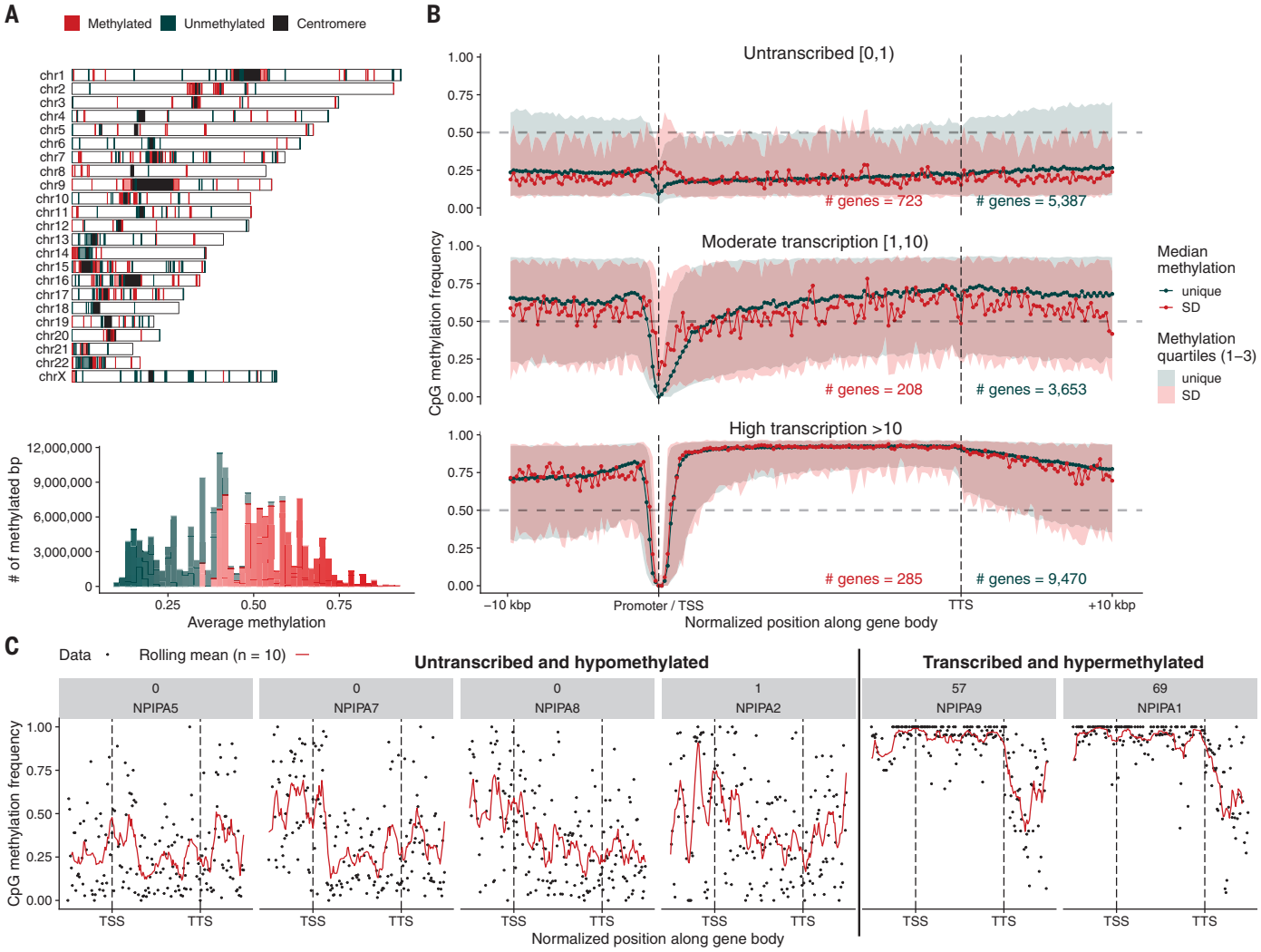


Fig. 6. SD methylation and gene transcription. (**A**) Methylated (red) or unmethylated (blue-gray) SD blocks in the CHM13 genome on the basis of processing ONT data. The histogram shows the distribution of average methylation across these regions. (**B**) Median methylation signal of SD (red) and unique (blue-gray) genes stratified by their Iso-Seq expression levels in CHM13. The filled intervals represent the 25 and 75 percentiles of the observed data. Vertical lines indicate the position of the transcription start site (TSS) and the transcription termination site (TTS). (**C**) Methylation signal across the recently duplicated *NPIPA* gene family in CHM13, showing increased methylation in transcriptionally active copies. Black points are individual methylation calls, and the red line is a rolling mean across 10 methylation sites. The labels in gray show the number of CHM13 Iso-Seq transcripts and the gene name.

recent work indicates that these large-scale variants are associated with infertility through increasing sperm aneuploidy, decreasing rates of embryonic cleavage (in vitro fertilization, IVF), and increasing miscarriages (*72*–*79*). Distinguishing between fixed and heteromorphic acrocentric SDs will facilitate such research as well as the characterization of breakpoints associated with Robertstonian translocations—the most common form of human translocation (*80*).

At a finer-grained level, the T2T-CHM13 reference and the use of long reads from other human genomes provides access to other complex forms of variation involving duplicated gene families. Short-read copy number variation analyses and single-nucleotide polymorphism microarrays have long predicted that SDs are enriched 10-fold for copy number variation, but the structural differences underlying these regions—as well as their functional consequences—have remained elusive (*10*, *81*). We reveal elevated levels of human genetic variation in genes important for human neurodevelopment (*TBC1D3*) and disease (*LPA*, *SMN*). Even between just two genomes (GRCh38 and CHM13), we find that 37% (81 Mbp) of SD bases are uncharacterized or structurally variable, and that this predicts 182 copy number variable genes between two human haplotypes (table S6).

In cases such as *TBC1D3*, we find that most human haplotypes vary structurally (64 to 78%). Different individuals carry different complements and arrangements of the *TBC1D3* gene family. The potential ramifications of this considerable expansion in humans versus chimpanzees and of such high structural heterozygosity among humans are notable given this gene's purported role in expansion of the frontal cortex (*7*). Similarly, we were able to reconstruct the complete structure of the *LPA* gene model in multiple human genotypes. Although *LPA* is only a single gene, variability in the tandemly repeated 5.2-kbp protein-encoding kringle IV domain underlies one of the most important genetic risk factors for cardiovascular disease. Sequence resolution of structural variation—as well as underlying amino acid differences—allow us to predict previously uncharacterized risk alleles for disease

(Fig. 5). Sequence-resolved structural variation improves genotyping and tests of selection (*49*, *82*, *83*), providing a path forward for understanding the disease and evolutionary implications of these complex forms of genetic variation.

Finally, the T2T-CHM13 reference—coupled with other long-read datasets—enables genome-wide functional characterization of recently duplicated genes. Both gene annotation and large-scale efforts to characterize the regulatory landscape of the human genome have typically excluded repetitive regions, including the 859 human genes that map to high-identity SDs (*84*, *85*). This is because the underlying short-read sequencing limits conventional RNA-seq or Chip-seq data from being assigned unambiguously to specific duplicated genes.

In this study, we used long-read full-length transcript data (Iso-Seq) (*86*) with long-read methylation data from ONT sequencing of the same genome to simultaneously investigate epigenetic and transcriptional data against a fully assembled reference genome. The long-read data from the same haploid source facilitated the unambiguous assignment of these functional readouts, allowing us to correlate methylation and transcript abundance. Our initial analyses suggest that a large fraction of duplicate genes are in fact epigenetically silenced (characterized by hypermethylation of the promoter and hypomethylation of the gene body) and that this epigenetic mark may be used to predict actively transcribed loci even when genes are virtually identical (Fig. 6 and fig. S23). Although more human genomes and diverse tissues need to be investigated to assess the implications of this observation, it is clear that phased genome assemblies (*49*) with long-read functional readouts such as methylation (*65*), transcription, or Fiber-seq (*87*, *88*) provide a powerful approach to understanding the regulatory landscape of duplicated and copy number polymorphic genes in the human genome.

However, there are several remaining challenges: First, not all human haplotypes corresponding to specific duplicated regions could be fully sequence-resolved with automated assembly of long-read HiFi sequencing technology. Most of the 250 unresolved regions of phased human genomes generated solely with HiFi long reads correspond to some of the largest and most variable duplicated regions of the human genome (*49*). For example, only 25% of *SMN1* and *SMN2* haplotypes were fully resolved by HiFi assembly, and these unresolved loci are predicted to carry some of the most complex structural variation patterns. In comparison, the T2T-CHM13 assembly used both accurate HiFi and ultra-long ONT data, and future assembly methods that combine these technologies will likely be critical for diploid T2T assembly and the complete characterization of SD haplotypes (*18*, *86*).

Another important challenge going forward will be accurate representation of these more complex forms of human genetic variation, including functional annotation where linear representations may be insufficient. Although a more complex pangenome reference graph could overcome these limitations, it is unclear how this will be achieved in practice or how it will be adopted by genomics and clinical communities. This highlights the importance of not only the construction of a pangenome reference but development of necessary tools that will distinguish paralogous and orthologous sequences within duplications to allow for comparison between haplotypes with different SD architectures. The work currently underway by the Human Pangenome Reference Consortium (HPRC), Human Genome Structural Variation Consortium (HGSVC), and Telomere-to-Telomere (T2T) Consortium will be key to developing these methods and completing our understanding of SDs and their role in human genetic variation.

Materials and methods summary

SDs in T2T-CHM13 were identified using SEDEF (*21*) after repeat masking with Tandem Repeats Finder (TRF) (*89*) and RepeatMasker (*90*). Syntenic one-to-one alignments were determined using halSynteny (*91*). Copy number prediction based on short-read data was performed with WSSD (*3*, *14*) and mrsFAST (*92*), and regions of comparable copy number were determined with the changepoint package in R (*93*). To generate gene annotations we used the tools Liftoff (*94*) and GffRead (*95*). Fosmid probes were selected from the ABC10 library (*28*, *96*) and two-color FISH was performed to experimentally validate acrocentric SDs (*97*–*99*). All assemblies (table S12) with the exception of T2T-CHM13 and GRCh38 were assembled with hifiasm v0.12 (*47*) using default parameters. Assembly validation of *TBC1D3* was performed using sample-matched ONT data by checking the consistency of read alignments to the assemblies (*100*, *101*). Phylogenetic analysis of *TBC1D3* was performed with MAFFT and RAxML (*102*–*105*). Assembling pangenome graphs for select loci was performed with minigraph (*61*). Methylation analysis was performed using the methods and data described in Gershman *et al.* (*106*) using Winnowmap2 and Nanopolish for mapping and methylation calling (*65*, *107*). Data visualization and figures (with the exception of Miropeats) (*108*) were primarily made in R making use of GenomicRanges (*109*), Tidyverse (*110*), karyoploteR (*111*), and circlize (*112*). Pipelines used for large-scale data analysis were constructed with Snakemake (*113*–*115*). Detailed descriptions of materials and methods are available in the supplementary materials (*25*).

REFERENCES AND NOTES

1. S. Ohno, U. Wolf, N. B. Atkin, Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–187 (1968). doi: 10.1111/j.1601-5223.1968.tb02169.x; pmid: 5662632
2. S. Ohno, *Evolution by Gene Duplication* (Springer, 1970).
3. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001). doi: 10.1101/gr.187101; pmid: 11381028
4. G. M. Cooper *et al.*, A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011). doi: 10.1038/ng.909; pmid: 21841781
5. M. Y. Dennis *et al.*, Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012). doi: 10.1016/j.cell.2012.03.033; pmid: 22559943
6. M. Florio *et al.*, Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife* **7**, e32332 (2018). doi: 10.7554/eLife.32332; pmid: 29561261
7. X. C. Ju *et al.*, The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *eLife* **5**, e18197 (2016). doi: 10.7554/eLife.18197; pmid: 27504805
8. I. T. Fiddes *et al.*, Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369.e22 (2018). doi: 10.1016/j.cell.2018.03.051; pmid: 29856954
9. P. H. Sudmant *et al.*, 1000 Genomes Project, Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010). doi: 10.1126/science.1197005; pmid: 21030649
10. P. H. Sudmant *et al.*, Global diversity, population stratification, and selection of human copy-number variation. *Science***349**, aab3761 (2015). doi: 10.1126/science.aab3761; pmid: 26249230
11. J. C. Venter *et al.*, The sequence of the human genome. *Science* **291**, 1304–1351 (2001). doi: 10.1126/science.1058040; pmid: 11181995
12. X. She *et al.*, Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004). doi: 10.1038/nature03062; pmid: 15496912
13. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). doi: 10.1038/35057062; pmid: 11237011
14. J. A. Bailey *et al.*, Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002). doi: 10.1126/science.1072047; pmid: 12169732
15. E. Eichler, U. Surti, R. Ophoff, Proposal for Construction a Human Haploid BAC library from Hydatidiform Mole Source Material (2002). www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf.
16. D. Fredman *et al.*, Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.* **36**, 861–866 (2004). doi: 10.1038/ng1401; pmid: 15247918
17. M. J. Chaisson *et al.*, Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). doi: 10.1038/nature13907; pmid: 25383537
18. K. H. Miga *et al.*, Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi: 10.1038/s41586-020-2547-7; pmid: 32663838
19. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010). doi: 10.1126/science.1188021; pmid: 20448178
20. S. Nurk *et al.*, The complete sequence of a human genome. *Science* **376**, 44–53 (2022). doi: 10.1126/science.abj6987
21. I. Numanagic *et al.*, Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**, i706–i714 (2018). doi: 10.1093/bioinformatics/bty586; pmid: 30423092
22. A. D. Bell, C. L. Usher, S. A. McCarroll, Analyzing Copy Number Variation with Droplet Digital PCR. *Methods Mol. Biol.* **1768**, 143–160 (2018). doi: 10.1007/978-1-4939-7778-9_9; pmid: 29717442
23. I. L. Gonzalez, J. E. Sylvester, Complete sequence of the 43-kb human ribosomal DNA repeat: Analysis of the intergenic spacer. *Genomics* **27**, 320–328 (1995). doi: 10.1006/geno.1995.1049; pmid: 7557999
24. J. H. Kim *et al.*, Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read

sequencing. *Nucleic Acids Res.* **46**, 6712–6725 (2018). doi: 10.1093/nar/gky442; pmid: 29788454
25. Supplementary materials and methods for segmental duplications and their variation in a complete human genome.
26. Z. Jiang, R. Hubley, A. Smit, E. E. Eichler, DupMasker: A tool for annotating primate segmental duplications. *Genome Res.* **18**, 1362–1368 (2008). doi: 10.1101/gr.078477.108; pmid: 18502942
27. Z. Jiang *et al*., Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007). doi: 10.1038/ng.2007.9; pmid: 17922013
28. J. M. Kidd *et al*., Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008). doi: 10.1038/nature06862; pmid: 18451855
29. M. K. Bhasin, Bhasin, Human population cytogenetics: A review. *Int. J. Hum. Genet.* **5**, 83–152 (2005). doi: 10.1080/09723757.2005.11885918
30. L. Y. F. Hsu *et al*., Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: A large prenatal study. *Am. J. Med. Genet.* **26**, 95–101 (1987). doi: 10.1002/ajmg.1320260116; pmid: 3812584
31. J. C. K. Barber, Euchromatic heteromorphism or duplication without phenotypic effect? *Prenat. Diagn.* **14**, 323–324 (1994). doi: 10.1002/pd.1970140418; pmid: 8066046
32. M. J. P. Chaisson *et al*., Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019). doi: 10.1038/s41467-018-08148-z; pmid: 30992455
33. A. D. Sanders *et al*., Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* **26**, 1575–1587 (2016). doi: 10.1101/gr.201160.115; pmid: 27472961
34. D. Porubský *et al*., Recurrent inversion toggling and great ape genome evolution. *Nat. Genet.* **52**, 849–858 (2020). doi: 10.1038/s41588-020-0646-x; pmid: 32541924
35. I. K. Suzuki *et al*., Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell* **173**, 1370–1384.e16 (2018). doi: 10.1016/j.cell.2018.03.067; pmid: 29856955
36. S. Mallick *et al*., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi: 10.1038/nature18964; pmid: 27654912
37. R. Clarke *et al*., Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N. Engl. J. Med.* **361**, 2518–2528 (2009). doi: 10.1056/NEJMoa0902604; pmid: 20032323
38. S. Coassin *et al*., A comprehensive map of single-base polymorphisms in the hypervariable *LPA* kringle IV type 2 copy number variation region. *J. Lipid Res.* **60**, 186–199 (2019). doi: 10.1194/jlr.M090381; pmid: 30413653
39. F. Kronenberg, G. Utermann, Lipoprotein(a): Resurrected by genetics. *J. Intern. Med.* **273**, 6–30 (2013). doi: 10.1111/j.1365-2796.2012.02592.x; pmid: 22998429
40. K. Schmidt, A. Noureen, F. Kronenberg, G. Utermann, Structure, function, and genetics of lipoprotein (a). *J. Lipid Res.* **57**, 1339–1359 (2016). doi: 10.1194/jlr.R067314; pmid: 27074913
41. J. R. Gum *et al*., Molecular cloning of cDNAs derived from a novel human intestinal mucin gene. *Biochem. Biophys. Res. Commun.* **171**, 407–415 (1990). doi: 10.1016/0006-291X(90)91408-K; pmid: 2393399
42. W. S. Pratt *et al*., Multiple transcripts of MUC3: Evidence for two genes, MUC3A and MUC3B. *Biochem. Biophys. Res. Commun.* **275**, 916–923 (2000). doi: 10.1006/bbrc.2000.3406; pmid: 10973822
43. K. Kyo, T. Muto, H. Nagawa, G. M. Lathrop, Y. Nakamura, Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease. *J. Hum. Genet.* **46**, 5–20 (2001). doi: 10.1007/s100380170118; pmid: 11289722
44. C. A. Paulding, M. Ruvolo, D. A. Haber, The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 2507–2511 (2003). doi: 10.1073/pnas.0437015100; pmid: 12604796
45. S. Cantsilieris *et al*., An evolutionary driver of interspersed segmental duplications in primates. *Genome Biol.* **21**, 202 (2020). doi: 10.1186/s13059-020-02074-4; pmid: 32778141
46. T. Marques-Bonet, E. E. Eichler, The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 355–362 (2009). doi: 10.1101/sqb.2009.74.011; pmid: 19717539
47. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021). doi: 10.1038/s41592-020-01056-5; pmid: 33526886
48. S. Nurk *et al*., HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020). doi: 10.1101/gr.263566.120; pmid: 32801147
49. P. Ebert *et al*., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science***372**, eabf7117 (2021). doi: 10.1126/science.abf7117; pmid: 33632895
50. L. Bertilsson, M. L. Dahl, P. Dalén, A. Al-Shurbaji, Molecular genetics of CYP2D6: Clinical relevance with focus on psychotropic drugs. *Br. J. Clin. Pharmacol.* **53**, 111–122 (2002). doi: 10.1046/j.0306-5251.2001.01548.x; pmid: 11851634
51. W. Hammer, F. Sjöqvist, Plasma levels of monomethylated tricyclic antidepressants during treatment with imipramine-like compounds. *Life Sci.* **6**, 1895–1903 (1967). doi: 10.1016/0024-3205(67)90218-4; pmid: 6052684
52. B. Alexanderson, D. A. Evans, F. Sjöqvist, Steady-state plasma levels of nortriptyline in twins: Influence of genetic factors and drug therapy. *BMJ* **4**, 764–768 (1969). doi: 10.1136/bmj.4.5686.764; pmid: 5391106
53. R. C. Skoda, F. J. Gonzalez, A. Demierre, U. A. Meyer, Two mutant alleles of the human cytochrome P-450db1 gene (P450C2D1) associated with genetically deficient metabolism of debrisoquine and other drugs. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5240–5243 (1988). doi: 10.1073/pnas.85.14.5240; pmid: 2899325
54. M. L. Dahl, I. Johansson, M. P. Palmertz, M. Ingelman-Sundberg, F. Sjöqvist, Analysis of the CYP2D6 gene in relation to debrisoquin and desipramine hydroxylation in a Swedish population. *Clin. Pharmacol. Ther.* **51**, 12–17 (1992). doi: 10.1038/clpt.1992.2; pmid: 1346258
55. A. Gaedigk, M. Blum, R. Gaedigk, M. Eichelbaum, U. A. Meyer, Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. *Am. J. Hum. Genet.* **48**, 943–950 (1991). pmid: 1673290
56. I. Johansson *et al*., Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11825–11829 (1993). doi: 10.1073/pnas.90.24.11825; pmid: 7903454
57. J. Schmutz *et al*., The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, 268–274 (2004). doi: 10.1038/nature02919; pmid: 15372022
58. M. E. Butchbach, Copy Number Variations in the Survival Motor Neuron Genes: Implications for Spinal Muscular Atrophy and Other Neurodegenerative Diseases. *Front. Mol. Biosci.* **3**, 7 (2016). doi: 10.3389/fmolb.2016.00007; pmid: 27014701
59. T. W. Bebee, J. T. Gladman, D. S. Chandler, Splicing regulation of the Survival Motor Neuron genes and implications for treatment of spinal muscular atrophy. *Front. Biosci.* **15**, 1191 (2010). doi: 10.2741/3670; pmid: 20515750
60. M. J. Wainszelbaum *et al*., The hominoid-specific oncogene TBC1D3 activates Ras and modulates epidermal growth factor receptor signaling and trafficking. *J. Biol. Chem.* **283**, 13233–13242 (2008). doi: 10.1074/jbc.M800234200; pmid: 18319245
61. H. Li, X. Feng, C. Chu, The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020). doi: 10.1186/s13059-020-02168-z; pmid: 33066802
62. F. Kronenberg, Human Genetics and the Causal Role of Lipoprotein(a) for Various Diseases. *Cardiovasc. Drugs Ther.* **30**, 87–100 (2016). doi: 10.1007/s10557-016-6648-3; pmid: 26896185
63. J. Quick *et al*., Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016). doi: 10.1038/nature16996; pmid: 26840485
64. N. J. Loman, J. Quick, J. T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015). doi: 10.1038/nmeth.3444; pmid: 26076426
65. J. T. Simpson *et al*., Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017). doi: 10.1038/nmeth.4184; pmid: 28218898
66. M. P. Ball *et al*., Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009). doi: 10.1038/nbt.1533; pmid: 19329998
67. S. Saxonov, P. Berg, D. L. Brutlag, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 1412–1417 (2006). doi: 10.1073/pnas.0510310103; pmid: 16432200
68. M. E. Johnson *et al*., Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001). doi: 10.1038/35097067; pmid: 11586358
69. N. Arnheim, in *Evolution of Genes and Proteins*, M. Nei, R. K. Koehn, Eds. (Sinauer, 1983) pp. 38–61.
70. R. Lyle *et al*., Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. *Genome Res.* **17**, 1690–1696 (2007). doi: 10.1101/gr.6675307; pmid: 17895424
71. S. Luke, R. S. Verma, R. A. Conte, T. Mathews, Molecular characterization of the secondary constriction region (qh) of human chromosome 9 with pericentric inversion. *J. Cell Sci.* **103**, 919–923 (1992). doi: 10.1242/jcs.103.4.919; pmid: 1487504
72. J. C. Barber *et al*., Duplications of proximal 16q flanked by heterochromatin are not euchromatic variants and show no evidence of heterochromatic position effect. *Cytogenet. Genome Res.* **114**, 351–358 (2006). doi: 10.1159/000094225; pmid: 16954678
73. F. I. Sahin *et al*., Chromosome heteromorphisms: An impact on infertility. *J. Assist. Reprod. Genet.* **25**, 191–195 (2008). doi: 10.1007/s10815-008-9216-3; pmid: 18461436
74. M. Codina-Pascual *et al*., Behaviour of human heterochromatic regions during the synapsis of homologous chromosomes. *Hum. Reprod.* **21**, 1490–1497 (2006). doi: 10.1093/humrep/del028; pmid: 16484310
75. A. O. Caglayan, I. Ozyazgan, F. Demiryilmaz, M. T. Ozgun, Are heterochromatin polymorphisms associated with recurrent miscarriage? *J. Obstet. Gynaecol. Res.* **36**, 774–776 (2010). doi: 10.1111/j.1447-0756.2010.01207.x; pmid: 20666944
76. P. F. Madon, A. S. Athalye, F. R. Parikh, Polymorphic variants on chromosomes probably play a significant role in infertility. *Reprod. Biomed. Online* **11**, 726–732 (2005). doi: 10.1016/S1472-6483(10)61691-4; pmid: 16417737
77. S. Minocherhomji *et al*., A case-control study identifying chromosomal polymorphic variations as forms of epigenetic alterations associated with the infertility phenotype. *Fertil. Steril.* **92**, 88–95 (2009). doi: 10.1016/j.fertnstert.2008.05.071; pmid: 18692838
78. Y. Hong, Y. W. Zhou, J. Tao, S. X. Wang, X. M. Zhao, Do polymorphic variants of chromosomes affect the outcome of in vitro fertilization and embryo transfer treatment? *Hum. Reprod.* **26**, 933–940 (2011). doi: 10.1093/humrep/deq333; pmid: 21266453
79. P. Kalantari, H. Sepehri, F. Behjati, Z. O. Ashtiani, M. T. Akbari, Chromosomal studies in infertile men. *Tsitol. Genet.* **35**, 50–54 (2001). pmid: 11944328
80. E. S. Wilch, C. C. Morton, Historical and Clinical Perspectives on Chromosomal Translocations, *Adv. Exp. Med. Biol.* **1044**, 1–14 (2018). doi: 10.1007/978-981-13-0593-1_1; pmid: 29956287
81. D. P. Locke *et al*., BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. *J. Med. Genet.* **41**, 175–182 (2004). doi: 10.1136/jmg.2003.013813; pmid: 14985376
82. J. Ebler *et al*., Pangenome-based genome inference. bioRxiv 2020.11.11.378133v1 [Preprint] Cold Spring Harbor Laboratory (2020); doi: 10.1101/2020.11.11.378133
83. P. Hsieh *et al*., Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366**, eaax2083 (2019). doi: 10.1126/science.aax2083; pmid: 31624180
84. A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery, GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC

Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). doi: 10.1038/nature24277; pmid: 29022597
85. M. L. Dougherty *et al*., Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* **28**, 1566–1576 (2018). doi: 10.1101/gr.237610.118; pmid: 30228200
86. G. A. Logsdon *et al*., The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021). doi: 10.1038/s41586-021-03420-7; pmid: 33828295
87. A. B. Stergachis, B. M. Debo, E. Haugen, L. S. Churchman, J. A. Stamatoyannopoulos, Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020). doi: 10.1126/science.aaz1646; pmid: 32587015
88. N. J. Abdulhay *et al.*, Massively multiplex single-molecule oligonucleosome footprinting. elife 9, e59404 (2020), doi: 10.7554/eLife.59404
89. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). doi: 10.1093/nar/27.2.573; pmid: 9862982
90. A.F.A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0. (2015). www.repeatmasker.org.
91. K. Krasheninnikova, M. Diekhans, J. Armstrong, A. Dievskii, B. Paten, S. O'Brien, halSynteny: A fast, easy-to-use conserved synteny block construction method for multiple whole-genome alignments. *Gigascience***9**, giaa047 (2020). doi: 10.1093/gigascience/giaa047; pmid: 32463100
92. F. Hach *et al*., mrsFAST: A cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7**, 576–577 (2010). doi: 10.1038/nmeth0810-576; pmid: 20676076
93. H. Killick, F. Eckley, P. Lee, changepoint: R package version 0.4. (2016). https://cran.r-project.org/web/packages/changepoint/changepoint.pdf.
94. A. Shumate, S. L. Salzberg, Liftoff: Accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021). doi: 10.1093/bioinformatics/btaa1016; pmid: 33320174
95. G. Pertea, M. Pertea, G. F. F. Utilities, GffRead and GffCompare. *F1000 Res.* **9** (2020) [version 2; peer review: 3 approved]. doi: 10.12688/f1000research.23297.2; pmid: 32489650
96. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). doi: 10.1016/S0022-2836(05)80360-2; pmid: 2231712
97. M. F. Cardone *et al*., Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs. *Genome Biol.* **7**, R91 (2006). doi: 10.1186/gb-2006-7-10-r91; pmid: 17040560
98. A. D. Sanders, E. Falconer, M. Hills, D. C. J. Spierings, P. M. Lansdorp, Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017). doi: 10.1038/nprot.2017.029; pmid: 28492527
99. Standing Committee on Human Cytogenetic Nomenclature, *ISCN 1995: An International System for Human Cytogenetic Nomenclature: Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature* (Karger Medical and Scientific Publishers, 1995).
100. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
101. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi: 10.1093/bioinformatics/bty191; pmid: 29750242
102. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002). doi: 10.1093/nar/gkf436; pmid: 12136088
103. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). doi: 10.1093/bioinformatics/btu033; pmid: 24451623
104. M. P. Berger, P. J. Munson, A novel randomized iterative strategy for aligning multiple protein sequences. *Comput. Appl. Biosci.* **7**, 479–484 (1991). doi: 10.1093/bioinformatics/7.4.479; pmid: 1747779
105. O. Gotoh, Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.* **9**, 361–370 (1993). doi: 10.1093/bioinformatics/9.3.361; pmid: 8324637
106. A. Gershman *et al*., Epigenetic Patterns in a Complete Human Genome. *Science* **376**, eabj5089 (2022).doi: 10.1126/science.abj5089
107. C. Jain *et al*., Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36** (Suppl. 1), i111–i118 (2020). doi: 10.1093/bioinformatics/btaa435; pmid: 32657365
108. J. D. Parsons, Miropeats: Graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995). pmid: 8808577
109. M. Lawrence *et al*., Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013). doi: 10.1371/journal.pcbi.1003118; pmid: 23950696
110. H. Wickham *et al*., Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019). doi: 10.21105/joss.01686
111. B. Gel, E. Serra, karyoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017). doi: 10.1093/bioinformatics/btx346; pmid: 28575171
112. Z. Gu, L. Gu, R. Eils, M. Schlesner, B. Brors, circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014). doi: 10.1093/bioinformatics/btu393; pmid: 24930139
113. J. Köster, S. Rahmann, Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012). doi: 10.1093/bioinformatics/bts480; pmid: 22908215
114. J. Köster, S. Rahmann, Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600 (2018). doi: 10.1093/bioinformatics/bty350; pmid: 29788404
115. F. Mölder *et al*., Sustainable data analysis with Snakemake. *F1000 Res.* **10**, 33 (2021). doi: 10.12688/f1000research.29032.2; pmid: 34035898
116. M. R. Vollger, Assemblies and data generated for "Segmental duplications and their variation in a complete human genome" (2021), https://zenodo.org/record/4726156.
117. M. R. Vollger, mrvollger/Data-Analysis-for-SDs-in-T2T-CHM13, *Zenodo* (2021); https://zenodo.org/record/5498994).
118. M. R. Vollger, Version v0.3, Interconnected snakemake workflows for annotation and analysis of assemblies, *Zenodo* (2021); https://zenodo.org/record/5499093.

ACKNOWLEDGMENTS

The authors thank T. Brown and A. Lo for help in editing this manuscript. **Funding:** This work was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (S.N., S.K., and A.M.P.), grants from the U.S. National Institutes of Health (NIH grants 5R01HG002385, 5U01HG010971, and 1U01HG010973 to E.E.E.; 1R01HG011274 to K.H.M.; 5R01HG009190 to W.T.; and U41HG007234 to M.D.), and a grant from Futuro in Ricerca (2010-RBFR103CE3 to M.V.). E.E.E. is an investigator of Howard Hughes Medical Institute. **Author contributions:** Identification of SDs in T2T-CHM13 and analysis: M.R.V. PacBio genome sequence generation: K.M.M., A.P.L., and K.H. FISH experiments and analysis: L.M., M.V., M.R.V., and E.E.E. Iso-Seq analysis: P.C.D., M.R.V., and R.L. *TBC1D3* analysis: X.G. and M.R.V. Copy number analysis: M.R.V. and W.T.H. inversion analysis: D.P. and M.R.V. T2T-CHM13 assembly generation: S.N., S.K., and A.M.P. Refinement of SD annotations near centromeres: K.H.M. and M.R.V. UCSC browser: M.D., W.T.H., and M.R.V. Methylation analysis: M.R.V., A.G., W.T., and E.E.E. Analysis of regions with genomic instability: M.R.V. and A.S. Organization of tables: M.R.V., P.C.D., and X.G. Organization of supplementary material: M.R.V. Display items: M.R.V., X.G., and P.C.D. Manuscript writing: M.R.V., E.E.E., and X.G. with input from all authors. **Competing interests:** E.E.E. is an SAB member of Variant Bio, Inc; S.K. and K.H.M. received travel funds to speak at events hosted by Oxford Nanopore Technologies; W.T. has licensed two patents to Oxford Nanopore Technologies (US 8748091 and 8394584). **Data and materials availability:** CHM13hTERT cells were obtained for research use via a material transfer agreement with the University of Pittsburg. PacBio HiFi data has been deposited into NCBI Sequence Read Archive (SRA) under the following accessions: SRX7897688, SRX7897687, SRX7897686, and SRX7897685 for CHM13; SRR14407677 and SRR14407676 for CHM1; SRR10382244, SRR10382245, SRR10382248, and SRR10382249 for HG002; PRJNA540705 for NA12878; PRJEB36100 for HG00733 and HG00514; ERX4787609, ERX4787607, ERX4787606, ERX4782632, and ERX4781730 for NA19240; PRJNA701308 for HG01109, HG01243, HG02080, HG02723, HG03125, and HG03492; and PRJNA659034 and PRJNA691628 for all nonhuman primate samples. The human lymphoblastoid cell lines GM24385, GM19240, HG00514, and HG00733 used in the FISH experiments were obtained from Coriell. The T2T-CHM13 v1.0 assembly can be found on NCBI (GCA_009914755.2) and all associated read data were uploaded to SRA under the BioProject identifier PRJNA686988. Table S13 contains the accession information for all Iso-Seq data used in the paper. The canonical rDNA unit used to estimate copy number can be found at the NCBI nucleotide repository (KY962518.1). Human and nonhuman primate genome assemblies, SD annotations, methylation data, and Liftoff gene models can be found at Zenodo (*116*). Code for Snakemake pipelines, data analysis, and figure generation are also available at Zenodo (*117*, *118*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj6965
Extended Materials and Methods
Figs. S1 to S25
Tables S1 to S14

26 May 2021; accepted 13 December 2021
10.1126/science.abj6965

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

Complete genomic and epigenetic maps of human centromeres

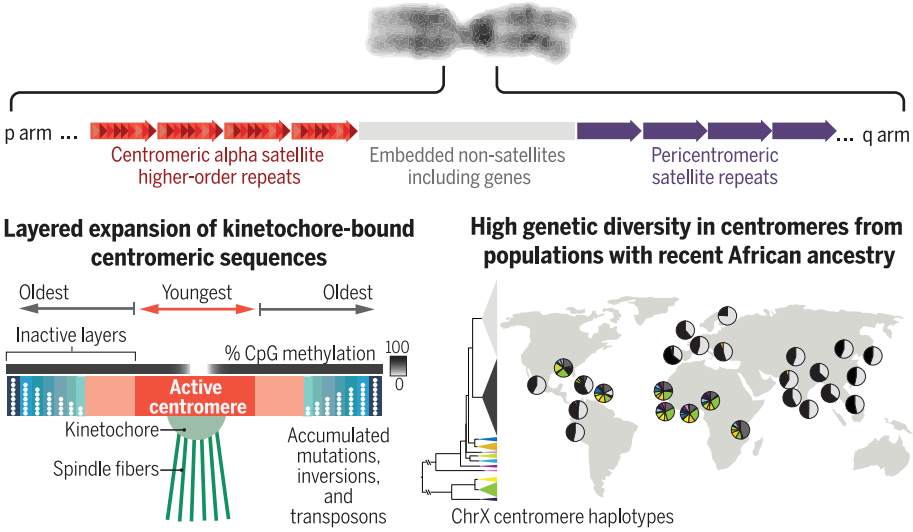
Nicolas Altemose, Glennis A. Logsdon†, Andrey V. Bzikadze†, Pragya Sidhwani†, Sasha A. Langley†, Gina V. Caldas†, Savannah J. Hoyt, Lev Uralsky, Fedor D. Ryabov, Colin J. Shew, Michael E. G. Sauria, Matthew Borchers, Ariel Gershman, Alla Mikheenko, Valery A. Shepelev, Tatiana Dvorkina, Olga Kunyavskaya, Mitchell R. Vollger, Arang Rhie, Ann M. McCartney, Mobin Asri, Ryan Lorig-Roach, Kishwar Shafin, Julian K. Lucas, Sergey Aganezov, Daniel Olson, Leonardo Gomes de Lima, Tamara Potapova, Gabrielle A. Hartley, Marina Haukness, Peter Kerpedjiev, Fedor Gusev, Kristof Tigyi, Shelise Brooks, Alice Young, Sergey Nurk, Sergey Koren, Sofie R. Salama, Benedict Paten, Evgeny I. Rogaev, Aaron Streets, Gary H. Karpen, Abby F. Dernburg, Beth A. Sullivan, Aaron F. Straight, Travis J. Wheeler, Jennifer L. Gerton, Evan E. Eichler, Adam M. Phillippy, Winston Timp, Megan Y. Dennis, Rachel J. O'Neill, Justin M. Zook, Michael C. Schatz, Pavel A. Pevzner, Mark Diekhans, Charles H. Langley, Ivan A. Alexandrov*, Karen H. Miga*

INTRODUCTION: To faithfully distribute genetic material to daughter cells during cell division, spindle fibers must couple to DNA by means of a structure called the kinetochore, which assembles at each chromosome's centromere. Human centromeres are located within large arrays of tandemly repeated DNA sequences known as alpha satellite (αSat), which often span millions of base pairs on each chromosome. Arrays of αSat are frequently surrounded by other types of tandem satellite repeats, which have poorly understood functions, along with nonrepetitive sequences, including transcribed genes. Previous genome sequencing efforts have been unable to generate complete assemblies of satellite-rich regions because of their scale and repetitive nature, limiting the ability to study their organization, variation, and function.

RATIONALE: Pericentromeric and centromeric (peri/centromeric) satellite DNA sequences have remained almost entirely missing from the assembled human reference genome for the past 20 years. Using a complete, telomere-to-telomere (T2T) assembly of a human genome, we developed and deployed tailored computational approaches to reveal the organization and evolutionary patterns of these satellite arrays at both large and small length scales. We also performed experiments to map precisely which αSat repeats interact with kinetochore proteins. Last, we compared peri/centromeric regions among multiple individuals to understand how these sequences vary across diverse genetic backgrounds.

RESULTS: Satellite repeats constitute 6.2% of the T2T-CHM13 genome assembly, with αSat representing the single largest component (2.8% of the genome). By studying the sequence relationships of αSat repeats in detail across each centromere, we found genome-wide evidence that human centromeres evolve through "layered expansions." Specifically, distinct repetitive variants arise within each centromeric region and expand through mechanisms that resemble successive tandem duplications, whereas older flanking sequences shrink and diverge over time. We also revealed that the most recently expanded repeats within each αSat array are more likely to interact with the inner kinetochore protein Centromere Protein A (CENP-A), which coincides with regions of reduced CpG methylation. This suggests a strong relationship between local satellite repeat expansion, kinetochore positioning, and DNA hypomethylation. Furthermore, we uncovered large and unexpected structural rearrangements that affect multiple satellite repeat types, including active centromeric αSat arrays. Last, by comparing sequence information from nearly 1600 individuals' X chromosomes, we observed that individuals with recent African ancestry possess the greatest genetic diversity in the region surrounding the centromere, which sometimes contains a predominantly African αSat sequence variant.

CONCLUSION: The genetic and epigenetic properties of centromeres are closely interwoven through evolution. These findings raise important questions about the specific molecular mechanisms responsible for the relationship between inner kinetochore proteins, DNA hypomethylation, and layered αSat expansions. Even more questions remain about the function and evolution of non-αSat repeats. To begin answering these questions, we have produced a comprehensive encyclopedia of peri/centromeric sequences in a human genome, and we demonstrated how these regions can be studied with modern genomic tools. Our work also illuminates the rich genetic variation hidden within these formerly missing regions of the genome, which may contribute to health and disease. This unexplored variation underlines the need for more T2T human genome assemblies from genetically diverse individuals. ■



Gapless assemblies illuminate centromere evolution. (**Top**) The organization of peri/centromeric satellite repeats. (**Bottom left**) A schematic portraying (i) evidence for centromere evolution through layered expansions and (ii) the localization of inner-kinetochore proteins in the youngest, most recently expanded repeats, which coincide with a region of DNA hypomethylation. (**Bottom right**) An illustration of the global distribution of chrX centromere haplotypes, showing increased diversity in populations with recent African ancestry.

The list of author affiliations is available in the fulll article online.
***Corresponding author. Email: livanalx@hotmail.com (I.A.A.); khmiga@ucsc.edu (K.H.M.)**
†These authors contributed equally to this work.
Cite this article as N. Altemose *et al.*, *Science* 376, eabl4178 (2022). DOI: 10.1126/science.abl4178

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abl4178

RESEARCH ARTICLE

HUMAN GENOMICS

Complete genomic and epigenetic maps of human centromeres

Nicolas Altemose[1], Glennis A. Logsdon[2]†, Andrey V. Bzikadze[3]†, Pragya Sidhwani[4]†, Sasha A. Langley[1]†, Gina V. Caldas[1]†, Savannah J. Hoyt[5,6], Lev Uralsky[7,8], Fedor D. Ryabov[9], Colin J. Shew[10], Michael E. G. Sauria[11], Matthew Borchers[12], Ariel Gershman[13], Alla Mikheenko[14], Valery A. Shepelev[8], Tatiana Dvorkina[14], Olga Kunyavskaya[14], Mitchell R. Vollger[2], Arang Rhie[15], Ann M. McCartney[15], Mobin Asri[16], Ryan Lorig-Roach[16], Kishwar Shafin[16], Julian K. Lucas[16], Sergey Aganezov[17]‡, Daniel Olson[18], Leonardo Gomes de Lima[12], Tamara Potapova[12], Gabrielle A. Hartley[5,6], Marina Haukness[16], Peter Kerpedjiev[19], Fedor Gusev[8], Kristof Tigyi[16,20], Shelise Brooks[21], Alice Young[21], Sergey Nurk[15], Sergey Koren[15], Sofie R. Salama[16,20], Benedict Paten[16,22], Evgeny I. Rogaev[7,8,23,24], Aaron Streets[25,26], Gary H. Karpen[1,27], Abby F. Dernburg[1,28,20], Beth A. Sullivan[29], Aaron F. Straight[4], Travis J. Wheeler[18], Jennifer L. Gerton[12,30], Evan E. Eichler[2,20], Adam M. Phillippy[15], Winston Timp[13,31], Megan Y. Dennis[10], Rachel J. O'Neill[5,6], Justin M. Zook[32], Michael C. Schatz[17], Pavel A. Pevzner[33], Mark Diekhans[16], Charles H. Langley[34], Ivan A. Alexandrov[8,14,35]*, Karen H. Miga[16,22]*

Existing human genome assemblies have almost entirely excluded repetitive sequences within and near centromeres, limiting our understanding of their organization, evolution, and functions, which include facilitating proper chromosome segregation. Now, a complete, telomere-to-telomere human genome assembly (T2T-CHM13) has enabled us to comprehensively characterize pericentromeric and centromeric repeats, which constitute 6.2% of the genome (189.9 megabases). Detailed maps of these regions revealed multimegabase structural rearrangements, including in active centromeric repeat arrays. Analysis of centromere-associated sequences uncovered a strong relationship between the position of the centromere and the evolution of the surrounding DNA through layered repeat expansions. Furthermore, comparisons of chromosome X centromeres across a diverse panel of individuals illuminated high degrees of structural, epigenetic, and sequence variation in these complex and rapidly evolving regions.

For two decades, genome sequencing and assembly efforts have excluded an estimated 5 to 10% of the human genome, most of which is found in and around each chromosome's centromere (*1*, *2*). These large regions contain highly repetitive DNA sequences, which impede assembly from short DNA sequencing reads (*1*, *3*). Centromeres function to ensure proper distribution of genetic material to daughter cells during cell division, making them critical for genome stability, fertility, and healthy development (*4*). Nearly everything known about the sequence composition of human centromeres and their surrounding regions, called pericentromeres, stems from individual experimental observations (*5*–*8*), low-resolution classical mapping techniques (*9*, *10*), analyses of unassembled sequencing reads (*11*–*14*), or recent studies of centromeric sequences on individual chromosomes (*15*–*17*). As a result, millions of bases in the pericentromeric and centromeric regions (hereafter peri/centromeres) remain largely uncharacterized and omitted from contemporary genetic and epigenetic studies. Recently, long-read sequencing and assembly methods enabled the Telomere-to-Telomere Consortium to produce a complete assembly of an entire human genome (T2T-CHM13) (*2*). This effort relied on careful measures to correctly assemble, polish, and validate entire peri/centromeric repeat arrays (*2*, *18*). By deeply characterizing these recently assembled sequences, we present a high-resolution, genome-wide atlas of the sequence content and organization of human peri/centromeric regions.

Centromeres provide a robust assembly point for kinetochore proteins, which physically couple each chromosome to spindle fibers during cell division (*4*). Compromised centromere function can lead to nondisjunction, a major cause of somatic and germline disease (*19*). In many eukaryotes, the centromere is composed of tandemly repeated DNA sequences, called satellite DNA, but these sequences differ widely among species (*20*, *21*). In humans, centromeres are defined by alpha satellite DNA (αSat), an AT-rich repeat family composed of ~171 base pair (bp) monomers, which can occur as different subtypes repeated in a head-to-tail orientation for millions of bases (*22*, *23*). In the largest αSat arrays, different monomer subtypes belong to higher-order repeats (HORs); for example, monomer subtypes a, b, and c can repeat as abc-abc-abc (*24*, *25*). HOR arrays tend to be large and highly homogeneous, often containing thousands of nearly identical HOR units. However, kinetochore proteins associate with only a subset of these HOR units, usually within the largest HOR array on each chromosome, which is called the active array (*25*, *26*). Distinct αSat HOR arrays tend to differ in sequence and structure (*27*, *28*), and like other satellite repeats, they evolve rapidly through mechanisms such as unequal crossover and gene conversion (*29*, *30*). Consequently, satellite arrays frequently expand and contract in size and

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [3]Graduate Program in Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA, USA. [4]Department of Biochemistry, Stanford University, Stanford, CA, USA. [5]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA. [6]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [7]Sirius University of Science and Technology, Sochi, Russia. [8]Vavilov Institute of General Genetics, Moscow, Russia. [9]Moscow Polytechnic University, Moscow, Russia. [10]Genome Center, MIND Institute, and Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, Davis, Davis, CA, USA. [11]Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [12]Stowers Institute for Medical Research, Kansas City, MO, USA. [13]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. [14]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia. [15]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [16]UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [17]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [18]Department of Computer Science, University of Montana, Missoula, MT. USA. [19]Reservoir Genomics, Oakland, CA. [20]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [21]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [22]Department of Biomolecular Engineering, University of California Santa Cruz, CA, USA. [23]Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA. [24]Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia. [25]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA. [26]Chan Zuckerberg Biohub, San Francisco, CA, USA. [27]BioEngineering and BioMedical Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [28]Institute for Quantitative Biosciences (QB3), University of California, Berkeley, Berkeley, CA, USA. [29]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC, USA. [30]University of Kansas Medical School, Department of Biochemistry and Molecular Biology and Cancer Center, University of Kansas, Kansas City, KS, USA. [31]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [32]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA. [33]Department of Computer Science and Engineering, University of California at San Diego, San Diego, CA, USA. [34]Department of Evolution and Ecology, University of California Davis, Davis, CA, USA. [35]Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia.

***Corresponding author. Email: iivanalx@hotmail.com (I.A.A.); khmiga@ucsc.edu (K.H.M.)**

†These authors contributed equally to this work.

‡Present address: Oxford Nanopore Technologies, Oxford, UK.

generate a high degree of interindividual polymorphism (29–31). Active αSat HOR arrays are flanked by inactive pericentromeric regions, which often include (i) smaller arrays of diverged αSat monomers that lack HORs (27, 32); (ii) transposable elements (TEs); (iii) segmental duplications, which sometimes include expressed genes (33, 34); and (iv) non-αSat satellite repeat families (35), which have poorly understood functions. Given the opportunity to explore these regions in a complete human genome assembly, we investigated the precise localization of inner kinetochore proteins within large αSat arrays and surveyed sequence-based trends in the structure, function, variation, and evolution of peri/centromeric DNA.

A comprehensive map of peri/centromeric satellite DNA

Human peri/centromeric satellite DNAs represent 6.2% of the T2T-CHM13v1.1 genome (~189.9 Mb) (tables S1 and S2 and figs. S1 and S2). Nearly all of this sequence remains unassembled or belongs to simulated arrays called reference models (12) in the current GRCh38/hg38 reference sequence (hereafter, hg38), including pericentromeric satellite DNA families that extend into each of the five acrocentric short arms. From decades of individual observations, a framework for the overall organization of a typical human peri/centromeric region has been proposed (Fig. 1A). By annotating and examining the repeat content of these regions in the CHM13 assembly (Fig. 1, B and C; figs. S1 and S2; table S1; and database S1), we tested and largely confirmed this broad framework genome-wide at base-pair resolution. However, we uncovered unexpected large-scale structural rearrangements and previously unresolved satellite variants (fig. S1).

All centromeric regions contain long tracts, or arrays, of tandemly repeated αSat monomers (85.2 Mb total genome-wide) (Fig. 1, B and C) (36). Most chromosomes also contain classical human satellites 2 and/or 3 (HSat2 and HSat3, totaling 28.7 and 47.6 Mb, respectively). HSat2 and HSat3 are derived from the simple repeat (CATTC)n and constitute the largest contiguous satellite arrays found in the human genome, including a 27.6-Mb array on chromosome 9 (chr9) (Fig. 1, B and C) (11, 37, 38). Furthermore, two distinct satellite DNA families constitute the most AT-rich regions of the genome (37, 39), which we refer to as HSat1A (13.4 Mb total, found mostly on chr3, chr4, and chr13) and HSat1B (found mostly on chrY, with 1.2 Mb on the acrocentrics) (table S1). Two additional large families, beta satellite (βSat; 7.7 Mb total) and gamma satellite (γSat; 630 kb total), are more GC-rich than αSat and contain dense CpG methylation (fig. S3). All remaining annotated pericentromeric satellite DNAs total 5.6 Mb, with 1.2 Mb representing previously unresolved types of satellite DNA (table S1 and fig. S2) (40). Nonsatellite bases between satellite arrays and extending into the p-arms and q-arms are considered "centric transition" regions, which largely represent long tracts of segmental duplications, including expressed genes (Fig. 1C and fig. S1) (2, 41, 42). These annotations provide a complete and detailed map of all the peri/centromeric sequences in a human genome.

Complete assessment of αSat substructure and genomic organization

To better understand the organization and evolution of αSat arrays, we generated a genome-wide database of αSat monomers (42). We grouped these monomers into distinct classes belonging to 20 suprachromosomal families

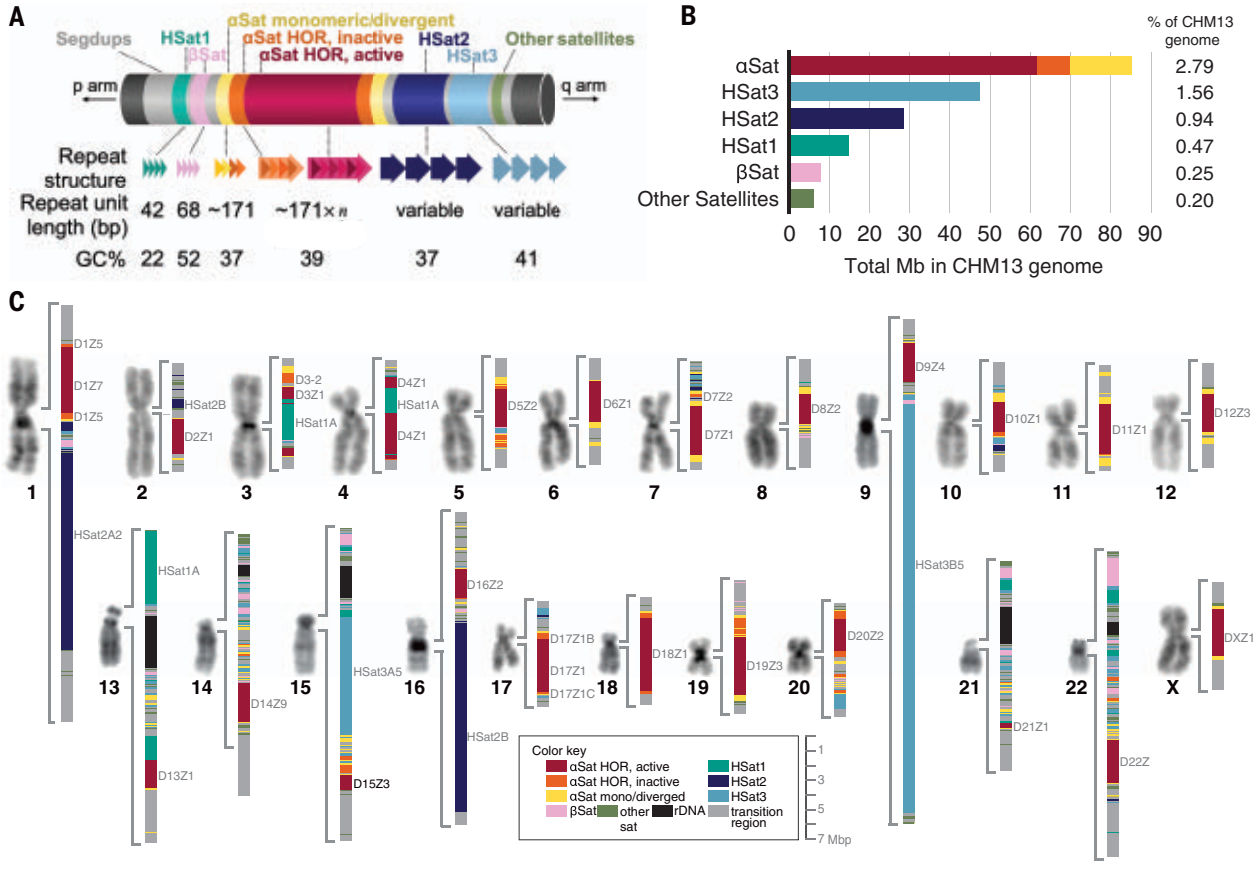


Fig. 1. Overview of all peri/centromeric regions in CHM13. (**A**) Schematic of a generalized human peri/centromeric region, identifying major sequence components and their properties (not to scale). HSat2,3 repeat unit lengths vary by genomic region. (**B**) Barplots of the total lengths of each major satellite family genome-wide. (**C**) Micrographs of representative 4′,6-diamidino-2-phenylindole (DAPI)–stained chromosomes from CHM13 metaphase spreads, next to a color-coded map of peri/centromeric satellite DNA arrays [available as a browser track (database S1)]. Large satellite arrays are labeled.

(SFs) (tables S2 and S3 and database S2) (*32*, *43*, *44*) and identified 80 different HOR arrays and more than 1000 different monomers in HORs across the genome (70 Mb total) (table S4 and database S3) (*38*). Although 18 out of 23 chromosomes contain multiple, distinct HOR arrays (up to nine), only one HOR array per chromosome is active, meaning that it consistently associates with the kinetochore across individuals (Fig. 1, B and C, and table S4) (*25*). The active array on each chromosome ranges in size from 4.8 Mb on chr18 down to 340 kb on chr21, which is near the low end of the estimated αSat size range for chr21 among healthy individuals (*45*). Inactive HOR arrays tend to be much smaller (8 Mb total genome-wide) (Fig. 1, B and C, and table S4). Adjacent to many homogeneous HOR arrays are regions of divergent αSat HORs, in which HOR periodicity is somewhat or even completely eroded (*44*), as well as highly divergent αSat monomeric layers that lack HOR structure (*32*), totaling 15.2 Mb in CHM13.

The completeness and quality of the T2T-CHM13 assembly also allowed us to resolve HOR arrays that are highly similar between chromosomes, such as those found on chromosomes 13/21, 14/22, and 1/5/19, which have confounded studies in the past (*7*, *27*, *36*). Within these arrays, we identified chromosome-specific sequence variants and patterns of structural variants, which we validated using flow-sorted chromosome libraries for the chromosome 1/5/19 arrays (fig. S4) (*42*). This enabled us to infer their evolutionary history, providing evidence that the 1/5/19 HOR first originated on chr19 (*42*).

Large structural rearrangements in peri/centromeric regions

Producing complete maps of peri/centromeric regions revealed the large-scale organization of satellite DNAs and their embedded non-satellite sequences, including TEs and genes (Fig. 2, A to E). Although divergent αSats contain many inversions (*46*) and TE insertions (*47*), such events within active HOR arrays are unexpected because they were considered to be homogeneous (*48*, *49*). Quantifying strand inversions across entire satellite arrays revealed unexpected anomalies (Fig. 2, A, B, and E; fig. S1; table S5; and databases S4 and S5). For example, we uncovered a 1.7-Mb inversion inside the active αSat HOR array on chr1 (Fig. 2A), along with inversions in inactive HOR arrays on chr3, chr16, and chr20 (figs. S1 and S5). Unexpectedly, the large pericentromeric HSat3 array on chr9 and the βSat arrays on chr1 and the acrocentrics contain more than 200 inversion breakpoints (Fig. 2A and fig. S5), whereas in other arrays inversions are rare.

Apart from inversions, two multimegabase HSat1A arrays appear to have inserted in and split the active HOR arrays on chr3 and chr4 (fig. S1 and table S6). We also found evidence for an ancient duplication event that predated African ape divergence and involved a large segment of the ancient chr6 centromere plus about 1 Mb of adjacent p-arm sequence (database S6) (*42*). This duplication created a different centromere locus that hosts the current active chr6 HOR array.

We also assigned HSat2 and HSat3 arrays to their respective sequence subfamilies from (*11*) and found previously unresolved chromosomal localizations of several HSat3 subfamilies (such as HSat3B1 on chr17) (Fig. 2, B and D). However, we also noticed a lack of HSat3B2 on chr1, contrary to expectations based on different cell lines (*11*), implying a large deletion of this subfamily on chr1 in CHM13.

To better understand whether these satellite inversions, insertions, and deletions are common outside of the CHM13 genome, we searched for them across 16 haplotype-resolved draft diploid assemblies of genetically diverse individuals from the Human Pangenome Reference Consortium (HPRC) (*50*). This revealed that the inversion in the active αSat HOR array on chr1 is polymorphic across individuals and evident in about half of ascertainable haplotypes (11 of 24) (fig. S6). However, the HSat1A insertions on chr3 and chr4 are present in all ascertainable haplotypes (32 of 32 and 33 of 33, respectively) (fig. S7). Furthermore, CHM13's missing chr1 HSat3B2 array is contained within a 400-kb polymorphic deletion, which we detected in 29% (8 of 28) of haplotypes examined (Fig. 2A and fig. S7). Thus, these peri/centromeric structural rearrangements are not specific to the CHM13 genome but are present either variably or fixed across humans.

TE and gene interspersion in peri/centromeric regions

Like inversions and insertions, TEs are virtually absent from homogeneous HOR arrays but are enriched in divergent αSat in CHM13 (Fig. 2E and database S7) (*47*, *51*). The CHM13 assembly also revealed regions where combinations of TE sequences have been tandemly duplicated, forming "composite satellites" [described in (*40*)]. We also found that other satellites—such as HSat1A/B, HSat3, and βSat—often include fragments of ancient TEs as part of their repeating units, a phenomenon rarely observed in αSat HOR arrays (Fig. 2, A, B, and E, and fig. S8) (*39*, *52*, *53*).

We also compared our pericentromeric maps with gene annotations reported for T2T-CHM13, revealing 676 gene and pseudogene annotations embedded between large satellite arrays, including 23 protein coding genes and 141 long noncoding RNAs (lncRNAs) (excluding the acrocentric short arms) (table S7 and database S8) (*2*). One region on chr17, located between the large HSat3 and αSat arrays (Fig. 2C), contains two protein-coding genes: *KCNJ17*, which encodes a disease-associated potassium channel in muscle cells (*54*), and *UBBP4*, which encodes a functional ubiquitin variant that may play a role in regulation of nuclear lamins (*55*). *KCNJ17* is missing from GRCh38, which likely has caused inaccurate and missed variant calls in paralogous genes *KCNJ12* and *KCNJ18* (*56*). This region also contains a lncRNA annotation (*LINC02002*), which starts inside an SST1 element and continues into an adjacent 33-kb array of divergent αSat. Furthermore, we identified a processed paralog of an apoptosis-related protein-coding gene, *BCLAF1* (BCL2 Associated Transcription Factor 1), as part of a segmental duplication embedded within an inactive αSat HOR array on chr16 (fig. S9).

The fine repeat structure of satellite DNA arrays

To further chart the structure of peri/centromeric regions at high resolution, we compared individual repeat units within and between different satellite arrays. We decomposed each αSat HOR array first into individual monomers and then into entire HORs, revealing the positions of full-size canonical HORs and structural variant HORs resulting from insertions or deletions (databases S9 and S10) (*42*). Whereas some chromosomes, such as chr7, are composed almost entirely of canonical HOR units, others, such as chr10, contain many structural variant HOR types, with high variation in the relative frequency of these structural variants across individuals (Fig. 3A and fig. S10).

Unlike αSat, some families such as HSat2 and HSat3 have inconsistent or unknown repeat unit lengths and often contain an irregular hierarchy of smaller repeating units. We refer to these repeat units as nested tandem repeats (NTRs), a more general term than HORs, which are composed of discrete numbers of monomers of similar lengths. To expand our ability to annotate repeat structure within assembled satellite DNA arrays, we developed NTRprism, an algorithm to discover and visualize satellite repeat periodicity [(*42*), similar to (*57*)]. Using this tool, we discovered HORs in HSat1 and βSat arrays, as well as NTRs in multiple HSat2,3 arrays, such as a 2235-bp repeating unit in the HSat3B1 array on chr17 (Fig. 3B and fig. S11). We also applied this tool in smaller windows across individual arrays, showing that repeat periodicity can vary across an array, which is consistent with NTRs evolving and expanding hyper-locally in some cases (fig. S11).

Genome-wide evidence of layered expansions in centromeric arrays

The T2T-CHM13 assembly also provides an opportunity to examine how peri/centromeric sequences evolve. A "layered expansion" model for centromeric αSat evolution has been

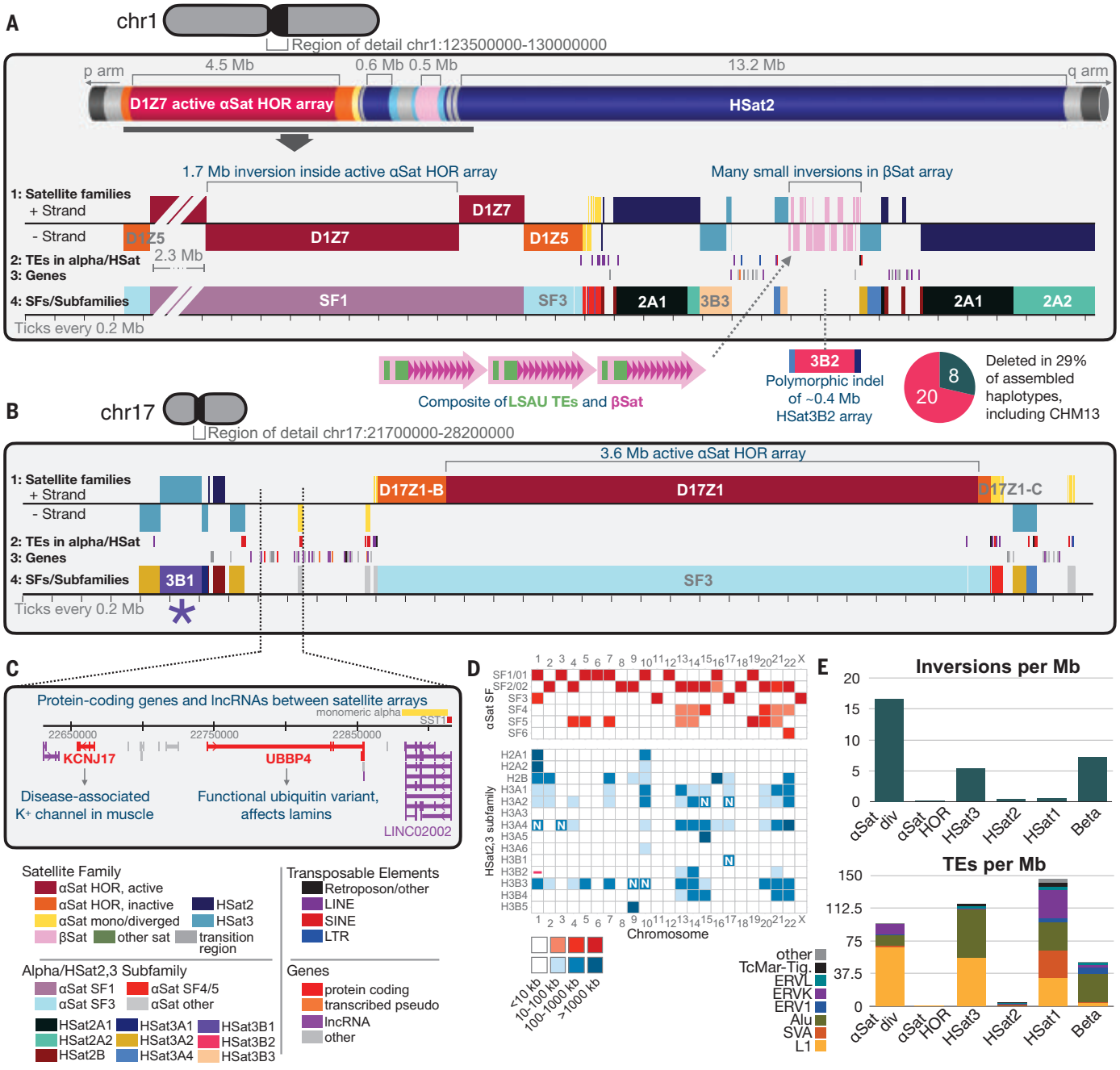


Fig. 2. Structural rearrangements, genes, and TEs in peri/centromeric regions. (**A**) The peri/centromeric region of chr1 (cylindrical schematic at top), zooming into the transition region between the large αSat and HSat2 arrays (tracks 1 to 4). Track 1, satellite families (color key at bottom left), with vertical placement indicating the strand with canonical satellite repeat polarity. Track 2, positions of TEs overlapping αSat or HSat1,2,3, colored by TE type. Track 3, annotated transcription start sites, colored by gene type. Track 4, HSat2,3 subfamily assignments [as in (*11*)] and αSat SF assignments, with large arrays labeled. (**B**) As in (A) but for chr17, with the previously unresolved HSat3B1 array indicated with an asterisk. (**C**) Gene annotations between the αSat and HSat3 arrays on chr17. (**D**) Heatmap showing the major and minor localizations of each αSat HOR SF (top; red) and each HSat2,3 subfamily (bottom; blue). "N" indicates localizations not described in (*11*). Dash "—" indicates the chr1 HSat3B2 array deleted in CHM13. HSat3A3 and 3A6 are predominantly found on chrY (not in CHM13). (**E**) Barplots illustrate the number of inversion breakpoints (strand switches) or the number and type of TEs detected per megabase within different satellite families. div, divergent αSat (dHORs + monomeric).

hypothesized on the basis of limited observations of the most diverged αSat sequences in the human genome [reviewed in (*36*)]. This model postulates that distinct αSat repeats periodically emerge and expand within an active array, displacing the older repeats sideways and becoming the site of kinetochore assembly. The newer, expanding αSat sequences can originate from within the same array (*32*) or from a different array (intra- versus interarray seeding) (*58*, *59*). As this process iterates over time, the displaced sequences form distinct layers that flank the active centromere with mirror symmetry (Fig. 3C), and these flanking layers rapidly shrink and decay (*17*, *32*, *44*). We used the T2T-CHM13 assembly to infer the evolutionary dynamics of αSat repeat arrays genome-wide to test the layered

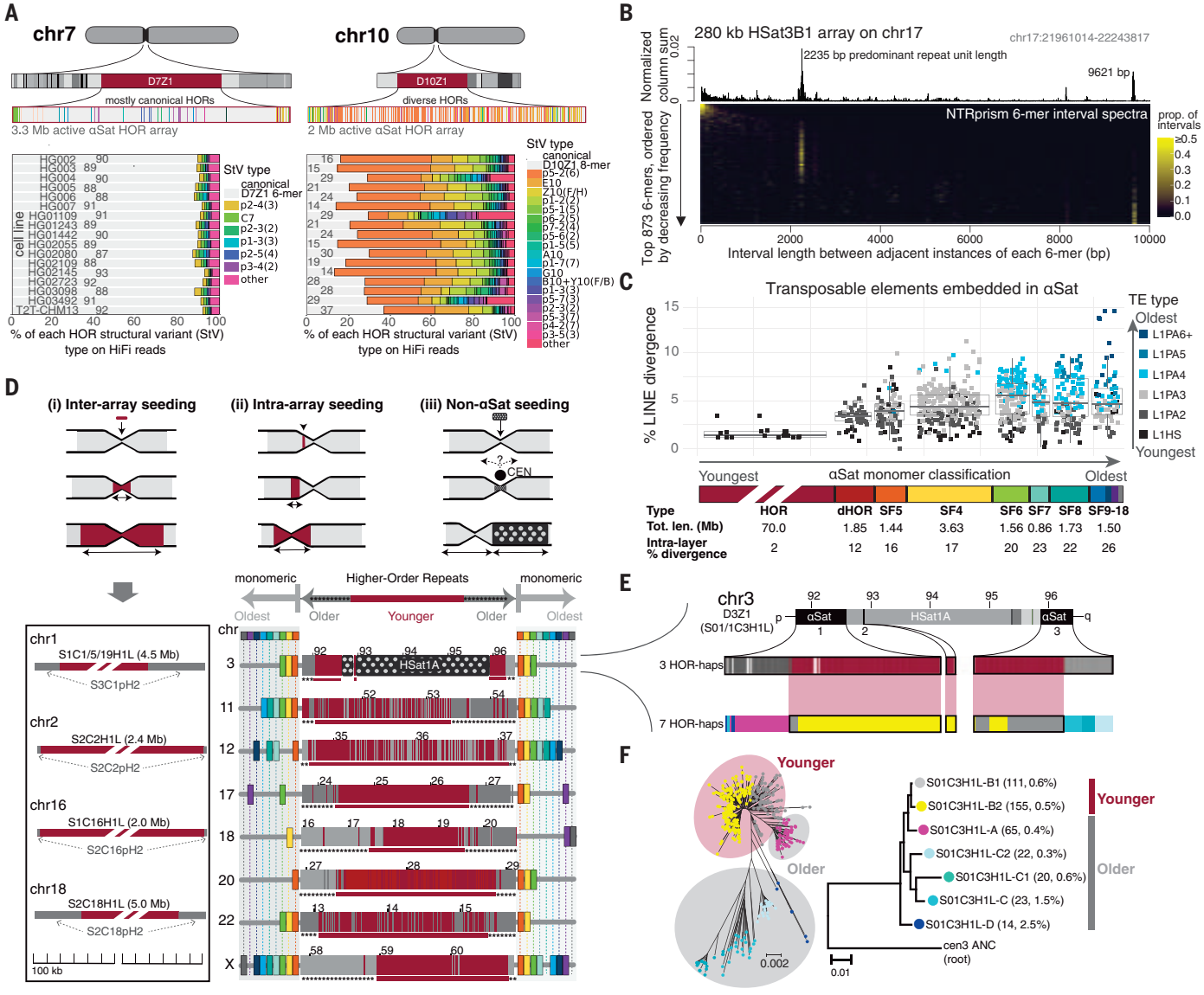


Fig. 3. Genome-wide evidence of layered expansions in centromeric αSat arrays. (**A**) (Top) HOR structural variant positions across the active αSat arrays on chr7 and chr10 (gray, canonical HORs; other colors, structural variants). (Bottom) Percentages of HOR structural variant types on HiFi sequencing reads from 16 HPRC cell lines. Variant nomenclature is described in (*42*); canonical HOR percentages are listed on the plot. (**B**) Repeat periodicities identified with NTRprism for the HSat3B1 array on chr17. (**C**) Comparison of the age and divergence of LINE TEs embedded in different αSat SF layers. (**D**) (i) Four centromeres in which an active HOR array of distinct origin appears to have expanded within a now-inactive HOR array. (ii) and (iii) Monomeric SFs (rainbow colors) surrounding active HOR arrays on eight chromosomes, with major HOR-haps shown (k = 2 to 3). Red, younger, emphasized below with red rectangles; gray, older, emphasized below with asterisks. (**E**) Zoomed-in view of chr3 αSat HOR arrays, divided into finer symmetrical HOR-haps (k = 7). (**F**) (Left) Minimum evolution tree showing the phylogenetic relationships between all HORs, colored by fine (k = 7) HOR-hap assignments. Red and gray ellipses group major HOR-hap divisions into younger and older variants, respectively (*42*). (Right) Phylogenetic tree built from HOR-hap consensus sequences derived from branches in the left tree, rooted with a reconstructed ancestral cen3 active HOR sequence (ANC) (*42*). Branch lengths indicate base substitutions per position.

expansion model and understand how it may relate to centromere function. In doing so, we detected evidence of layered expansions across all αSat sequences, from the most diverged fringes of monomeric αSat to the cores of active HOR arrays.

First, we confirmed that two types of divergent αSat symmetrically flank HOR arrays across the genome: divergent HORs (dHORs) (database S11) and monomeric αSats (table S8), which represent ancient, decayed centromeres of primate ancestors (*32*). We classified divergent αSat into distinct SFs and dHOR families and demonstrated how these sequences accumulate mutations, inversions, TE insertions, and non-αSat satellite expansions over time (Fig. 3C; tables S5, S6, and S9; and databases S4, S5, and S7). We also found gradients of size and intra-array divergence (17 to 26%) in monomeric αSat layers, a steep (~10%) divergence increase between HORs and dHORs, and a gradient of embedded TE quantity and age that parallels the age of monomeric layers (Figs. 2E and 3C, table S9, and database S7) (*17*, *32*, *44*).

We next asked whether the layered expansion pattern extends into the active αSat HOR

arrays. On four chromosomes, the active HOR array is surrounded symmetrically by inactive HORs of a distinct type, which is consistent with interarray seeding [chr1 (*60*), chr2, chr16, and chr18] (Fig. 3D). In the assembled centromeres from chrX (*16*, *61*, *62*) and chr8 (*17*), the central part of the active array was found to contain HOR variants slightly different from those on the flanks. To test whether this array structure is typical, we aligned individual HOR units within the same array and clustered them on the basis of their shared sequence variants (*49*, *63*, *64*) into "HOR-haplotypes" or "HOR-haps" (*42*). Initial broad classifications of HOR units into two to four distinct HOR-haps per array revealed symmetrical layering, which typically expands from the middle of the array and is consistent with intra-array seeding and expansion (Fig. 3D, dark red versus gray). Further classification into a larger number of HOR-haps (5 to 10) found additional evidence for symmetric patterns (Fig. 3E) (*42*).

By building rooted phylogenetic trees of consensus HOR-haps, we confirmed that the middle HOR-haps are the most recently evolved (Fig. 3F) (*42*). We also verified this using complete phylogenetic analysis of all HOR units on chr3, chr8, and chrX (shown for chr3 in Fig. 3F) (*42*). In addition, the intra-array divergence in central HOR-haps is often slightly lower than in the flanking arrays, indicating that the central HOR-haps have expanded more recently (Fig. 3F) (*42*). Together, these findings present genome-wide evidence that active αSat HOR arrays evolve rapidly through layered expansions, raising the question of how this dynamic evolutionary process relates to the positioning of the centromere.

Precise mapping of sites of kinetochore assembly

Human centromeres are defined epigenetically as the specific subregion bound by inner kinetochore proteins within each active αSat HOR array (*21*, *65*). Centromeres contain a combination of epigenetic marks that distinguish them from the surrounding pericentromeric heterochromatin. For example, the histone variant Centromere Protein A (CENP-A) is constitutively present at centromeres (*66*) and is often accompanied by "centrochromatin"-associated modifications to canonical histones (*67*). Active αSat arrays also have generally high CpG methylation compared with that of neighboring inactive arrays (*26*) and contain local regions of reduced CpG methylation called centromere dip regions (CDRs) (*16*, *17*, *26*). To study HOR organization at sites of kinetochore assembly, we identified discrete regions of CENP-A enrichment within each active array using sequencing data from native chromatin immunoprecipitation (NChIP-seq) [data from (*17*)] and from CUT&RUN [data from this study (*42*)] (table S10) (*68*). To map these short sequencing reads within αSat arrays, we developed specialized, repeat-sensitive alignment approaches (*42*).

We confirmed that CENP-A binding is almost exclusively localized within αSat HOR arrays, with one active array per chromosome (tables S4 and S11) (*25*). We also found the strongest CENP-A enrichment near and within CDRs on all chromosomes (*17*, *26*). We found that the complete span of each centromere position, defined as a window with high CENP-A enrichment, extends outside of the CDR and totals 190 to 570 kb on each chromosome (Fig. 4, A and B, and table S11). Each CENP-A span occupies 7 to 24% of the total length of the active HOR array in which it is embedded (table S11), which is contrary to predictions from previous work on chrX and chrY in different cell lines (*69*). However, we cannot exclude the possibility that lower levels of CENP-A extend beyond these windows of strong enrichment, or that the sizes of these windows vary among cells or cell types. We detected smaller regions of CENP-A enrichment outside of the primary CDR, with some overlapping a minor, secondary CDR (chr 4, chr16, and chr22) or no CDR at all (chr18) (Fig. 4B, fig. S12, and table S11). Furthermore, similar dips in CpG methylation, although infrequent, do occur outside CENP-A–associated regions, as observed in a 5*S* RNA composite satellite array (*40*) and within a 10-kb region in the active αSat HOR array on chr5 (fig. S12).

We also found that CENP-A is typically enriched in young, recently expanded HOR-haps (Fig. 4, A to D, and table S11). For example, in the active array on chr12, CENP-A is enriched on only one of two large macro-repeat structures, both of which contain similar young HOR-haps (Fig. 4A and fig. S13). Further investigation revealed that CENP-A and the CDR coincide with a zone of very recent HOR expansions (eight sites of nearly identical duplications within a ~365-kb region) (fig. S13) (*42*) that distinguish one macro-repeat region from the other (Fig. 4, A and D). On most other chromosomes, we similarly observed a predominant zone of recently expanded young HOR-haps (*42*), which tends to associate with CENP-A (eight more examples are shown in fig. S14 and table S11).

However, we identified a few notable exceptions to this general trend. On chr4, which has two CENP-A regions occurring on either side of a 1.7-Mb HSat1A array, we found that the larger CENP-A region spans a slightly younger HOR-hap, and the minor CENP-A region spans an older HOR-hap (Fig. 4, B and D). On chr5, chr7, and chr13, CENP-A overlaps young HOR-haps but not near the predominant zone of recent expansions on that chromosome (fig. S15 and table S11) (*42*). Inversely, CENP-A overlaps the zone of recent expansion on chr2, but this zone is composed of older HOR-haps (fig. S15). On chr6, we observed CENP-A enrichment within an older HOR-hap layer, more than a megabase away from the major zone of recent duplications and expansions in this centromere (Fig. 4, C and D). Last, chr21 shows enrichment across the entire active HOR array (the smallest in CHM13) (table S11). We observed that human centromeres and CDRs are typically, although not universally, positioned over young and/or recently expanded layers within active HOR arrays in CHM13, indicating that centromere function is closely related to the rapid evolution of αSat sequences.

Genetic variation across human X centromeres

Satellite DNA arrays are highly variable in size across individuals. The extremes of satellite size variation are often plainly visible under the microscope in chromosomal karyotypes (*30*), yet the clinical relevance of these variants remains unknown and largely unexplored. Studies have provided low-resolution sequencing-based evidence for variability in both satellite array lengths and in the frequency of certain sequence and structural variants within human populations (*11*–*13*, *29*). However, satellite array variation and evolution have remained poorly understood at base-level resolution owing to a lack of complete centromere assemblies.

Therefore, we characterized and compared centromere array assemblies from chrX across seven XY individuals with diverse genetic ancestry [lymphoblastoid cell lines from (*70*)] (Fig. 5A, fig. S16, and table S12). We assigned repeats in the cenX active array to seven HOR-haps, revealing both localized and broad variation within each array (*42*). For example, we identified duplications spanning hundreds of kilobases in two assemblies relative to CHM13 (HG01109 and HG03492) (Fig. 5A and fig. S17). Four of the seven arrays contain zones of recent HOR expansion in the younger HOR-hap (CHM13, HG01109, HG02145, and HG03098). The remaining three assemblies show a trend of recent expansion within older HOR-haps closer to the p-arm (HG03492, HG01243, and HG02055). We also found evidence for a recently expanded HOR-hap type (HOR-hap 6) present in three individuals with recent African ancestry but absent in the other individuals, including CHM13 (Fig. 5A, dark red).

Next, we studied how this variation within αSats relates to variation across single-nucleotide variants (SNVs) that tend to be co-inherited with the centromere. Because meiotic crossover rates are low in peri/centromeric regions (*71*), centromeres are embedded in long haplotypes, called cenhaps (*72*). Cenhaps are identified by clustering pericentromeric SNVs into phylogenetic trees and then splitting them into large clades of shared descent. We divided a group of 1599 XY individuals genotyped using published short-read sequencing data (*73*) into 12 cenhaps (with 98 individuals

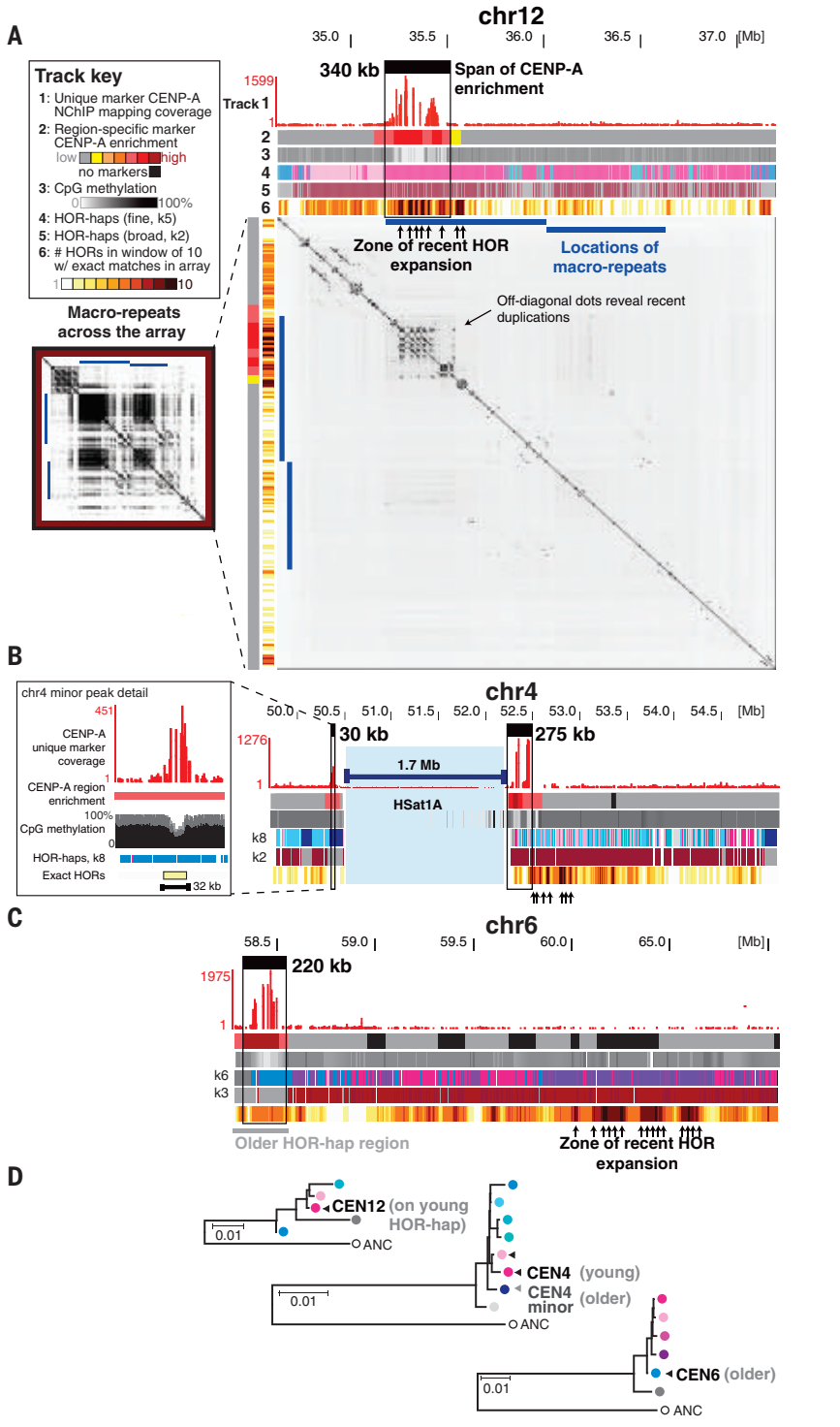


Fig. 4. Inner kinetochore associates with recently expanded αSat HORs. (**A**) Active αSat HOR array on chr12 (coordinates at top). Track 1, CENP-A NChIP-seq marker-assisted mapping coverage. Track 2, reference-free region-specific marker enrichment (black indicates no markers in bin) (*42*). Track 3, percent of CpG sites methylated. Tracks 4 and 5, HOR-haps (k = 5 or 2 clusters, respectively). Track 6, number of HOR units (out of 10 per bin) that have at least one identical copy in the array. (Bottom) Self-alignment dotplot (exact-match word size 2000), with arrows pointing to a zone of recent duplication. (Inset) Smaller dotplot of the entire array (word size 500, allowing for detection of older duplications), with positions of two large macro-repeats indicated with blue lines. (**B**) As in (A) but for chr4. (Inset) Highlighting of a secondary CENP-A enrichment site and minor CDR on the other side of the interrupting HSat1A array. (**C**) As in (A) but for chr6, with CENP-A enrichment over an older HOR-hap region. (**D**) Rooted HOR-hap consensus phylogenetic trees as in Fig. 3F, with CENP-A–enriched region(s) indicated with arrows.

remaining unclassified) (Fig. 5B, fig. S18, and database S12). We also used these short-read sequencing data to estimate the absolute size of each individual's chrX active HOR array (fig. S19 and database S12) (*12*, *72*), along with the relative proportion of that individual's array belonging to each HOR-hap (*42*). This analysis revealed that distinct cenhaps have different αSat array size distributions and different average HOR-hap compositions (Fig. 5B and fig. S18). For example, HOR arrays belonging to cenhaps 1 and 2 tend to be larger overall than those belonging to cenhaps 3 to 12. We found a recent duplication in the chrX HOR array, representing hundreds of kilobases, that is common in cenhap 1 and can explain the relatively larger average array sizes in this cenhap (Fig. 5B).

Two of the 12 cenhaps (1 and 2) are very common in non-African populations (49 and 47% of individuals, respectively) and rare in African populations (1.7 and 3.5%, respectively) (Fig. 5C). The remaining 10 cenhaps are almost exclusive to African populations as well as those with recent African admixture (ASW, PUR, CLM, and ACB). The relatively low cenhap diversity in non-African populations is consistent with their lower overall genetic diversity, which is attributable to demographic bottlenecks during early human migrations out of Africa (*70*). This analysis also revealed that HOR-hap 6 appears to be almost exclusively found in cenhaps 10 to 12, which form an anciently diverged clade within African populations (Fig. 5B). These findings demonstrate that centromere-linked SNVs can be used to tag and track the evolution of αSat, and they underline the need for greater representation of African genomes in pan-genome assembly efforts.

Last, to dissect the sequence differences between two arrays from the same cenhap, we compared two finished centromere assemblies from CHM13 and HG002, a cell line whose chrX array had been constructed by use of T2T assembly methods and whose array structure had been experimentally validated (*2*). We found both genomes to be highly concordant across the array, apart from three regions, where we observed recent amplifications and/or deletions of repeats (Fig. 5D and fig. S20). These comparisons of completely assembled centromeres demonstrate that satellite DNA variation is common at both coarse and fine scales, raising the question of how this genetic variation relates to possible epigenetic variation in centromere positioning.

Epigenetic variation across human X centromeres

To examine how centromere positioning varies among individuals, we compared patterns of CENP-A CUT&RUN enrichment on the fully assembled chrX centromeres from HG002 and

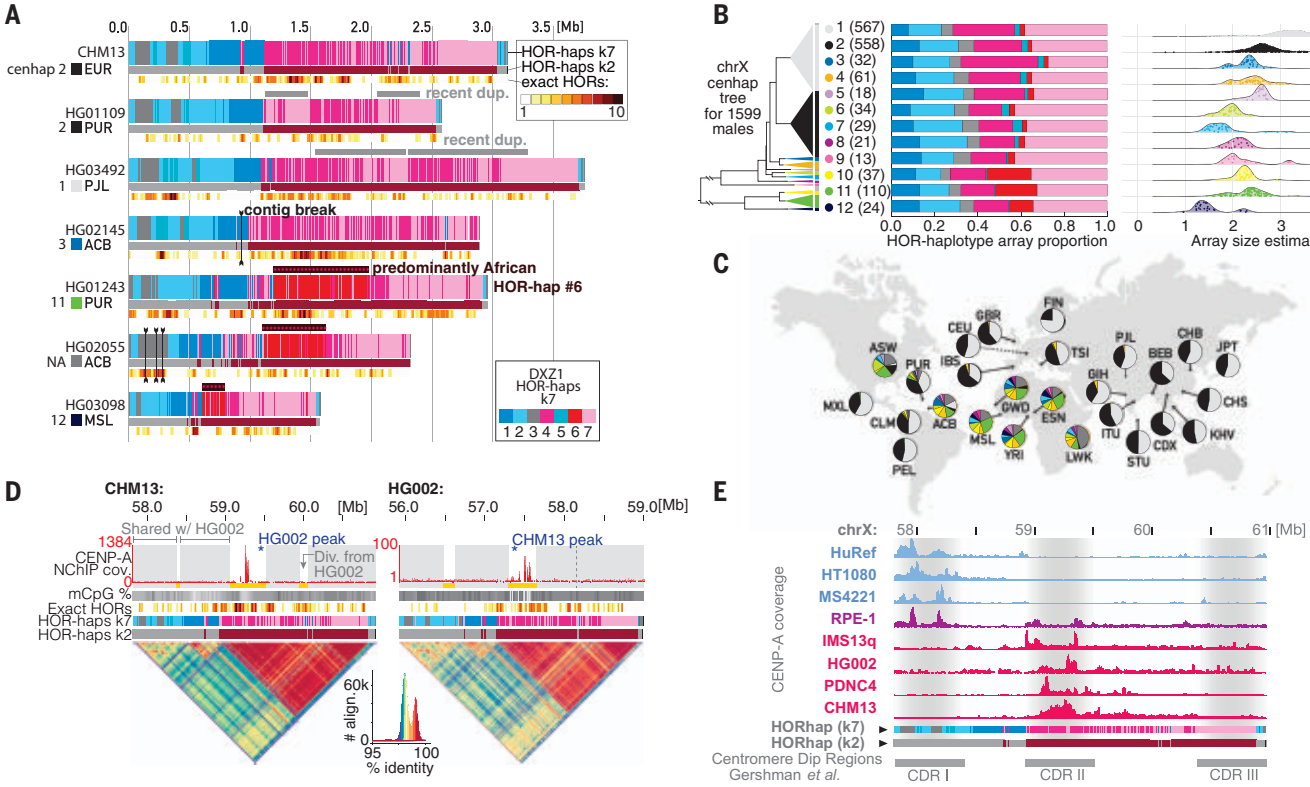


Fig. 5. Substantial genetic and epigenetic variation in and around the chrX centromere. (**A**) Comparing the active αSat HOR array on chrX (DXZ1) between (top) CHM13 and six HPRC cell line HiFi read assemblies. Tracks indicate HOR-haps (top, *k* = 7; bottom, *k* = 2) and recent HOR duplication events (bottom, as in Fig. 4A). (**B**) (Left) Phylogenetic tree illustrating the relationships of 12 cenhaps defined by using short-read data from 1599 XY genomes from (*70*, *73*) plus HG002, CHM13, and HuRef. Triangle vertical length is proportional to the number of individuals in that cenhap (98 individuals, labeled NA and colored dark gray, belong to small clades not among the 12 major cenhaps). (Middle) Barplots illustrating the average HOR-hap compositions for all individuals within each cenhap, colored as in (A). (Right) Ridgeline plots indicating the distribution of estimated total array sizes for all individuals within each cenhap, with individual values represented as jittered points. (**C**) Populations represented among the 1599 XY genomes, with pie charts indicating the proportion of cenhap assignments within each population, with the same colors used as in the tree in (B). Population descriptions are in (*42*). (**D**) Comparison of the DXZ1 assembly for CHM13 and HG002, which are both in cenhap 2. Tracks are as in (A), with the addition of a top track to indicate regions that align closely (gray) or are diverged (yellow) between the two individuals. Vertical dotted line indicates the homologous site of a CHM13 expansion on the HG002 array. (Bottom) StainedGlass dotplots representing the percent identity of self-alignments within the array, with a color-key and histogram below (*88*). (**E**) A comparison of CENP-A coverage (NChIP-seq or CUT&RUN) in eight cell lines relative to the CHM13 chrX centromere assembly. Each track is normalized to its maximum peak height in the array. Below are CDR positions from (*26*).

CHM13 (*26*). The region with the strongest CENP-A enrichment in both arrays coincides with the most pronounced sequence differences between CHM13 and HG002, mostly because of structural rearrangements (Fig. 5D, yellow, and fig. S20). Despite these local structural differences, CENP-A remains positioned over CDRs and young HOR-haps in both individuals.

Last, we asked whether CENP-A enrichment patterns were consistently found in younger HOR-haps, as observed in CHM13 and HG002, across seven additional cell lines with publicly available CENP-A NChIP-seq and CUT&RUN datasets (Fig. 5E and fig. S21). Unlike CHM13, in three XY individuals we observed CENP-A enrichment within the older HOR-hap subregion, proximal to the p-arm, indicating the presence of an epiallele [HuRef (*74*), HT1080b (*75*), and MS4221 (*76*)]. This coincides with an alternative CDR observed in the HG03098 cell line [CDR I from (*26*)] (Fig. 5E). Further, we examined two independent CUT&RUN experiments from the RPE-1 cell line (XX) (*77*) and found enrichment on both older and younger HOR-haps, which could be explained if the two chrX homologs carry different functional epialleles. Three additional XX cell lines were consistent with CHM13, providing evidence that the same CENP-A–enriched HOR-hap is shared across both chrX homologs in each line (IMS13q, PDNC4, and K562) (Fig. 5E and fig. S21) (*78*). These overlap a CDR also seen in the HG01109 cell line [CDR II from (*26*)] (Fig. 5E). A third CDR proximal to the q-arm was observed in the HG01243 and HG03492 cell lines (*26*), which is indicative of a third possible CENP-A epiallele. These findings uncover frequent variation in the position of the chrX centromere, with some XX individuals potentially harboring heterozygous epialleles.

Discussion

This study provides comprehensive maps of recently assembled human peri/centromeric regions to facilitate exploration of their function, variation, and evolution. Using this resource, we uncovered strong evidence that the genetic and epigenetic fates of centromeres are intertwined through evolution: αSat arrays evolve through layered expansions, and the inner-kinetochore protein CENP-A tends to associate with the most recently expanded sequences. The kinetochore frequently shifts to new loci, and the old loci rapidly shrink and decay.

One possible explanation for this relationship is that αSat expansions occur independently of the kinetochore, but the kinetochore

maintains an affinity for some property of recently expanded sequences, such as their homogeneity (the "independent expansion hypothesis"). Kinetochore-independent expansion is feasible in light of our observation of large duplications and localized repeat expansions in noncentromeric satellites such as HSat3 arrays, which are not associated with kinetochores (fig. S11). Another possibility is that kinetochore proteins—or other proteins that may associate with the centromere such as loading, replication, recombination, or repair factors—play a causal role in the expansion of particular HOR variants [the "kinetochore selection hypothesis" (*36*)]. This aligns with the proposed recombination-based homogenization process in *Arabidopsis* (*79*). Further, experiments in model organisms have demonstrated that extreme array sequence variants increase meiotic and mitotic nondisjunction rates and can promote both mutational drive and/or female meiotic drive (*20*, *80*–*82*). Similar drive mechanisms (*83*), along with selection for variants that promote high-fidelity chromosome transmission, may also play a role in shaping centromeric sequence evolution in humans. Exploring these evolutionary models, as well as studying why CENP-A colocalizes with CDRs, will require precise experimental methods for measuring interactions between kinetochore proteins and repetitive DNA [such as DiMeLo-seq (*84*)].

Fully assembled peri/centromeric regions also provide a reference against which sequencing information from multiple individuals can be aligned and compared. By doing so, we uncovered a 400-kb polymorphic deletion of an entire HSat3 array and a 1.7-Mb polymorphic inversion in an active αSat HOR array, both on chr1. We also detected an expansion of a particular αSat sequence variant on chrX in individuals with recent African ancestry. This high degree of satellite DNA polymorphism underlines the need to produce T2T assemblies from genetically diverse individuals, to fully capture the extent of human variation in these regions, and to shed light on their recent evolution. Measuring this variation will also be essential to understand the functional consequences of satellite variation on centromere function or, in the case of HSat3, on phenomena such as satellite transcription in response to stress [reviewed in (*38*)].

Along with genetic variation, we identified epigenetic variation in the location of CENP-A within the αSat array on chrX, similar to a rare but well-studied epiallele on chr17 (*85*–*87*). CENP-A is typically positioned on young HOR-haps on chrX, as seen for most chromosomes in CHM13. However, in some cell lines, CENP-A appears to be positioned over older chrX HOR-haps more than a megabase away (Fig. 5E), which is similar to the positioning of the chr6 CENP-A locus in CHM13. Thus, although CENP-A tends to localize to the most recently expanded HORs, there are exceptions on at least some chromosomes in some individuals. Studying centromere positioning across many samples, across families, and across different tissues from the same individuals will reveal the extent of this epigenetic plasticity in centromere localization and how this epigenetic variation relates to genetic variation and evolution. This will potentially illuminate how human cells maintain essential centromere functions despite the rapid evolution of centromeric DNA and inner-kinetochore proteins, an anomaly referred to as the "centromere paradox" (*20*).

Materials and methods

A very brief methods overview is provided here. Detailed methods are provided in (*42*). Repeats in the T2T-CHM13 assembly were annotated by parsing and combining output from RepeatMasker [provided in (*40*)] along with custom-built pipelines for annotating αSat and HSat2,3 (*42*). Regions identified as "SAR" by RepeatMasker were annotated as HSat1A, and regions annotated as "HSATI" by RepeatMasker were annotated as HSat1B. αSat HOR-haps were identified by (i) generating multiple alignments of all HOR units (or subregions of HOR units) from an array, (ii) deriving a consensus sequence, (iii) recoding the individual sequences into binary vectors based on matches to the consensus, and (iv) clustering these binary vectors by use of *k*-means clustering. Phylogenetic analyses of αSat sequences were performed with MEGA5. Dotplots colored by percent identity were produced with StainedGlass (*88*).

To analyze short-read NChIP-seq and CUT&RUN data, two parallel methods were developed: (i) marker-assisted mapping to the T2T-CHM13 reference and (ii) reference-free region-specific marker enrichment. For marker-assisted mapping, reads were aligned to the reference then filtered to include only alignments that overlap precomputed nucleotide oligomers of length *k* (*k*-mers) that occur in only one distinct position in the reference. For reference-free enrichment analysis, a set of *k*-mers that are enriched in CENP-A–targeted sequencing reads (relative to reads from input or immunoglobulin G controls) were first identified. Next, these enriched *k*-mers were compared with precomputed *k*-mers in the reference that occur exclusively within a single window of a given size ("region-specific markers"). Windows with multiple matches to enriched *k*-mers were reported as enriched for CENP-A. We performed a similar analysis using HOR-hap–specific markers on chrX, to reveal the broad enrichment of CENP-A on each HOR-hap across multiple individuals (fig. S21).

REFERENCES AND NOTES

1. E. E. Eichler, R. A. Clark, X. She, An assessment of the sequence gaps: Unfinished business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004). doi: 10.1038/nrg1322; pmid: 15143317
2. S. Nurk *et al*., The complete sequence of a human genome. *Science* **376**, 44 (2022).
3. K. H. Miga, Completing the human genome: The progress and challenge of satellite DNA assembly. *Chromosome Res.* **23**, 421–426 (2015). doi: 10.1007/s10577-015-9488-2; pmid: 26363799
4. K. L. McKinley, I. M. Cheeseman, The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016). doi: 10.1038/nrm.2015.5; pmid: 26601620
5. R. Wevrick, H. F. Willard, Physical map of the centromeric region of human chromosome 7: Relationship between two distinct alpha satellite arrays. *Nucleic Acids Res.* **19**, 2295–2301 (1991). doi: 10.1093/nar/19.9.2295; pmid: 2041770
6. M. S. Jackson, P. Slijepcevic, B. A. Ponder, The organisation of repetitive sequences in the pericentromeric region of human chromosome 10. *Nucleic Acids Res.* **21**, 5865–5874 (1993). doi: 10.1093/nar/21.25.5865; pmid: 8290346
7. H. E. Trowell, A. Nagy, B. Vissel, K. H. Choo, Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: Identification of a narrow domain containing two key centromeric DNA elements. *Hum. Mol. Genet.* **2**, 1639–1649 (1993). doi: 10.1093/hmg/2.10.1639; pmid: 8268917
8. C. Tyler-Smith, Structure of repeated sequences in the centromeric region of the human Y chromosome. *Development* **101** (suppl.), 93–100 (1987). doi: 10.1242/dev.101.Supplement.93; pmid: 3503726
9. I. Tagarro, A. M. Fernández-Peralta, J. J. González-Aguilera, Chromosomal localization of human satellites 2 and 3 by a FISH method using oligonucleotides as probes. *Hum. Genet.* **93**, 383–388 (1994). doi: 10.1007/BF00201662; pmid: 8168808
10. N. Archidiacono *et al*., Comparative mapping of human alphoid sequences in great apes using fluorescence in situ hybridization. *Genomics* **25**, 477–484 (1995). doi: 10.1016/0888-7543(95)80048-Q; pmid: 7789981
11. N. Altemose, K. H. Miga, M. Maggioni, H. F. Willard, Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLOS Comput. Biol.* **10**, e1003628 (2014). doi: 10.1371/journal.pcbi.1003628; pmid: 24831296
12. K. H. Miga *et al*., Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014). doi: 10.1101/gr.159624.113; pmid: 24501022
13. Y. Suzuki, E. W. Myers, S. Morishita, Rapid and ongoing evolution of repetitive sequence structures in human centromeres. *Sci. Adv.* **6**, eabd9230 (2020). doi: 10.1126/sciadv.abd9230; pmid: 33310858
14. C. Alkan *et al*., Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLOS Comput. Biol.* **3**, 1807–1818 (2007). doi: 10.1371/journal.pcbi.0030181; pmid: 17907796
15. M. Jain *et al*., Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018). doi: 10.1038/nbt.4109; pmid: 29553574
16. K. H. Miga *et al*., Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi: 10.1038/s41586-020-2547-7; pmid: 32663838
17. G. A. Logsdon *et al*., The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021). doi: 10.1038/s41586-021-03420-7; pmid: 33828295
18. A. M. M. Cartney *et al*., Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *bioRxiv* 450803 [Preprint] (2021); doi: 10.1101/2021.07.02.450803
19. S. I. Nagaoka, T. J. Hassold, P. A. Hunt, Human aneuploidy: Mechanisms and new insights into an age-old problem. *Nat. Rev. Genet.* **13**, 493–504 (2012). doi: 10.1038/nrg3245; pmid: 22705668
20. S. Henikoff, K. Ahmad, H. S. Malik, The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001). doi: 10.1126/science.1062939; pmid: 11498581
21. G. H. Karpen, R. C. Allshire, The case for epigenetic effects on centromere identity and function. *Trends Genet.* **13**, 489–496 (1997). doi: 10.1016/S0168-9525(97)01298-5; pmid: 9433139
22. J. C. Wu, L. Manuelidis, Sequence definition and organization of a human repeated DNA. *J. Mol. Biol.* **142**, 363–386 (1980). doi: 10.1016/0022-2836(80)90277-6; pmid: 6257909

23. H. F. Willard, The genomics of long tandem arrays of satellite DNA in the human genome. *Genome* **31**, 737–744 (1989). doi: 10.1139/g89-132; pmid: 2698839
24. H. F. Willard, J. S. Waye, Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987). doi: 10.1016/0168-9525(87)90232-0
25. S. M. McNulty, B. A. Sullivan, Alpha satellite DNA biology: Finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018). doi: 10.1007/s10577-018-9582-3; pmid: 29974361
26. A. Gershman *et al*., Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
27. I. Alexandrov, A. Kazakov, I. Tumeneva, V. Shepelev, Y. Yurov, Alpha-satellite DNA of primates: Old and new families. *Chromosoma* **110**, 253–266 (2001). doi: 10.1007/s004120100146; pmid: 11534817
28. H. F. Willard, Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37**, 524–532 (1985). pmid: 2988334
29. K. H. Miga, Centromeric satellite DNAs: Hidden sequence variation in the human population. *Genes* **10**, 352 (2019). doi: 10.3390/genes10050352; pmid: 31072070
30. A. P. Craig-Holmes, M. W. Shaw, Polymorphism of human constitutive heterochromatin. *Science* **174**, 702–704 (1971). doi: 10.1126/science.174.4010.702; pmid: 5123418
31. M. M. Mahtani, H. F. Willard, Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: High-frequency polymorphisms and array size estimate. *Genomics* **7**, 607–613 (1990). doi: 10.1016/0888-7543(90)90206-A; pmid: 1974881
32. V. A. Shepelev, A. A. Alexandrov, Y. B. Yurov, I. A. Alexandrov, The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLOS Genet.* **5**, e1000641 (2009). doi: 10.1371/journal.pgen.1000641; pmid: 19749981
33. X. She *et al*., The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864 (2004). doi: 10.1038/nature02806; pmid: 15318213
34. G. Genovese *et al*., Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406–414, 414e1–2 (2013). doi: 10.1038/ng.2565; pmid: 23435088
35. C. Lee, R. Wevrick, R. B. Fisher, M. A. Ferguson-Smith, C. C. Lin, Human centromeric DNAs. *Hum. Genet.* **100**, 291–304 (1997). doi: 10.1007/s004390050508; pmid: 9272147
36. K. H. Miga, I. A. Alexandrov, Variation and evolution of human centromeres: A field guide and perspective. *Annu. Rev. Genet.* **55**, 583–602 (2021). doi: 10.1146/annurev-genet-071719-020519; pmid: 34813350
37. J. Prosser, M. Frommer, C. Paul, P. C. Vincent, Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* **187**, 145–155 (1986). doi: 10.1016/0022-2836(86)90224-X; pmid: 3701863
38. N. Altemose, A classical revival: Human satellite DNAs enter the genomics era. *Preprints* 10.20944/preprints202202.0009.v1 (2022), doi: 10.20944/preprints202202.0009.v1
39. M. Frommer, J. Prosser, P. C. Vincent, Human satellite I sequences include a male specific 2.47 kb tandemly repeated unit containing one Alu family member per repeat. *Nucleic Acids Res.* **12**, 2887–2900 (1984). doi: 10.1093/nar/12.6.2887; pmid: 6324132
40. S. J. Hoyt *et al*., From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
41. M. R. Vollger *et al*., Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
42. Materials and methods are available as supplementary materials.
43. V. A. Shepelev *et al*., Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genom. Data* **5**, 139–146 (2015). doi: 10.1016/j.gdata.2015.05.035; pmid: 26167452
44. L. I. Uralsky *et al*., Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. *Data Brief* **24**, 103708 (2019). doi: 10.1016/j.dib.2019.103708; pmid: 30989093
45. A. W. Lo, G. C. Liao, M. Rocchi, K. H. Choo, Extreme reduction of chromosome-specific alpha-satellite array is unusually common in human chromosome 21. *Genome Res.* **9**, 895–908 (1999). doi: 10.1101/gr.9.10.895; pmid: 10523519
46. M. K. Rudd, H. F. Willard, Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **20**, 529–533 (2004). doi: 10.1016/j.tig.2004.08.008; pmid: 15475110
47. A. E. Kazakov *et al*., Interspersed repeats are found predominantly in the "old" α satellite families. *Genomics* **82**, 619–627 (2003). doi: 10.1016/S0888-7543(03)00182-4; pmid: 14611803
48. P. E. Warburton, H. F. Willard, Genomic analysis of sequence variation in tandemly repeated DNA. Evidence for localized homogeneous sequence domains within arrays of alpha-satellite DNA. *J. Mol. Biol.* **216**, 3–16 (1990). doi: 10.1016/S0022-2836(05)80056-7; pmid: 2122000
49. P. E. Warburton, H. F. Willard, Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: Evidence for concerted evolution along haplotypic lineages. *J. Mol. Evol.* **41**, 1006–1015 (1995). doi: 10.1007/BF00173182; pmid: 8587099
50. K. H. Miga, T. Wang, The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021). doi: 10.1146/annurev-genom-120120-081921; pmid: 33929893
51. M. G. Schueler, A. W. Higgins, M. K. Rudd, K. Gustashaw, H. F. Willard, Genomic and genetic definition of a functional human centromere. *Science* **294**, 109–115 (2001). doi: 10.1126/science.1065042; pmid: 11588252
52. R. Bandyopadhyay, C. McQuillan, S. L. Page, K. H. Choo, L. G. Shaffer, Identification and characterization of satellite III subfamilies to the acrocentric chromosomes. *Chromosome Res.* **9**, 223–233 (2001). doi: 10.1023/A:1016648404388; pmid: 11330397
53. A. Agresti *et al*., Linkage in human heterochromatin between highly divergent Sau3A repeats and a new family of repeated DNA sequences (HaeIII family). *J. Mol. Biol.* **205**, 625–631 (1989). doi: 10.1016/0022-2836(89)90308-2; pmid: 2538633
54. D. P. Ryan *et al*., Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell* **140**, 88–98 (2010). doi: 10.1016/j.cell.2009.12.024; pmid: 20074522
55. M.-L. Dubois *et al*., UBB pseudogene 4 encodes functional ubiquitin variants. *Nat. Commun.* **11**, 1306 (2020). doi: 10.1038/s41467-020-15090-6; pmid: 32161257
56. S. Aganezov *et al*., A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
57. V. Paar, I. Basar, M. Rosandić, M. Gluncić, Consensus higher order repeats and frequency of string distributions in human genome. *Curr. Genomics* **8**, 93–111 (2007). doi: 10.2174/138920207780368169; pmid: 18660848
58. I. A. Alexandrov, S. P. Mitkevich, Y. B. Yurov, The phylogeny of human chromosome specific alpha satellites. *Chromosoma* **96**, 443–453 (1988). doi: 10.1007/BF00303039; pmid: 3219915
59. H. F. Willard, J. S. Waye, Chromosome-specific subsets of human alpha satellite DNA: Analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* **25**, 207–214 (1987). doi: 10.1007/BF02100014; pmid: 2822935
60. P. Finelli *et al*., Structural organization of multiple alphoid subsets coexisting on human chromosomes 1, 4, 5, 7, 9, 15, 18, and 19. *Genomics* **38**, 325–330 (1996). doi: 10.1006/geno.1996.0635; pmid: 8975709
61. A. V. Bzikadze, P. A. Pevzner, Automated assembly of centromeres from ultra-long error-prone reads. *Nat. Biotechnol.* **38**, 1309–1316 (2020). doi: 10.1038/s41587-020-0582-4; pmid: 32665660
62. K. H. Miga, Centromere studies in the era of 'telomere-to-telomere' genomics. *Exp. Cell Res.* **394**, 112127 (2020). doi: 10.1016/j.yexcr.2020.112127; pmid: 32504677
63. P. E. Warburton, R. Wevrick, M. M. Mahtani, H. F. Willard, Pulsed-field and two-dimensional gel electrophoresis of long arrays of tandemly repeated DNA: Analysis of human centromeric alpha satellite. *Methods Mol. Biol.* **12**, 299–317 (1992). doi: 10.1385/0-89603-229-9:299; pmid: 21409641
64. S. J. Durfy, H. F. Willard, Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: Evidence for short-range homogenization of tandemly repeated DNA sequences. *Genomics* **5**, 810–821 (1989). doi: 10.1016/0888-7543(89)90123-7; pmid: 2591964
65. M. D. Blower, B. A. Sullivan, G. H. Karpen, Conserved organization of centromeric chromatin in flies and humans. *Dev. Cell* **2**, 319–330 (2002). doi: 10.1016/S1534-5807(02)00135-1; pmid: 11879637
66. A. A. Van Hooser *et al*., Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. *J. Cell Sci.* **114**, 3529–3542 (2001). doi: 10.1242/jcs.114.19.3529; pmid: 11682612
67. B. A. Sullivan, G. H. Karpen, Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat. Struct. Mol. Biol.* **11**, 1076–1083 (2004). doi: 10.1038/nsmb845; pmid: 15475964
68. P. J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017). doi: 10.7554/eLife.21856; pmid: 28079019
69. L. L. Sullivan, C. D. Boivin, B. Mravinac, I. Y. Song, B. A. Sullivan, Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosome Res.* **19**, 457–470 (2011). doi: 10.1007/s10577-011-9208-5; pmid: 21484447
70. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245
71. M. Nambiar, G. R. Smith, Repression of harmful meiotic recombination in centromeric regions. *Semin. Cell Dev. Biol.* **54**, 188–197 (2016). doi: 10.1016/j.semcdb.2016.01.042; pmid: 26849908
72. S. A. Langley, K. H. Miga, G. H. Karpen, C. H. Langley, Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *eLife* **8**, e42989 (2019). doi: 10.7554/eLife.42989; pmid: 31237235
73. M. Byrska-Bishop *et al*., High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* 430068 (2021); doi: 10.1101/2021.02.06.430068
74. J. G. Henikoff, J. Thakur, S. Kasinathan, S. Henikoff, A unique chromatin complex occupies young α-satellite arrays of human centromeres. *Sci. Adv.* **1**, e1400234 (2015). doi: 10.1126/sciadv.1400234; pmid: 25927077
75. J. Thakur, S. Henikoff, CENPT bridges adjacent CENPA nucleosomes on young human α-satellite dimers. *Genome Res.* **26**, 1178–1187 (2016). doi: 10.1101/gr.204784.116; pmid: 27384170
76. D. Hasson *et al*., The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat. Struct. Mol. Biol.* **20**, 687–695 (2013). doi: 10.1038/nsmb.2562; pmid: 23644596
77. M. Dumont *et al*., Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. *EMBO J.* **39**, e102924 (2020). doi: 10.15252/embj.2019102924; pmid: 31750958
78. S. J. Falk *et al*., Chromosomes. CENP-C reshapes and stabilizes CENP-A nucleosomes at the centromere. *Science* **348**, 699–703 (2015). doi: 10.1126/science.1259308; pmid: 25954010
79. M. Naish *et al*., The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021). doi: 10.1126/science.abi7489; pmid: 34762468
80. L. Fishman, J. K. Kelly, Centromere-associated meiotic drive and female fitness variation in Mimulus. *Evolution* **69**, 1208–1218 (2015). doi: 10.1111/evo.12661; pmid: 25873401
81. L. E. Kursel, H. S. Malik, The cellular mechanisms and consequences of centromere drive. *Curr. Opin. Cell Biol.* **52**, 58–65 (2018). doi: 10.1016/j.ceb.2018.01.011; pmid: 29454259
82. L. Chmátal *et al*., Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr. Biol.* **24**, 2295–2300 (2014). doi: 10.1016/j.cub.2014.08.017; pmid: 25242031
83. W. R. Rice, A game of thrones at human centromeres II. A new molecular/evolutionary model. *bioRxiv* 731471 [Preprint] (2019); doi: 10.1101/731471
84. N. Altemose *et al*., DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide. *bioRxiv* 451383 [Preprint] (2021); doi: 10.1101/2021.07.06.451383
85. K. A. Maloney *et al*., Functional epialleles at an endogenous human centromere. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13704–13709 (2012). doi: 10.1073/pnas.1203126109; pmid: 22847449
86. M. E. Aldrup-MacDonald, M. E. Kuo, L. L. Sullivan, K. Chew, B. A. Sullivan, Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* **26**, 1301–1311 (2016). doi: 10.1101/gr.206706.116; pmid: 27510565
87. K. E. Hayden *et al*., Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33**, 763–772 (2013). doi: 10.1128/MCB.01198-12; pmid: 23230266
88. M. R. Vollger, P. Kerpedjiev, A. M. Phillippy, E. E. Eichler, StainedGlass: Interactive visualization of massive tandem

repeat structures with identity heatmaps. *Bioinformatics*, btac018 (2022). doi: 10.1093/bioinformatics/btac018; pmid: 35020798
89. B. J. Raney *et al*., Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003–1005 (2014) doi: 10.1093/bioinformatics/btt637; pmid: 24227676
90. W. J. Kent *et al*., The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi: 10.1101/gr.229102; pmid: 12045153
91. J. Lucas, kmiga/alphaAnnotation: HumAS-HMMER_for_AnVIL. *Zenodo* (2021); doi: 10.5281/zenodo.5715444
92. N. Altemose, altemose/NTRprism: NTRprism-v1.0.0. *Zenodo* (2021); doi: 10.5281/zenodo.5715473
93. L. Uralsky, Phylogenetic trees and supp. alignments for Altemose et al. 2021, T2T Consortium. *FigShare* (2022); doi: 10.6084/m9.figshare.19299857.v1

ACKNOWLEDGMENTS

This work used the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. **Funding:** This work was supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (S.N., S.K., A.R., A.M.M., S.B., A.Y., and A.M.P.), the Intramural funding at the National Institute of Standards and Technology (J.M.Z.), HHMI Hanna H. Gray Fellowship (N.A.), Damon Runyon Postdoctoral Fellowship, and Pew Latin American Fellowship (G.V.C.). Grants supporting this work are from the US National Institutes of Health (NIH/NIGMS F32 GM134558 to G.A.L.; NIH/NHGRI R01 GM074728 to P.S. and A.F.S.; NIH R01GM123312-02 to S.J.H., G.A.H., and R.J.O.; NIH R21CA240199 to R.J.O.; NIH/NHGRI F31HG011205 to C.J.S.; NIH/NHGRI R01HG009190 to A.G. and W.T.; NIH/NHGRI R01HG010485, U41HG010972, and U01HG010961 to K.S.; NIH/NIGMS R01GM132600, P20GM103546, and NIH/NHGRI U24HG010136 to D.O. and T.J.W.; NIH/NHGRI R01HG010329 to S.R.S.; NIH/NHGRI R01HG010485, U41HG010972, U01HG010961, U24HG011853, and OT2OD026682 to B.P.; NIH R01AG054712 to E.I.R.; NIH/GM R35GM139653 and R01GM117420 to G.H.K.; NIH R01GM124041, R01GM129263, and R21CA238758 to B.A.S.; NIH/NHGRI R01HG010169 and U01HG010971 to E.E.E.; NIH/OD/NIMH DP2MH119424 to M.Y.D.; NIH/NHGRI U24HG010263; NHGRI U24HG006620; NCI U01CA253481; NIDDK R24DK106766-01A1 to M.C.S.; NIH/NHGRI U41HG007234 to M.D.; NIH/NHGRI R01HG011274-01 and NIH/NHGRI R21HG010548-01 to K.H.M.), National Science Foundation (NSF 1613806 to S.J.H., G.A.H., and R.J.O.; NSF 1643825 to R.J.O.; NSF DBI-1627442, NSF IOS-1732253, and NSF IOS-1758800 to M.C.S.); Mark Foundation for Cancer Research to SA and MCS (19-033-ASP); Russian Science Foundation RSF 19-75-30039 (analysis of genomic repeats) (I.A.A.); St. Petersburg State University (grant ID PURE 73023573 to A.M., T.D., and I.A.A. and grant ID PURE 51555639 to O.K.); Supported by the Sirius University (L.U.); Ministry of Science and Higher Education of the Russian Federation [075-10-2020-116 (13.1902.21.0023)] (F.G.); Connecticut Innovations to R.J.O.; and Stowers Institute for Medical Research to J.L.G.; A.S. is a Chan Zuckerberg Biohub Investigator. E.E.E. and A.F.D. are investigators of the Howard Hughes Medical Institute.
Author contributions: αSat sequence characterization: A.V.B., L.U., F.D.R., A.M., V.A.S., T.D., O.K., F.G., E.I.R., P.A.P., N.A., I.A.A., and K.H.M. Pericentromeric satellite characterization: N.A., G.A.L., S.J.H., M.E.G.S., D.O., T.J.W., L.G.d.L., A.M.P., R.J.O., and K.H.M. CUT&RUN experiments, mapping, and enrichment analyses: G.V.C., N.A., G.A.L., P.S., S.J.H., A.M.M., A.R., M.E.G.S., K.T., S.R.S., A.S., A.F.S., B.A.S., A.F.D., G.H.K., A.G., W.T., and K.H.M. Cenhap analysis and interpretation: S.A.L., N.A., C.H.L., I.A.A., and K.H.M. Array length prediction: M.B., J.L.G., M.C.S., and J.M.Z. Methylation analysis: A.G. and W.T.; Chromosome imaging and flow sorting: T.P., J.L.G., S.B., A.Y., and A.M.P. Dotplot analysis: L.U., F.D.R., M.R.V., R.L., P.K., A.M.P., and I.A.A. TE analysis: N.A., S.J.H., G.A.H., R.J.O., L.U., and I.A.A. CHM13 satellite assembly and het analysis: N.A., G.A.L., S.N., S.K., A.R., A.M.M., A.M.P. UCSC genome browser and annotation workflow: M.D. Formalizing code for satellite annotation workflows: J.K.L., F.G., N.A., S.L., K.H.M. HiFi assemblies and quality assessment of HPRC panel: N.A., M.A., R.L., K.S., A.M., A.V.B., S.A., J.M.Z., M.C.S., B.P., E.E.E., and A.M.P. Gene annotation and expression: C.J.S., M.R.V., M.H., M.Y.D., and M.D. Manuscript writing: N.A., I.A.A., and K.H.M., with input from all authors. **Competing interests:** S.K. and K.H.M. have received travel funds to speak at symposia organized by Oxford Nanopore Technologies. W.T. has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore Technologies. S.A. is an employee of Oxford Nanopore Technologies. P.K. owns and receives income from Reservoir Genomics. E.E.E. is a scientific advisory board (SAB) member of Variant Bio. K.H.M. is a SAB member of Centaura. No other competing interests are declared from other authors. **Data and materials availability:** CHM13hTERT cells were obtained for research use through a materials transfer agreement with U. Surti and the University of Pittsburgh. Genomic and CUT&RUN sequencing data are available through BioProjects PRJNA559484 (CHM13) and PRJNA752795 (HG002). Additional data described in the paper are presented, curated, or archived in the supplementary materials. The Human Pangenome Reference Consortium (HPRC) data for 16 human samples (HG002, HG003, HG004, HG005, HG006, HG007, HG01243, HG02055, HG02109, HG02723, HG03492, HG01109, HG01442, HG02080, HG02145, and HG03098) were obtained from BioProject PRJNA730822. Nonhuman primate reference sequences cited in database S6 were obtained from Genbank (accessions NW_019932813, NW_019932814, NW_019933443, NW_019932985, NW_019932799, NW_022149050, NW_022149051, SRLZ01006095, NW_022149053, NW_022149054, NW_019937219, and NC_041757). Data tracks and satellite annotations can be visualized on the UCSC Genome Browser (*89*, *90*), linked from https://github.com/marbl/chm13. Code used for analyzing the data are available at (*91*, *92*). Phylogenetic data are available at (*93*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl4178
Materials and Methods
Supplementary Text
Figs. S1 to S43
Tables S1 to S16
References (*94*–*155*)
MDAR Reproducibility Checklist
Databases S1 to S21

12 July 2021; accepted 17 February 2022
10.1126/science.abl4178

Your Legacy to Science

AN ESTATE GIFT TO THE
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE





Since 1848, our founding year, the American Association for the Advancement of Science (AAAS) has been deeply committed to advancing science, engineering and innovation around the world for the benefit of all people.

By making AAAS a beneficiary of your will, trust, retirement plan or life insurance policy, you become a member of our 1848 Society, joining Thomas Edison, Alexander Graham Bell and the many distinguished individuals whose vision led to the creation of AAAS and our world-renowned journal, *Science*, so many years ago.

Unlike many of its peers, *Science* is not for-profit. Your estate gift would provide long-term financial stability and durable annual income that will support operations and competitive innovation for years to come. **This support is vital.**

"As a teacher and instructor, I bear responsibility for the younger generations. If you have extra resources, concentrate them on organizations, like AAAS, that are doing work for all."

—Prof. Elisabeth Ervin-Blankenheim, 1848 Society member

If you intend to include AAAS in your estate plans, provide this information to your lawyer or financial adviser:

Legal Name: American Association for the Advancement of Science
Federal Tax ID Number: 53-0196568
Address: 1200 New York Avenue, NW, Washington, DC 20005

If you would like more information on making an estate gift to AAAS, cut out and return the form below or send an email to philanthropy@aaas.org. Additional details are also available online at **www.aaas.org/1848Society**.

1848 SOCIETY AAAS

cut here

Yes, I would like more information about joining the AAAS 1848 Society.

PLEASE CONTACT ME AT:

Name: _____

Address: _____

City: _____ State: _____ Zip code: _____ Country: _____

Email: _____ Phone: _____

RETURN THIS FORM TO:
AAAS Office of Philanthropy and Strategic Partnerships • 1200 New York Avenue, NW • Washington, DC 20005 USA

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

From telomere to telomere: The transcriptional and epigenetic state of human repeat elements

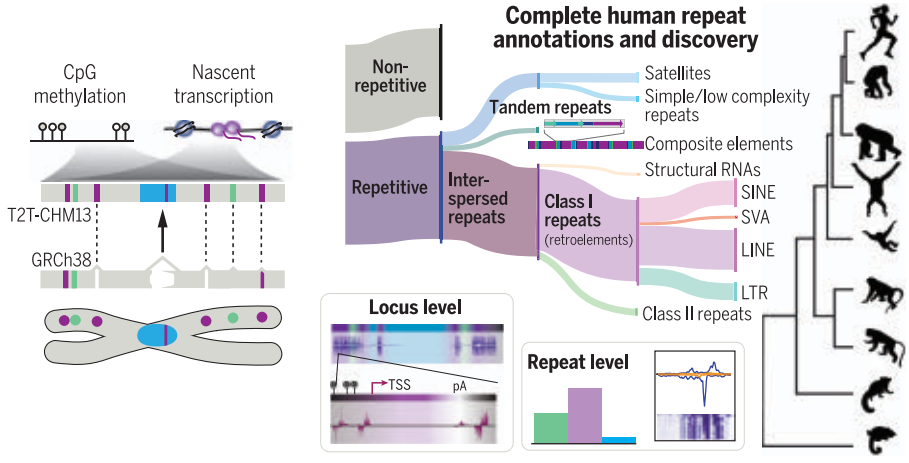
Savannah J. Hoyt, Jessica M. Storer†, Gabrielle A. Hartley†, Patrick G. S. Grady†, Ariel Gershman†, Leonardo G. de Lima, Charles Limouse, Reza Halabian, Luke Wojenski, Matias Rodriguez, Nicolas Altemose, Arang Rhie, Leighton J. Core, Jennifer L. Gerton, Wojciech Makalowski, Daniel Olson, Jeb Rosen, Arian F. A. Smit, Aaron F. Straight, Mitchell R. Vollger, Travis J. Wheeler, Michael C. Schatz, Evan E. Eichler, Adam M. Phillippy, Winston Timp, Karen H. Miga, Rachel J. O'Neill*

INTRODUCTION: Transposable elements (TEs), repeat expansions, and repeat-mediated structural rearrangements play key roles in chromosome structure and species evolution, contribute to human genetic variation, and substantially influence human health through copy number variants, structural variants, insertions, deletions, and alterations to gene transcription and splicing. Despite their formative role in genome stability, repetitive regions have been relegated to gaps and collapsed regions in human genome reference GRCh38 owing to the technological limitations during its development. The lack of linear sequence in these regions, particularly in centromeres, resulted in the inability to fully explore the repeat content of the human genome in the context of both local and regional chromosomal environments.

RATIONALE: Long-read sequencing supported the complete, telomere-to-telomere (T2T) assembly of the pseudo-haploid human cell line CHM13. This resource affords a genome-scale assessment of all human repetitive sequences, including TEs and previously unknown repeats and satellites, both within and outside of gaps and collapsed regions. Additionally, a complete genome enables the opportunity to explore the epigenetic and transcriptional profiles of these elements that are fundamental to our understanding of chromosome structure, function, and evolution. Comparative analyses reveal modes of repeat divergence, evolution, and expansion or contraction with locus-level resolution.

RESULTS: We implemented a comprehensive repeat annotation workflow using previously known human repeats and de novo repeat modeling followed by manual curation, including assessing overlaps with gene annotations, segmental duplications, tandem repeats, and annotated repeats. Using this method, we developed an updated catalog of human repetitive sequences and refined previous repeat annotations. We discovered 43 previously unknown repeats and repeat variants and characterized 19 complex, composite repetitive structures, which often carry genes, across T2T-CHM13. Using precision nuclear run-on sequencing (PRO-seq) and CpG methylated sites generated from Oxford Nanopore Technologies long-read sequencing data, we assessed RNA polymerase engagement across retroelements genome-wide, revealing correlations between nascent transcription, sequence divergence, CpG density, and methylation. These analyses were extended to evaluate RNA polymerase occupancy for all repeats, including high-density satellite repeats that reside in previously inaccessible centromeric regions of all human chromosomes. Moreover, using both mapping-dependent and mapping-independent approaches across early developmental stages and a complete cell cycle time series, we found that engaged RNA polymerase across satellites is low; in contrast, TE transcription is abundant and serves as a boundary for changes in CpG methylation and centromere substructure. Together, these data reveal the dynamic relationship between transcriptionally active retroelement subclasses and DNA methylation, as well as potential mechanisms for the derivation and evolution of new repeat families and composite elements. Focusing on the emerging T2T-level assembly of the HG002 X chromosome, we reveal that a high level of repeat variation likely exists across the human population, including composite element copy numbers that affect gene copy number. Additionally, we highlight the impact of repeats on the structural diversity of the genome, revealing repeat expansions with extreme copy number differences between humans and primates while also providing high-confidence annotations of retroelement transduction events.

CONCLUSION: The comprehensive repeat annotations and updated repeat models described herein serve as a resource for expanding the compendium of human genome sequences and reveal the impact of specific repeats on the human genome. In developing this resource, we provide a methodological framework for assessing repeat variation within and between human genomes. The exhaustive assessment of the transcriptional landscape of repeats, at both the genome scale and locally, such as within centromeres, sets the stage for functional studies to disentangle the role transcription plays in the mechanisms essential for genome stability and chromosome segregation. Finally, our work demonstrates the need to increase efforts toward achieving T2T-level assemblies for nonhuman primates and other species to fully understand the complexity and impact of repeat-derived genomic innovations that define primate lineages, including humans. ■



Telomere-to-telomere assembly of CHM13 supports repeat annotations and discoveries. The human reference T2T-CHM13 filled gaps and corrected collapsed regions (triangles) in GRCh38. Combining long read–based methylation calls, PRO-seq, and multilevel computational methods, we provide a compendium of human repeats, define retroelement expression and methylation profiles, and delineate locus-specific sites of nascent transcription genome-wide, including previously inaccessible centromeres. SINE, short interspersed element; SVA, SINE–variable number tandem repeat–*Alu*; LINE, long interspersed element; LTR, long terminal repeat; TSS, transcription start site; pA, polyadenylation signal.

The list of author affiliations is available in the full article online.
***Corresponding author. Email: rachel.oneill@uconn.edu**
†These authors contributed equally to this work.
Cite this article as S. J. Hoyt *et al.*, *Science* 376, eabk3112 (2022). DOI: 10.1126/science.abk3112

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abk3112

RESEARCH ARTICLE

HUMAN GENOMICS

From telomere to telomere: The transcriptional and epigenetic state of human repeat elements

Savannah J. Hoyt[1], Jessica M. Storer[2]†, Gabrielle A. Hartley[1]†, Patrick G. S. Grady[1]†, Ariel Gershman[3]†, Leonardo G. de Lima[4], Charles Limouse[5], Reza Halabian[6], Luke Wojenski[1], Matias Rodriguez[6], Nicolas Altemose[7], Arang Rhie[8], Leighton J. Core[1,9], Jennifer L. Gerton[4], Wojciech Makalowski[6], Daniel Olson[10], Jeb Rosen[2], Arian F. A. Smit[2], Aaron F. Straight[5], Mitchell R. Vollger[11], Travis J. Wheeler[10], Michael C. Schatz[12], Evan E. Eichler[11,13], Adam M. Phillippy[8], Winston Timp[3,14], Karen H. Miga[15], Rachel J. O'Neill[1,9,16]*

Mobile elements and repetitive genomic regions are sources of lineage-specific genomic innovation and uniquely fingerprint individual genomes. Comprehensive analyses of such repeat elements, including those found in more complex regions of the genome, require a complete, linear genome assembly. We present a de novo repeat discovery and annotation of the T2T-CHM13 human reference genome. We identified previously unknown satellite arrays, expanded the catalog of variants and families for repeats and mobile elements, characterized classes of complex composite repeats, and located retroelement transduction events. We detected nascent transcription and delineated CpG methylation profiles to define the structure of transcriptionally active retroelements in humans, including those in centromeres. These data expand our insight into the diversity, distribution, and evolution of repetitive regions that have shaped the human genome.

Studies of mobile elements and repeat arrays have long shown that eukaryotic genomes are in constant flux (*1*). Transposable element (TE) insertions and repeat-mediated structural rearrangements can influence gene regulation, create new coding structure, and affect chromosome stability. Transposition, expansion, and contraction of repeats generate species-specific genomic innovations (*1*, *2*), major evolutionary transitions (*3*), and human- and primate-specific adaptations (*4*). Together, TEs and other forms of repetitive DNA, constituting more than half of the human genome, are the largest contributor to human genetic variation and affect human health (*5*) owing to their roles in deleterious copy number variants (CNVs), structural variants (SVs), insertions, deletions, and alterations to gene transcription and splicing.

A major challenge in tracking and understanding repeat structure, function, and variation is that large complex repeats, sequences in tandem arrays, and recent insertions by TEs have been largely impenetrable to available sequencing and assembly technologies. Despite this challenge, a species-agnostic repeat database (the Dfam database) (*6*), manual curation (*7*), and the development of improved algorithms for repeat discovery (*8*, *9*) have laid the groundwork underlying efforts to create and finish a complete map and catalog of the repertoire of human repeats.

Previous assemblies of a reference human genome contained gaps and collapsed repeats (*10*). Capitalizing on recent advances in ultralong sequencing and assembly methods, the Telomere-to-Telomere (T2T) Consortium generated a complete human reference genome on the basis of the pseudo-haploid genome of an androgenetic hydatidiform mole (CHM13hTERT cell line, hereafter CHM13) (*11*). This assembly, T2T-CHM13v1.1, resulted in the addition of more than 200 mega–base pairs (Mbp) of DNA and resolution of collapsed and unassembled regions in previous reference genomes. The gap-filled and decompressed regions, representing 8% of the human genome, are dominated by tandemly arrayed repeats [such as in the alpha satellite arrays that are found in higher-order repeat arrays (HORs) within centromeres (*12*)] and complex repeats in pericentromeres, subtelomeres, and some chromosome arms (i.e., acrocentrics). T2T-CHM13 supported additional annotations for human repetitive sequences residing in previously unassembled regions of the human reference GRCh38 and added repeat annotations for low copy repeats genome-wide. In total, we identify 53.9% of the T2T-CHM13 assembly as repetitive. Here we highlight key advances from this resource, while illustrating the power of combining multiple approaches and tools to enhance genomic discoveries.

Eukaryotic repeats are classified into two main types on the basis of their genomic organization: tandem repeats and interspersed repeats (Fig. 1A) (*6*, *13*). Tandem repeats are further subdivided into satellites and simple repeats; satellites are often further defined by their regional chromosomal distribution (centromeric, for example). With the exception of pseudogenes retroposed from structural RNAs (tRNAs, ribosomal RNAs, etc.), interspersed repeats largely refer to TEs, which are classified on the basis of their mechanism of propagation (*6*, *14*–*16*). Class I elements are spread within genomes via retrotransposition and are further subdivided into two broad subclasses. One subclass consists of long interspersed elements (LINEs) and long terminal repeat (LTR) elements, which typically encode their own catalyzing enzymes. The other consists of short interspersed elements (SINEs) and the composite retroelement SINE-VNTR-*Alus* (SVAs), both of which are nonautonomous, relying on LINE-encoded proteins for retrotransposition. Class II elements are those that are mobilized through transposase, helicase, or recombinase and include TEs such as Tc1-Mariner and hAT. These varied repeat types constitute a major portion, and in some cases the majority [for example, 85% of wheat genomes (*17*)], of eukaryotic genome sequences.

The varied modes of propagation of such repeats, from simple insertion events to promoting nonallelic recombination, facilitate genomic diversity, often in bursts of activity followed by periods of neutral evolution. Furthermore, organismal defense mechanisms that have evolved to counter the deleterious effects of mobilization, such as DNA methylation, can influence the sequence evolution of targeted elements. Repeats represent the nexus of evolutionary forces, the selfishness of mobile elements, and the cellular mechanisms marshaled to silence them. The genomic turbulence engendered by repeats makes them the most challenging genomic regions to study. However, insights from studies of these regions have revealed regulatory and coding domains critical to organismal life histories and human health. A full accounting of repeat domains permitted by a gapless telomere-to-telomere

[1]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [2]Institute for Systems Biology, Seattle, WA, USA. [3]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. [4]Stowers Institute for Medical Research, Kansas City, MO, USA. [5]Department of Biochemistry, Stanford University, Stanford, CA, USA. [6]Institute of Bioinformatics, Faculty of Medicine, University of Münster, Münster, Germany. [7]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA. [8]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [9]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA. [10]Department of Computer Science, University of Montana, Missoula, MT, USA. [11]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [12]Department of Computer Science and Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [13]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [14]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA [15]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [16]Department of Genetics and Genome Sciences, UConn Health, Farmington, CT, USA.

***Corresponding author. Email: rachel.oneill@uconn.edu**

†These authors contributed equally to this work.

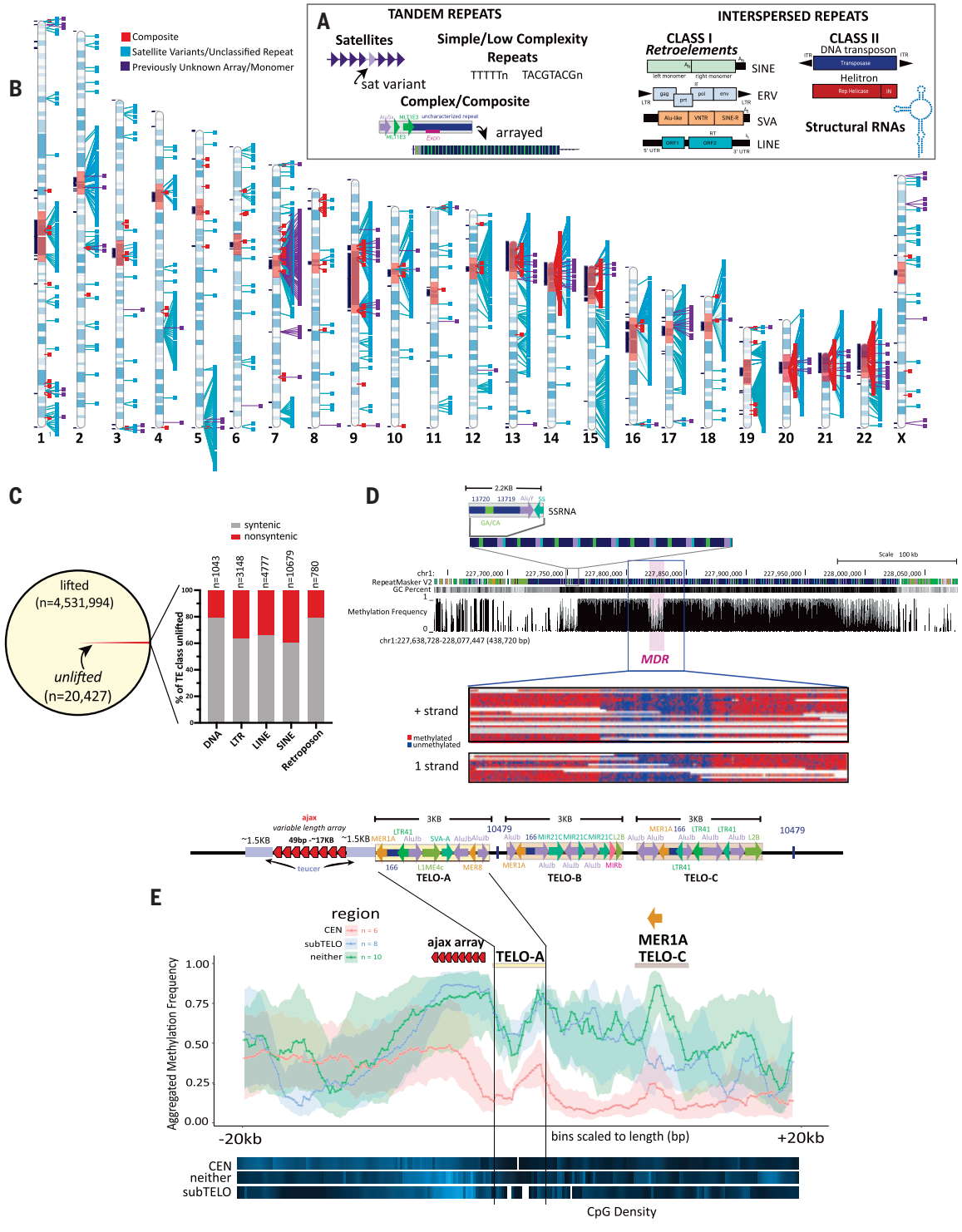


Fig. 1. T2T-CHM13 assembly supports identification of previously unknown repeat families and complex epigenetic signatures. (**A**) Schematic illustrating examples of tandem repeats, including satellites, simple and low complexity repeats and composites, and interspersed repeats, including class I and class II TEs, and structural RNAs. (**B**) Ideogram of CHM13 indicating the locations of annotated composite elements (red), satellite variants and unclassified repeats (aqua), and arrays or monomers of sequences found within those arrays (purple). Gaps in GRCh38 with no synteny to T2T-CHM13 (*11*) are shown in black boxes to the left of each chromosome, centromere blocks [including centromere transition regions (*12*)] are indicated in orange. (**C**) (Left) The number of TEs lifted and unlifted from T2T-CHM13 to GRCh38. (Right) Bar plot showing percentage of TEs by class (DNA, LTR, LINE, SINE, and retroposon) that were unlifted from T2T-CHM13 gap-filled regions (nonsyntenic, red) and syntenic regions (gray); the *n* values show the number of elements within each class affected. (**D**) (Top) T2T-CHM13 genome browser showing the 5SRNA_Comp subunit structure and array. RepeatMaskerV2 track, CG percentage, and methylation frequency tracks are shown. The MDR is indicated. (Bottom) A zoomed image of individual nanopore reads showing consistent hypomethylation in the MDR (chr1:227,818,289-227,830,789) and hypermethylation in the flanking regions (chr1:227,804,021-227,845,689). Both positive (top) and negative (bottom) strand aligning reads show the same methylation pattern. (**E**) (Top) Each T2T-CHM13 TELO-composite element consists of a duplication of a teucer repeat (blue) separated by a variable 49-bp (ajax) repeat array (red arrowheads) and three different composite subunits (TELO-A, -B, and -C). Repeat and TE annotations are shown. Some copies of TELO-composite contain the previously unknown repeat "10479" between the TELO-A and TELO-C subunits and/or after the TELO-C subunit. (Bottom) Metaplot of aggregated methylation frequency (average methylation of each bin across the region, 100 bins total) centered on the TELO-A subunit, ±20 kbp, grouped by chromosomal location (orange, centromeric; blue, subtelomeric; green, interstitial). CpG density for each group is indicated at the bottom (white, no CpG; dark blue, low CpG; bright blue, high CpG). The location of the ajax repeat array and the MER1A element within the TELO-C subunit are indicated.

DNA sequence is therefore essential to a full understanding of the origins and function of the human genome.

Results

Comprehensive repeat annotations for a complete human genome

We developed a computational pipeline to discover previously unknown repeat annotations and tandem arrays while reducing false positives from pseudogenes, segmental duplications, and Dfam overlaps (*18*) (fig. S1). At each step, computational analysis was supplemented by manual curation and polishing. In total, 49 previously unidentified repeat types from RepeatModeler were curated, including 27 repeats (Fig. 1B and fig. S2) as well as 22 potentially older TE repeats whose alignment scores precluded classification and were thus set aside (table S1). Among the 27 identified repeats were one previously unknown centromeric satellite [86.6% of base pairs found within centromere regions defined in (*11*, *12*)] and 10 repeats classified into five variants of known satellites [three centromere transition satellites (GSATII, HSAT5v1, and HSAT5v2) and two interstitial satellites (SATR1 and SST1)] and five previously unknown repetitive sequences. Manual curation identified an additional 13 interstitial satellite arrays (and monomers of the satellites); three repetitive sequences, all of which were previously unknown and unclassified; and 19 composite elements (including 16 curated composite subunits), defined as a repeating unit consisting of three or more repeated sequences, including TEs, simple repeats, composite subunits, and satellites (fig. S2). In total, 62 repeat entries were classified and submitted to Dfam as previously unannotated human repeats, with 19 elements added as a "composite" track for the T2T-CHM13v1.1 genome browser (Table 1 and table S2).

This updated repeat library yielded annotations of human repeats within regions previously unresolved in GRCh38 and provided copy number support to identify additional, previously unnoticed repeat elements genome-wide (Fig. 1B and fig. S2). Using this T2T-CHM13–based repeat library, the T2T-CHM13v1.1 assembly was fully annotated for all repeat classes, resulting in 1.65 giga–base pairs (Gbp) of repeat annotations (53.94% of the genome), of which 168.3 Mbp are found within the 182.1 Mbp of gap-filled T2T-CHM13 genomic sequence (92.4%), representing added annotations, and 5.5 Mbp of which are previously unknown human repeats that we identified genome-wide (Table 1; tables S2 and S3; and fig. S3). Reannotation of GRCh38 (without the Y chromosome) using the T2T-CHM13 repeat database resulted in annotation of 2,114,766 bp of previously uncataloged repeats (Table 2 and table S3), demonstrating the utility of a T2T-level assembly in supporting more comprehensive repeat annotations. Additionally, reannotation of the GRCh38 Y chromosome revealed previously unidentified annotations consisting of six composite elements, eight satellite arrays, 156 satellite variants, and six unclassified repeats, totaling 161,055 bp in repeat annotations discovered through this study (fig. S4 and table S4).

The reannotated GRCh38 and annotated T2T-CHM13 were compared with reverse liftOver coordinates (CHM13 to GRCh38) to identify TE insertions specific to CHM13 (*18*). TEs found in CHM13 but not in GRCh38 were further grouped into those that are in gap-filled regions (nonsyntenic overlap) or those that are potentially polymorphic between these two genomes or were collapsed in the GRCh38 assembly (syntenic overlap but missing in GRCh38) (Fig. 1C).

Across 4,531,994 TEs with lifted coordinates (i.e., shared between T2T-CHM13 and GRCh38), 118,787 lifted TE pair annotations were discordant between the two genomes (fig. S5A); 82.3% of these (97,719 discordant liftOver pairs) were typically short loci with low scores and therefore of questionable discordance (*19*), and/or subtle subfamily reclassifications (fig. S5B and table S5). Among the 20,427 unlifted TEs specific to T2T-CHM13, all TE classes are represented (Fig. 1C), with 35.2% of TE sites specific to gap-filled regions in T2T-CHM13 (7194 total TEs) (tables S5 and S6). Unlifted TE sites are found genome-wide, with a higher density on the acrocentric chromosomes 13, 14, 15, 21, and 22 (fig. S5C and table S7).

Table 1. Complete genome assembly supported discovery and refinement of human repeat annotations. Repeats identified through RepeatModeler and manual curation (RMv2) shown in counts and base pairs, by category, for T2T-CHM13v1.1 and GRCh38 (excluding the Y chromosome) (Fig. 1B and table S2).

Repeat category	CHM13v1.1		GRCh38 (excluding Y)	
	RMv2		RMv2	
	Count	Bp	Count	Bp
Composite subunits	4,446	2,805,296	1,162	536,979
Unclassified repeats	1,234	1,025,084	783	570,985
Satellite variants	730	568,841	671	591,294
Monomers of arrayed satellites	11,900	1,127,758	2,651	415,508
Total	18,310	5,526,979	5,267	2,114,766

Composite elements shape the human genome and local methylation

Composite structural elements contribute to human diversity and disease through structural variation and copy number variation, particularly when exonic regions are "captured" in a core unit (*20*). We annotated 19 composite repeat elements (table S2 and figs. S6 to S11) in T2T-CHM13, each composed of three or more repeated sequences, including TEs, simple repeats, composite subunits, and satellites (*18*). Most composites are found in a tandem array only on a single chromosome (figs. S6, A to F, and S7, B to G), and in eight cases, each core unit contains protein-coding annotations (fig. S7), indicating that unequal crossing-over events and concerted evolution among composite units contribute to the expansion or contraction of gene families within humans (table S2).

One composite, 5SRNA_Comp, consists of a portion of the *5S* RNA, an *AluY*, and two subunit repeats as an array of 128 repeating units with high sequence similarity (most share 98 to 100% identity) on chromosome 1 (fig. S9, A and B). Using methylation profiles developed for T2T-CHM13 and long read–based methylation clusters (*21*), we find that the methylation pattern of the 5SRNA_Comp is not consistent across the array; rather we find a drop in methylation, which we called a methylation dip region (MDR), internal to the array, similar to the centromere dip region (CDR) identified in higher-order arrays of alpha satellites in T2T-CHM13 (*21*) (Fig. 1D). The location of the MDR is not linked to DNA sequence, as neither the GC content nor sequence identity is variable across repeat units in this array (Fig. 1D and fig. S9B), suggesting that other epigenetic factors may facilitate the drop in methylation.

We annotated a highly complex composite, TELO_Comp, that consists of multiple satellite arrays and other composites (Fig. 1E), with instances found on 10 chromosomes (figs. S12 and S13) at interstitial, pericentromeric, and subtelomeric loci. The canonical TELO_Comp consists of three 3-kbp (kilo–base pair) composites (TELO-A, -B, and -C subunits), each containing multiple TEs, downstream of a variable-length array of a 49-bp satellite repeat unit, ajax, bounded by a duplicated sequence, teucer (Fig. 1E). In-depth analysis of the overall structure of the subunits across all loci and phylogenetic analyses of the TELO-A subunit (*18*) (fig. S12 and table S8) indicate that subtelomeric units are a monophyletic group of recent origin, likely by segmental duplication

Table 2. Repeat annotations are more refined for CHM13v1.1. Kilo–base pairs of repeat annotations, by repeat class and family, for different human genome assemblies with T2T-CHM13v1.1 RMv2, GRCh38 RMv2, and GRCh38 Dfam3.3 only. Note that *Alu*Jb is included in the *Alu* repeat family category.

Repeat class	Repeat family	CHM13v1.1		GRCh38 (excluding Y)			
		RMv2		RMv2		Dfam3.3	
		Kbp	% of assembly	Kbp	% of assembly	Kbp	% of assembly
SINE	5S-Deu-L2	221.0	0.0072	218.2	0.0075	210.9	0.0072
	Alu	308,309.7	10.0926	304,734.4	10.4283	305,457.4	10.4531
	MIR	80,937.9	2.6495	80,726.2	2.7625	79,989.0	2.7373
	tRNA-Deu	48.4	0.0016	48.1	0.0016	46.1	0.0016
	tRNA-RTE	549.0	0.0180	546.9	0.0187	525.9	0.0180
	tRNA	1.6	0.0066	1.6	0.0069	1.5	0.0067
Retroposon	SVA	4,654.7	0.1524	4,507.8	0.1543	4,520.7	0.1547
LINE	CR1	10,817.9	0.3541	10,805.0	0.3698	10,571.1	0.3618
	Dong-R4	120.1	0.0039	121.0	0.0041	115.2	0.0039
	I-Jockey	15.7	0.0005	15.6	0.0005	14.8	0.0005
	L1	512,421.5	16.7742	507385.6	17.3633	507,866.7	17.3797
	L1-Tx1	49.6	0.0016	50.3	0.0017	49.1	0.0017
	L2	104,083.4	3.4072	103,819.6	3.5528	102,055.0	3.4924
	RTE-BovB	872.7	0.0286	875.3	0.0300	843.2	0.0289
	RTE-X	3,195.6	0.1046	3,190.0	0.1092	3,097.8	0.1060
	Penelope	68.0	0.0022	68.4	0.0023	63.1	0.0022
LTR	ERV1	83,480.5	2.7327	82,370.2	2.8188	82,641.0	2.8281
	ERVK	8,611.5	0.2819	8,370.9	0.2865	8,468.0	0.2898
	ERVL	59,049.8	1.9330	58,682.4	2.0082	58,646.0	2.0069
	ERVL-MaLR	110,751.8	3.6255	110,098.6	3.7677	109,957.5	3.7629
	Gypsy	4,843.6	0.1586	4,826.3	0.1652	4,629.3	0.1584
	Undefined	3,172.7	0.1039	3,176.1	0.1087	3,081.7	0.1055
DNA	Crypton	44.4	0.0015	45.1	0.0015	44.5	0.0015
	Crypton-A	21.7	0.0007	22.0	0.0008	20.6	0.0007
	Kolobok	65.7	0.0021	65.7	0.0022	63.9	0.0022
	MULE-MuDR	1,008.0	0.0330	985.5	0.0337	986.8	0.0338
	Merlin	40.3	0.0013	40.6	0.0014	40.0	0.0014
	PIF-Harbinger	68.5	0.0022	70.0	0.0024	67.6	0.0023
	PiggyBac	540.6	0.0177	541.4	0.0185	539.5	0.0185
	TcMar	44.9	0.0015	45.5	0.0016	45.2	0.0015
	TcMar-Mariner	2,961.4	0.0969	2,888.2	0.0988	2,872.1	0.0983
	TcMar-Pogo	4.5	0.0001	4.2	0.0001	4.0	0.0001
	TcMar-Tc1	135.8	0.0044	135.8	0.0046	134.1	0.0046
	TcMar-Tc2	1,678.1	0.0549	1,666.6	0.0570	1,665.6	0.0570
	TcMar-Tigger	37,999.9	1.2439	37,725.0	1.2910	37,557.8	1.2853
	hAT	561.8	0.0184	561.6	0.0192	543.1	0.0186
	hAT-Ac	653.4	0.0214	634.7	0.0217	610.5	0.0209
	hAT-Blackjack	2,608.8	0.0854	2,603.9	0.0891	2,589.8	0.0886
	hAT-Charlie	46,980.9	1.5379	46,779.6	1.6008	46,468.2	1.5902
	hAT-Tag1	476.1	0.0156	476.5	0.0163	463.2	0.0159
	hAT-Tip100	12,250.5	0.4010	12,034.5	0.4118	11,915.0	0.4077
	hAT-hAT19	1.8	0.0001	1.7	0.0001	1.6	0.0001
	Undefined	1,201.8	0.0393	1,202.9	0.0412	1163.1	0.0398
Kilo–base pairs of TEs		1,405,826.3		1,393,370.3		1,390,841.6	
Kilo–base pairs of non-TEs		241,986.3		122,998.7		118,576.2	
Assembled kilo–base pairs		3,054,815.5		2,922,175.7		2,922,175.7	
TEs masked		46.02		47.68		47.60	
Non-TEs masked		7.92		4.21		4.06	
% repeatmasked		53.941		51.892		51.654	

events (fig. S13A and tables S8 and S9), whereas interstitial and pericentromeric units are polyphyletic. Moreover, each subtelomeric unit contains the ajax array proximal to the telomere, indicating that inverted orientations are favored at subtelomeric loci. Location-specific repeat diversification in subunit content and structure as well as ajax and teucer repeat copy numbers, which each retain high sequence identity (figs. S13, B and C, S14, and S15, and tables S10 and S11), reveal differential evolutionary forces acting on TELO_Comp loci on the basis of chromosome location.

Meta-analysis of aggregated methylation frequency across the TELO_Comp units (±20 kbp)

(Fig. 1E) shows that the ajax satellite array is hypermethylated across all elements, with a discernible drop in methylation across TELO-A subunits and peak of methylation in the MER1A unit in elements containing TELO-C. Subtelomeric and interstitial TELO_Comp elements share similar methylation profiles, with higher methylation levels across the entire element, whereas pericentromeric TELO_Comp units have lower overall methylation levels. This indicates that local epigenetic states affect overall methylation levels but do not change relative levels within the ajax array and TELO subunits. Comparison of aggregated methylation frequency across TELO_Comp units at the same loci in the human diploid assembly for HG002 (fig. S16) (*21*) show that overall methylation levels are higher across TELO_Comp elements, including those found in centromeres, as expected from global differences in methylation level between T2T-CHM13 and HG002. However, the overall methylation pattern for the TELO_Comp elements (Fig. 1E and fig. S16) is retained, indicating it is an epigenetic signature of this repeat in humans.

Transcriptional, epigenetic, and structural differences define TEs across the human genome

Precision nuclear run-on sequencing (PRO-seq) (*22*) detects nascent transcription from RNA polymerases with nucleotide resolution at genome scale. The resulting read density profiles quantitatively reflect the occupancy of active polymerases across the genome. Sites of accumulating RNA polymerase activity (*22*, *23*), such as promoter-proximal pause sites, 3′ cleavage and polyA regions, splice junctions, and enhancers, indicate points of transcription regulation (*22*, *24*). In addition, because PRO-seq captures RNA synthesis before mechanisms that affect RNA stability take place, unprocessed and unstable RNAs can be detected with high sensitivity. Capitalizing on the single-base resolution of PRO-seq and CpG methylation profiles (*21*), we define profiles of RNA polymerase activity that distinguish different families of retroelements (Fig. 1A). We assessed PRO-seq signal, CpG methylation density, CpG site density, and sequence divergence from the consensus for each element within each subfamily, further classified as full-length or truncated and grouped by relative age (fig. S17 and tables S12 and S13) (*18*). For each element type, density profiles were correlated with known features of specific repeats.

Across all full-length retroelements in T2T-CHM13, PRO-seq density profiles show signals of RNA polymerase accumulation (Fig. 2, A to E, and fig. S18). *AluY* elements show two signal peaks; the first corresponds to the known RNA pol III promoter site within the first monomer, while the second, broader peak within the second monomer indicates the site of a second, ancient 7SL RNA promoter (*25*), whose presence might promote polymerase pausing (Fig. 2A). The peak distribution closely mimics the relative size of the left and right *Alu* monomers and thus reflects the dimerization of *Alu*. Although active transcription continues in truncated *AluY* elements, there is no longer a visible signal of promoter exclusivity, and RNA polymerase signal spreads across the element. Full-length *AluY* elements retain a similar methylation profile and show low divergence levels corresponding with low, single-copy *k*-mer density. Truncated and older elements (*Alu*J and *Alu*S) (table S13 and figs. S18 to S22) show broad methylation profiles with low CpG content and higher divergence (Fig. 2A). Transcriptionally active retroelement families wherein the majority of full-length elements show high PRO-seq signal (*AluY*, SVA, and L1Hs; Fig. 2, purple lines in parallel plots) do have some full-length members that exhibit the full diversity of transcriptional activity, likely influenced by local chromatin or epigenetic features of surrounding insertion sites.

Whereas PRO-seq signal is detected in truncated HERV-Ks (human endogenous retrovirus type K) that retain LTRs [LT (less than 7500 bp in length)/LTR+] (*18*), signal is reduced and completely lost in truncated elements without LTRs, as expected (*26*). Full-length HERV-Ks [GT (greater than 7500 bp in length)/LTR+] (*18*) generally have low methylation levels despite higher CpG content than the LT HERV-Ks, albeit with nonsignificant *P* values (Fig. 2B and figs. S18 to S20). Given the low number of HERV-K elements and high identity among 5′ and 3′ LTRs of HERV-K elements (GT range 0.21 to 23%, average 12.05%; LT range 1.98 to 28.96%, average 11.58%), discerning a clear 5′ promoter signal was not possible. Furthermore, SVA_E and SVA_F elements, the only SVA elements in the human genome that retain mobility (*27*, *28*), both show similar PRO-seq peaks (Fig. 2, C and D), which distinguishes them from their truncated counterparts SVA_A, SVA_B, SVA_C, and SVA_D (figs. S18 to S22).

We find evidence for RNA polymerase promoter proximal pausing at the 5′ end of the SVA element at predicted transcription start sites (TSSs) (*29*). Notably, we find PRO-seq peak signal at the 3′ end within the HERV-K/LTR5a–derived portion of the element, overlapping with the Kruppel-associated box (KRAB)–containing zinc finger proteins (KZFPs) controlled enhancer activity (TEEnhancer) identified in this region (Fig. 2, C and D, gray arrowheads) that contributes to human-specific early embryonic transcription (*30*). While some truncated SVA_F elements retain the 5′ promoter signal, most SVA elements retain the 3′ signal (Fig. 2, C and D, and figs. S18, S20, and S21) and thus may also retain the ability to modulate gene expression.

L1Hs elements, a major contributor to human structural variation (*31*), show a strong promoter-proximal pause signal at the 5′ end (*32*) (Fig. 2E). This site also contains a methylation peak followed by a hypomethylated TSS, delineating full-length L1Hs elements from their truncated counterparts (Fig. 2E and figs. S18 to S22). As elements become inactivated through 5′ truncation (*33*, *34*) and increased divergence, CpG content and transcriptional signal drops considerably (Fig. 2, E and F, and figs. S18 to S22), indicating that CpGs are likely targeted for methylation and subsequent deamination from cytosine to thymine.

To extend our analyses and demonstrate the applicability of this approach in studying other complex repeats in the human genome, we focused on the TE-derived macrosatellite SST1 [also called MER22 (*35*) and NBL2 (*36*, *37*)]. SST1 has demonstrated meiotic instability (*38*), and its methylation status is of clinical relevance to multiple cancer types (*39*–*42*). SST1 arrays are variable in the human population (*38*), and our annotations identify about a two-fold increase over the 342 loci (315,515 bases) (table S14) identified in GRCh38 (excluding the Y chromosome, which carries an additional 587 loci) (fig. S4). Randomized Axelerated Maximum Likelihood (RAxML) phylogenetic analysis with representative loci subsampled from the 16 autosomes on which SST1 resides (*18*) (Fig. 3, A and B, and table S15) showed that the array situated on the long (q) arm of chromosome 19 represents the ancestral SST1 in the human genome and carries a propensity for centromere seeding and array size expansions or contractions across primate lineages (*35*, *43*).

The number of overlapping PRO-seq reads, average methylation, and percent divergence for each SST1 element in CHM13 were compared to delineate correlations among transcriptional, epigenetic, and structural features of SST1 across genomic loci. PRO-seq revealed that the SST1 arrays on chromosome 4 and centromeric monomers on chromosomes 9, 13, and 14 are highly transcribed in comparison to other SST1 loci and are grouped in a single phylogenetic cluster (Fig. 3, A, C, and D; fig. S23; and table S16), indicating that centromeric SST1 repeat arrays are transcriptionally inactive in CHM13.

Statistical analyses of SST1 repeats showed that the highly transcribed repeats are both longer and less diverged from the consensus sequence (*t* test, $P < 0.0001$) (Fig. 3C, fig. S24, and table S17) despite their basal location in the phylogenetic tree (Fig. 3A). CpG methylation levels are high (>50%) for SST1 within chromosome 4 and 19 arrays, low (<50%) for centromeric monomers, and variable (low and high) for centromeric arrays (Fig. 3, A and D; figs. S24 and S25; and tables S16 and S17). Metaplots of aggregated methylation frequency across SST1 repeat units support this

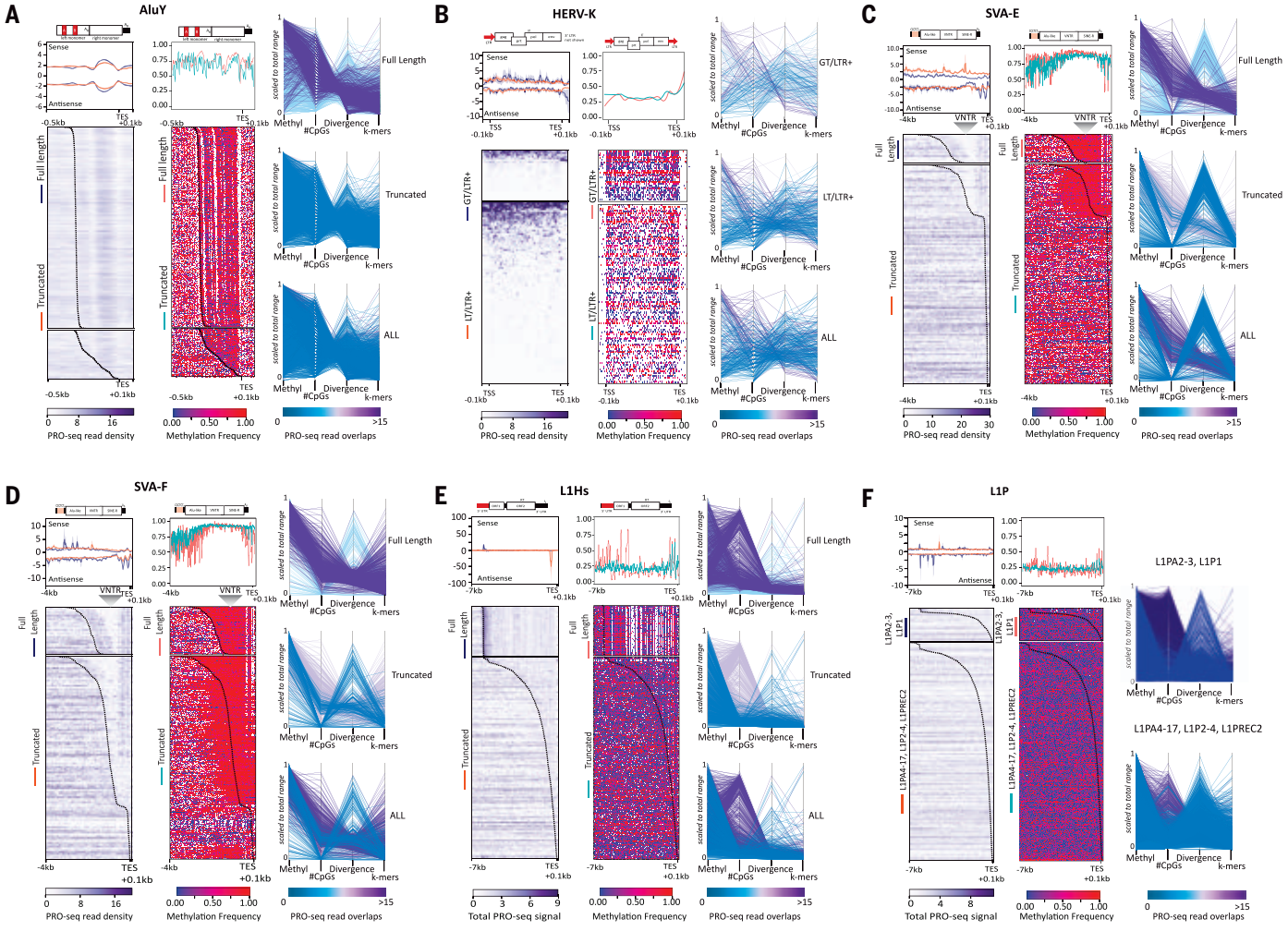


Fig. 2. Transcriptional profiles of TEs are highly correlated with sequence divergence and epigenetic features. (**A** to **F**) RNA polymerase occupancy, methylation levels, CpGs, and divergence for (A) *AluY*, (B) HERV-K, (C) SVA-E, (D) SVA-F, (E) L1Hs, and (F) L1P elements from CHM13. Heatmaps of (left panel) T2T-CHM13 PRO-seq density (Bowtie2 default "best match," purple scale) and average profiles showing sense and antisense strands (upper panels, standard error shown in gray) and (right panel) methylated CpGs (red–purple scale, aggregated frequency per site) for TEs grouped by their length [(A) to (E)] [full-length (FL) and truncated (TR)] or L1PA subfamily [(F), all truncated)]. HERV-K groups are delineated as follows: >7500 bp elements (GT) and <7500 bp elements (LT) with both 5′ and 3′ long-terminal repeats (LTR+). (HERV-K elements with only one or no LTR are shown in fig. S18C). Both GT and LT/LTR+ HERV-K elements are scaled. All other TEs are anchored to the 3′ end, with a specified distance from the anchor (bottom left). Standard error for composite (gray), TSS (transcription start site), TES (transcription end site), location of the VNTR (variable number tandem repeat) within SVA are indicated. A dotted line is included on the heatmap denoting the static −0.1 kbp from the end of the annotated element. Representative schematic of elements and respective subcomponents are shown above the composite profile, scaled to the TES; red blocks indicate previously known promoter regions. (Right side of each panel) Parallel plots for each TE are shown, highlighting each group of TEs (FL/TR, or L1P subfamily; HERV-K plots represent LTRs only). Vertical axes represent scaled values for average methylation, number of CpG sites, and divergence from RepeatMasker consensus sequences for each instance of the element. Coloration by the number of overlapping PRO-Seq reads where purple represents the highest read overlap and blue the lowest, on the scale matching each plot.

observation and indicate that while interstitial arrays and monomeric SST1s carry the same methylation frequency at their 5′ end, monomeric SST1s lose most methylation across the body of the element (Fig. 3E and figs. S25 and S26). Irrespective of this methylation pattern, heatmaps of PRO-seq density show that all highly transcribed SST1s have two internal peaks of high RNA polymerase occupancy that are closely spaced and in opposite orientations (Fig. 3D and fig. S25B), characteristic of RNA pol II promoters and enhancers.

Together, these data suggest selective pressure to retain the genomic integrity of older, less diverged SST1 arrays and monomers that are actively transcribed, whereas silenced repeats found in centromeric arrays are more susceptible to sequence variation. Contrary to expectations that CpG methylation renders repeats transcriptionally silent (*44*, *45*), we find that high levels of average methylation across interstitial, arrayed SST1s define these transcribed repeats on chromosome 4 (Fig. 3, A, D, and E) and bear a resemblance to methylation patterns observed over gene bodies (*46*, *47*). Chromosomal instability and cancerous phenotypes associated with demethylation and/or transcription of SST1 repeats have been reported (*48*, *49*), indicating a need to delineate patient-specific and locus-specific annotations for SST1 (*39*, *50*).

The transcriptional landscape of human centromeres

Centromere transcription is integral to proper centromere function, affecting the loading

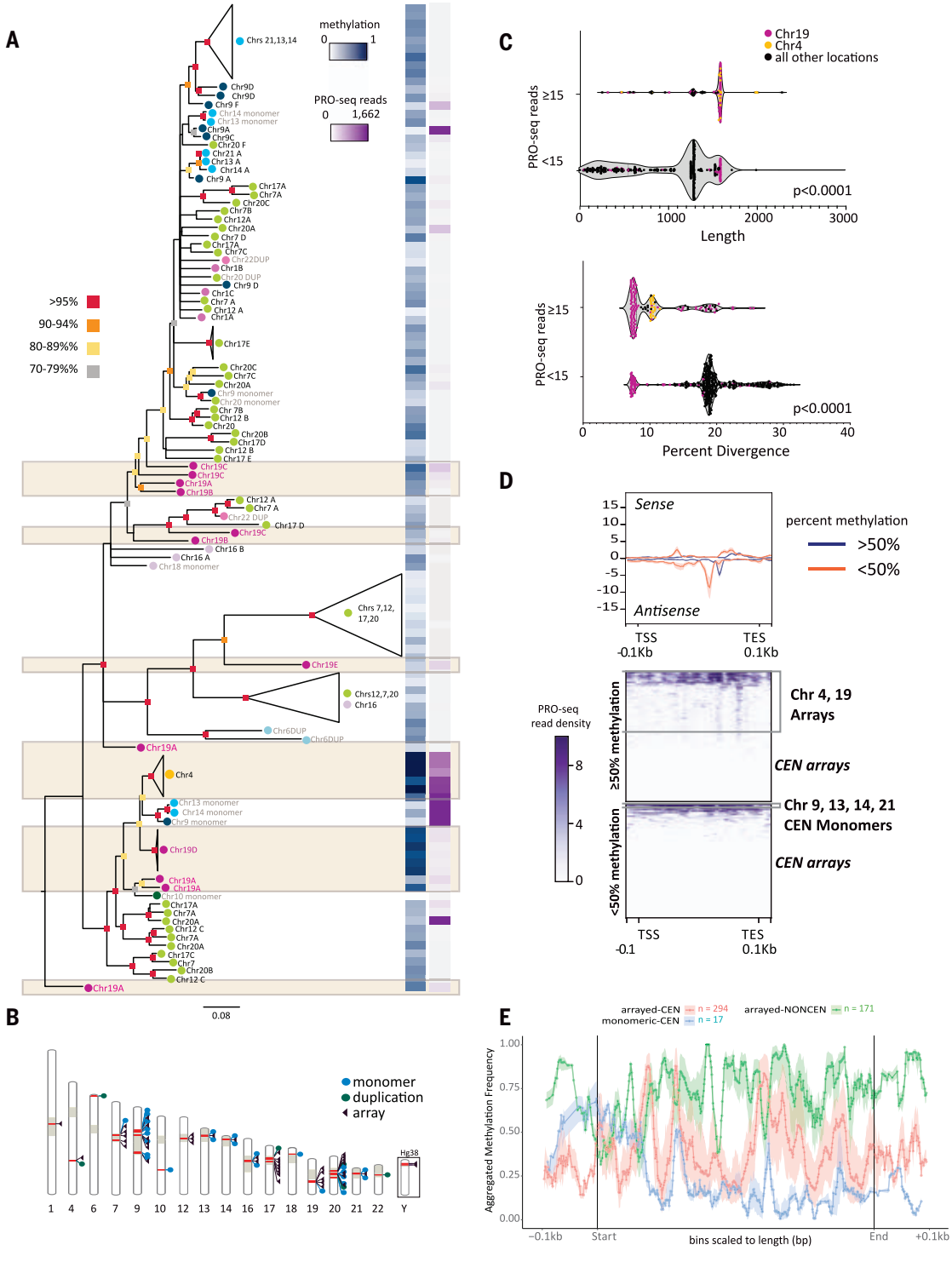


Fig. 3. Transcriptional, epigenetic, and structural differences define SST1 elements across the human genome. (**A**) RAxML phylogenetic analysis of SST1 elements [subsampled to represent each chromosomal location and aligned using MAFFT (*107*)] (tables S14 to S17). Bootstrap values are indicated by color (as per key to the left) at the base of each node. Branch lengths indicate distances and unresolved nodes were collapsed. "Chr#" followed by letters A to F indicates the array designation by T2T-CHM13 chromosome unless SST1 is present as a monomer or as duplicons (DUP) (indicated in gray text). Colored circles by chromosome labels indicate phylogenetic clusters (e.g., chromosomes 7, 12, 17, and 20 in green and chromosomes 13, 14, and 21 in aqua). (Right) For each SST1 sequence or group of collapsed sequences on the tree, average methylation frequency (0, hypomethylated; 1, hypermethylated) is indicated in blue, and PRO-seq read coverage is indicated in purple as per key inset. Tan boxes denote noncentromeric arrays. (**B**) The location of SST1 elements across T2T-CHM13 is indicated by red bars within the chromosome schematic (table S14). Tan blocks indicate centromeres and centromere transition regions as per (*12*). SST1 arrangement as a single monomer (blue dot), duplication (green dot), or array (purple triangle) is indicated. Locations of SST1 arrays on the Y chromosome are shown for GRCh38 (CHM13 is 46,XX). (**C**) Violin plot of SST1 elements shows statistically significant differences between expression levels (repeat overlap of PRO-seq reads, Bowtie2 default "best match") and length of the element (*t* test, $P < 0.0001$) as well as percent divergence (*t* test, $P < 0.0001$). Dot colors indicate interstitial arrays on chromosome 19 (purple) and chromosome 4 (yellow) with a read overlap higher than 15. All other locations with a read overlap lower than 15 are indicated in black. Fifteen read overlap cutoffs determined by analyzing the range of read overlap among all SST1s (fig. S23). (**D**) T2T-CHM13 PRO-seq profiles (Bowtie2 default "best match," upper panel) of SST1 grouped by average methylation levels (<50% and > 50%). Each element is scaled to a fixed size with standard error shading (gray), TSS, TES, and ±0.1 kbp are shown (bottom). Heatmaps (lower panels) of PRO-seq density (purple scale, normalized reads per million aggregate for sense and antisense) grouped by average methylation levels (>50%, top; <50%, bottom). Clusters of specific SST1 loci are indicated to the right. (**E**) Metaplot of aggregated methylation frequency (100 bins total) of SST1 elements (500 bp to 2 kbp), ±0.1 kbp, grouped by chromosomal location and arrayed versus monomeric or duplicated [orange, centromeric (CEN) array; blue, centromeric monomer; green, noncentromeric array]. Truncated noncentromeric/CEN monomers and duplications not shown; length filtering resulted in $n = 1$.

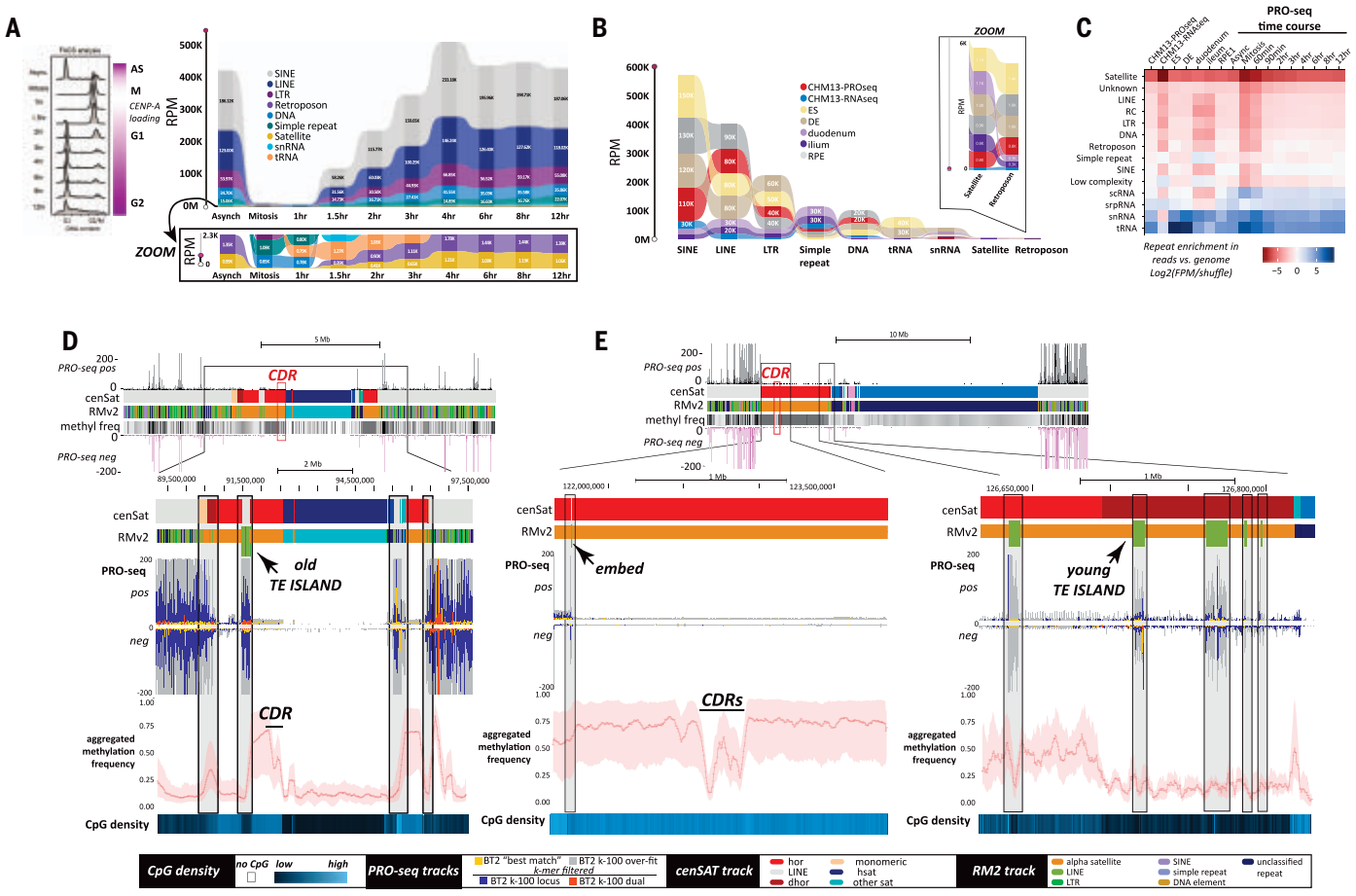


Fig. 4. Centromere landscape is characterized by the transcription of TEs rather than satellites. (**A**) (Left) Cell sorting data showing the stages of the cell cycle after synchronization and release. (Right) Ribbon plots of repeat abundance in PRO-seq data [shown as reads per million (RPM)] assessed by CASK method in asynchronous and synchronized HeLa cells collected at time points across the cell cycle (key in inset). A zoomed image shows the reads for the lower range of expressed repeats, including all satellites classified in T2T-CHM13 (tan). (**B**) Ribbon plot of repeat abundance in PRO/ChRO-seq data, shown as RPM, assessed by CASK method across different developmental stages and samples. Datasets include T2T-CHM13 PRO-seq and native RNA-seq, PRO-seq for RPE-1 (differentiated retinal pigment epithelial cells), and ChRO-seq for H9 ES (embryonic stem cells), DE (differentiated endoderm cells), duodenum tissue, and ileum tissue. A zoomed image shows the reads for the lowest of categories of repeats across all samples, including the satellites classified in T2T-CHM13. (**C**) Repeat enrichment across PRO-seq and RNA-seq datasets (all times points and tissues) ranked from least (red) to most enriched (blue) on the basis of *k*-mers normalized to genomic frequency in T2T-CHM13. (**D** and **E**) Recently active retroelements (green ticks in RM2 track) found embedded within alpha satellite HOR arrays (red) in (D) an "old" TE island derived from segmental duplications on chromosome 3 and (E) solo embedded TEs and "young" TE islands on chromosome 1. Stranded PRO-seq profiles (Bowtie2 default "best match") across chromosome 3 and 1 regions encompassing the centromere are shown (top). TEs are transcriptionally active (PRO-seq Bowtie2 "best match" mapping (yellow), k-100 overfit mapping (gray), and single (blue) and dual filtered (red) k-100 mapping data are indicated for both strands) and located (black boxes) at transitions in CpG methylation (metaplot at bottom; 200 bins total) and CpG density (blue, below) within the array. Key of elements in cenSAT and RM2 tracks indicated at bottom.

of newly synthesized centromere protein A (CENP-A) histones (*51*–*57*). Although evidence suggests that RNA is a critical component of the epigenetic cascade leading to faithful CENP-A assembly, an assessment of nascent transcription across human centromeres has been lacking. The availability of high-confidence centromere annotations for T2T-CHM13 (*12*, *58*, *59*) provides an opportunity to assess transcription and active RNA polymerase activity across previously unresolved regions of a human genome reference: the centromere and the pericentromere. To capitalize on the T2T-level assembly and the resolution of PRO-seq at single nucleotides, we developed genome-dependent and genome-independent approaches to define the landscape of centromere transcription (*18*) (figs. S27 and S28).

We observed low levels of satellite transcription (figs. S29 to S33 and table S18), indicating that RNA polymerase occupancy at centromeric satellites in CHM13 is lower than that observed for all other repeat types. The low levels of satellite transcription are not explained by differences in genomic abundance between satellite repeats and other repeats. Indeed, after normalizing the observed PRO-seq levels with shuffled reads, satellite transcription is the lowest among all other repeat types (fig. S33), indicating genome-wide repression of centromere satellite transcription, including the CENP-A–containing HORs (*12*).

Given that centromere transcription and CENP-A deposition are dynamic processes (*60*), we tested whether repeat transcription varied across the cell cycle. After synchronization and release into mitosis, we find that repeat transcription across the genome drops in mitosis (Fig. 4A and fig. S34). SINEs, LINEs, and LTRs increase transcription rates at the 1-hour time point and reach a steady state by 1.5 hours, coincident with the transition to G1 after CENP-A loading. Notably, satellite transcripts are detected, but at low levels across the cell cycle (Fig. 4A and figs. S29 to S34). We

used available datasets to determine whether the low level of satellite transcription was specific to CHM13 or its early developmental stage. Across cell types and developmental stages, retroelements show dynamic PRO-seq profiles, yet satellite transcription remains low (Fig. 4B and figs. S35 and S36). Across all cell types and time points, alpha satellites within the CENP-A–containing HORs (*12*) show generally higher PRO-seq signal than do degenerate HOR alpha satellite arrays (dHORs) and monomers or interstitial alpha satellites (MONs) (fig. S37). Thus, although nascent transcription is low, transcription from alpha satellites is detectable within the HOR domain that demarcates the active centromere (Fig. 4C). The low level of detectable transcripts within the active HOR domains contrasts with the transcriptional level of pericentromeric satellite arrays where satellite transcripts promote the recruitment of chromatin modifiers to maintain the heterochromatic status of these domains (*61*).

TE annotations for T2T-CHM13 show that members of retroelement subfamilies known to contain full-length and, in some cases, transpositionally active members are found within centromeric HOR satellite arrays and retain their PRO-seq signal (fig. S38 and table S19). We find evidence for multiple types of TE–alpha satellite associations across T2T-CHM13 (table S19); all chromosomes have TE insertions within alpha satellites, but several lack TEs within HORs (e.g., all acrocentric chromosomes). We also find "older" TE islands within HORs, derived from segmental duplications (Fig. 4D and fig. S39), recent insertions of TEs within HORs, and aggregates of TEs that appear to form emerging TE islands (Fig. 4E, right, and fig. S40). Single insertions of TEs found within HORs, dHORs, and monomeric regions (table S19) remain transcriptionally active (Fig. 4, D and E, and fig. S38) yet show limited evidence of transcription of adjacent alpha satellites (Fig. 4E and figs. S39 and S40), indicating that read-through transcription from embedded TEs may affect alpha satellites, but not in the arrays underlying the CDR, the region defined by CENP-A enrichment (Fig. 4E, left) (*12*).

Given the higher proportion of L1Hs insertions in HORs and work showing a link between L1 transcription and neocentromere formation (*57*, *62*), we compared embedded L1Hs within HORs to those found in dHORs, monomers, and chromosome arms to determine whether L1Hs embeds retained their TE signatures or were "overwritten" by their local chromatin environment. We find no statistical evidence that L1Hs within HORs and dHORs deviate in length, divergence, or average methylation from those found outside of these regions (figs. S41 and S42 and table S20). However, L1Hs within monomeric segments of alpha satellites are both more diverged and less methylated than L1Hs that are in HORs ($P < 0.05$), dHORs ($P < 0.01$), or not embedded at all ($P \leq 0.001$) and show less transcription than their counterparts elsewhere in the genome, including those in the HOR and dHOR (figs. S38 and S42).

Although we find no clear link between alpha satellite transcription and the CENP-A domain overlapping the CDR (*12*, *21*), transcription detected from embedded TEs marks shifts in methylation frequencies across satellite domains, establishing putative TE boundaries. Whether and how TEs facilitate these shifts is unknown. In previous work, the activity and copy number of TEs has been linked to alterations in methylation levels within centromeres in interspecific hybrids, resulting in chromosome instability (*63*), indicating that a balance of methylation is required for centromere stabilization. With the technological advances presented in the assembly and annotation of the T2T-CHM13 human reference, comparative studies across other species will aide in revealing how the structure of the satellite-dense centromeres of human differs from that of TE-enriched centromeres in other species (*64*) and how these differences affect centromere function and chromosome evolution.

Putative TE-driven genomic DNA transductions and their evolutionary consequences

The complete sequence provided by T2T-CHM13 revealed previously unknown patterns of repeat expansions across the short (p) arms of acrocentric chromosomes. In T2T-CHM13 (*11*), we discovered previously unannotated repeat arrays of a 64-nucleotide sequence (Fig. 1B) present in high copy numbers on the p arms of acrocentric chromosomes 14, 15, 21, and 22 (*11*) and in single or low copy number (<5) on eight other chromosomes (Fig. 5A and tables S2 and S21). A solo monomer resides on chromosome 10, with all other occurrences adjacent to an *Alu*Sx3 element (thus, with *Alu* satellite, or WaluSat). The lack of identity among 5′ and 3′ sequences of the chromosome 10 locus and the *Alu*Sx-WaluSat loci on all other chromosomes (fig. S43), coupled with phylogenetic analyses across primates (figs. S43 to S46), indicates that an ancestral duplication of the chromosome 10 locus was followed by a mobile element insertion to form the *Alu*Sx-WaluSat unit in the last shared ancestor with Catarrhini.

The WaluSat sequence exists as a single monomer at eight loci, as a duplication at three loci, and in one case as a pentamer. However, once segmental duplication events placed the *Alu*Sx-WaluSat on the p arms of chromosomes 14, 15, 21, and 22, WaluSat amplified into longer arrays, ranging from 26 copies (chromosome 15) to 5836 copies (chromosome 14) (Fig. 5A). We hypothesize that the high degree of sequence similarity and copy number variation among p arm WaluSat arrays is due to frequent non-allelic or ectopic recombination events on acrocentric chromosomes (*11*, *65*), which may be exacerbated by replication challenges associated with the predicted periodic G-quadruplex structures (*66*) identified at junctions of WaluSat sequences within arrays (*18*) (Fig. 5, B and C).

The low identity among the sequences adjacent to the chromosome 3 *Alu*Sx-WaluSat and other *Alu*Sx-WaluSat loci, along with the identification of putative target site duplications (TSDs), may indicate that a transduction event followed the *Alu* insertion and preceded the spread across the human genome via duplications. TE-mediated transduction (i.e., a TE transduction event), a process by which retroelements co-mobilize DNA flanking the element to new genomic loci (*67*–*70*), has been observed for L1 and SVA elements in humans (*67*–*73*). TE transduction events mediated by *Alu* elements are seemingly rare (*74*), likely because of efficient termination of RNA polymerase III on sequences with long poly-T tract lengths and nearby RNA secondary structures (*75*). Given the age of the initial insertion of the *Alu*Sx element, it is unknown if such an event was mediated by an RNA polymerase III or cryptic upstream RNA polymerase II promoter, or if other rearrangements specific to chromosome 3 degraded signal of shared identity with other segmental duplications (Fig. 5A, dashed box).

Beyond potentially seeding new repeat sequences across the genome, TE transductions can affect the genome through exon shuffling (*67*, *71*, *76*) and are a possible source of somatic mutations (*74*, *75*). Here, we applied a set of computational approaches (*18*) (fig. S47) to annotate putative TE transduction events in T2T-CHM13. In total, we analyzed 971,993 L1s and 7068 SVAs (figs. S48 to S51). After stringent filtering for potential artifacts, such as segmental duplications and putative duplications of truncated elements, we find 65 L1, five 3′ SVA, and 11 5′ SVA transduction events (tables S21 and S22 and figs. S50 and S51).

Of these 81 annotated transduction events, 78 are shared with GRCh38 (Fig. 5D and table S23), and three appear specific to T2T-CHM13. One T2T-CHM13 TE transduction is in a region of no synteny with GRCh38 and is caused by an L1PA4, representing an older event according to Kimura-2 distances (fig. S49). Of the remaining two T2T-CHM13 TE transductions, both events are derived from the youngest, human-specific TEs, L1Hs and SVA-F, and may represent polymorphic TE transductions. However, we find the offspring TE in both GRCh38 and T2T-CHM13, yet the transduced sequence is missing in GRCh38, owing to a collapse in the sequence, highlighting the utility of a T2T-level assembly in identifying putative TE transduction events.

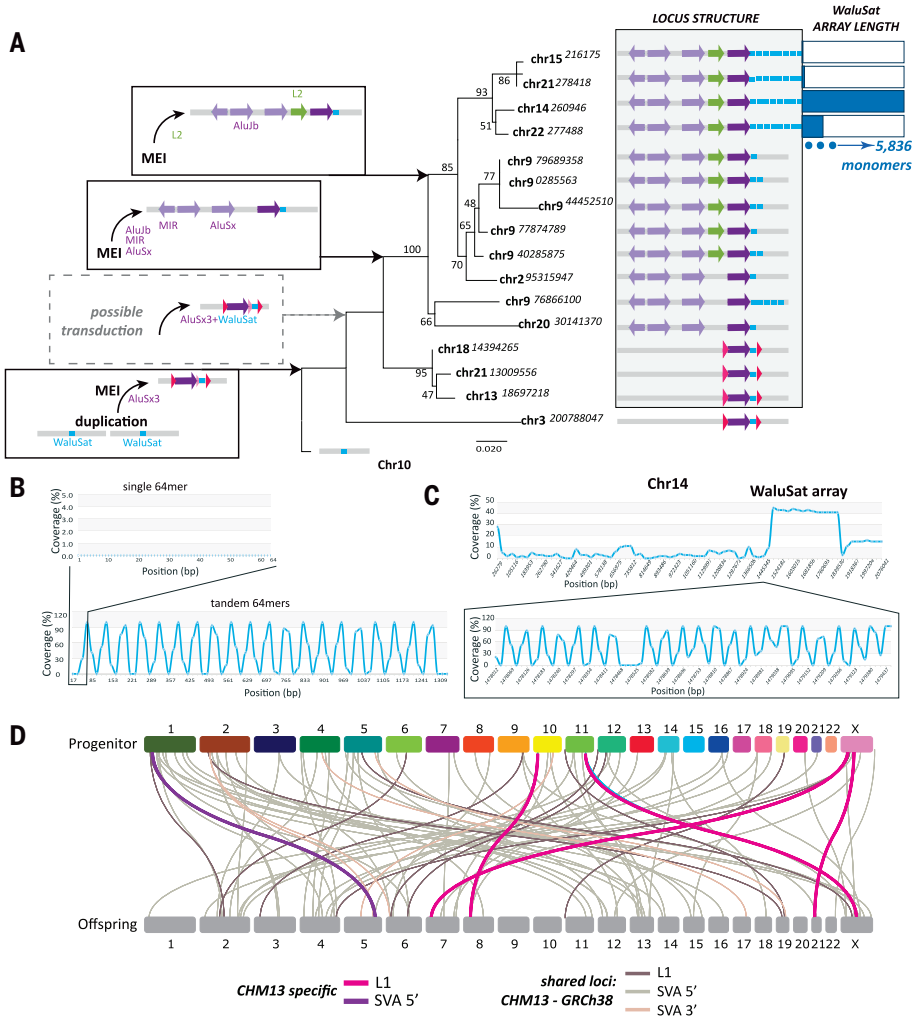


Fig. 5. TE activity affects genomic repeat diversity in CHM13. (**A**) Maximum likelihood (ML) phylogenetic analyses of the *Alu*Sx3-WaluSat locus across T2T-CHM13. Chromosome location is indicated (starting nucleotide position shown) at each branch. Bootstrap values shown at each node, distance indicated by length of branch. Left shows the sequential order of events, initiating with a duplication of the chromosome 10 WaluSat locus followed by mobile element insertion (MEI) of an *Alu*Sx3. The identification of putative TSDs (pink, fig. S43) and a lack of identity among sequences adjacent to WaluSat on chromosome 3 and all other loci (fig. S43) may indicate that a transduction event preceded the spread of *Alu*Sx3-WaluSat across the human genome (dotted box). MEI events upstream of the *Alu*Sx3-WaluSat are concordant with phylogenetic relationships among loci and indicate that the derivation of *Alu*Sx3-WaluSat loci across other chromosomes were the result of segmental duplication events (gray shaded box). Once the *Alu*Sx3-WaluSat was duplicated to the acrocentric chromosomes 14, 15, 21, and 22, a massive expansion of the WaluSat sequence (blue boxes) occurred. The number of WaluSat monomers within each acrocentric array is indicated on the right with monomer number relative to maximum monomer count 5836 on chromosome 14. (**B**) G-quadruplex (G4) analysis of a single 64-mer monomer of the WaluSat sequence showed no predicted G4 structures (top), while an in silico construct of a tandem array of the WaluSat shows high G4 coverage at the junction between individual WaluSat monomers across the array. (**C**) G4 analysis of the p arm of chromosome 14 shows a peak in G4 predictions coincident with the WaluSat array. Bottom is a zoom inset of a subset of the array showing that the junctions between most monomers carry predicted G4 structures. (**D**) Transduction events predicted for CHM13 (L1, pink; SVA 5′, purple) and shared between T2T-CHM13 and GRCh38 (gray shades) are shown. Chromosome connections link progenitor and offspring locations (fig. S49).

CHM13 serves as a reference for comparative repeat analyses across humans and other primate genomes

Studies of the link between TE activity and chromatin states can extend beyond local influences, as exemplified by LINE and SINE transcriptional activity and the chromosome-wide silencing of the X chromosome during X inactivation (*77*–*79*). Two noncoding RNAs on the X chromosome are central to the inactivation of one X in females, *Xist* and *Tsix* (*80*). These two loci overlap one another in a sense and antisense orientation but are in distinct topologically associating domains (TADs); *Tsix* is the antisense repressor of *Xist*, whose up-regulation leads to X inactivation (*81*). The bipartite structure of the locus in two TADs facilitates partitioning of the X inactivation center (XIC) and supports appropriate timing of X inactivation through *Xist* transcription in early development (*82*). Moreover, an early step in the formation of heterochromatin across the inactive X is the silencing of LINEs and SINEs within the *Xist* RNA compartment (*77*).

The scarcity of SNPs (*21*) in T2T-CHM13, coupled with the short reads of PRO-seq data, made it impossible to discern transcripts originating from one X allele versus the other within CHM13. However, we were able to phase reads into their individual alleles, supporting the assessment of methylation differences of TEs between the two X chromosomes in the XIC. PRO-seq signal was found across the *Xist* locus, whereas no signal was detected from the *Tsix* locus, indicating that X inactivation has proceeded, resulting in differential methylation profiles across alleles (Fig. 6A). Low methylation (Fig. 6A, blue block in cluster 2) marks the initiation of *Xist* transcription, followed by high methylation levels across the *Xist*/*Tsix* locus on this allele, inclusive of the interspersed repeats found across the locus (Fig. 6A and table S25). A distinct pause (indicated by a pileup of PRO-seq signal, boxed in Fig. 6A) after the termination signal of the *Xist* transcript unit was found that coincides with the TAD junction and delineates the *Xist* and *Tsix* domains. These data are inconsistent with a report that androgenetic hydatidiform moles lack X inactivation (*83*).

We also compared both the XIC and chromosome-wide repeat content of the chromosome X from T2T-CHM13 and HG002 (XY). As expected, the XIC in HG002 shows high methylation across the locus and only a single allelic cluster, with no detectable transcripts across the *Tsix*/*Xist* domain (Fig. 6B and table S25). Sequence comparison of the 269,020 repeats assessed between the haploid X of HG002 and T2T-CHM13 (Fig. 6C and tables S26 and S27), excluding the pseudoautosomal region (see fig. S52 for T2T-CHM13 PAR annotations), currently unassembled in HG002, uncovered 778 repeat differences, of which 70% were simple repeats and 21% were TEs (64 of which were length outliers) (*18*) (fig. S53). Collectively, these data demonstrate that the depth of repeat annotations based on the T2T-CHM13 assembly can serve as a reference for studying human variation inclusive of repeats that affect local and regional chromatin, gene expression, and gene copy numbers.

While many of the previously unidentified repeat classifications coincide with gaps filled

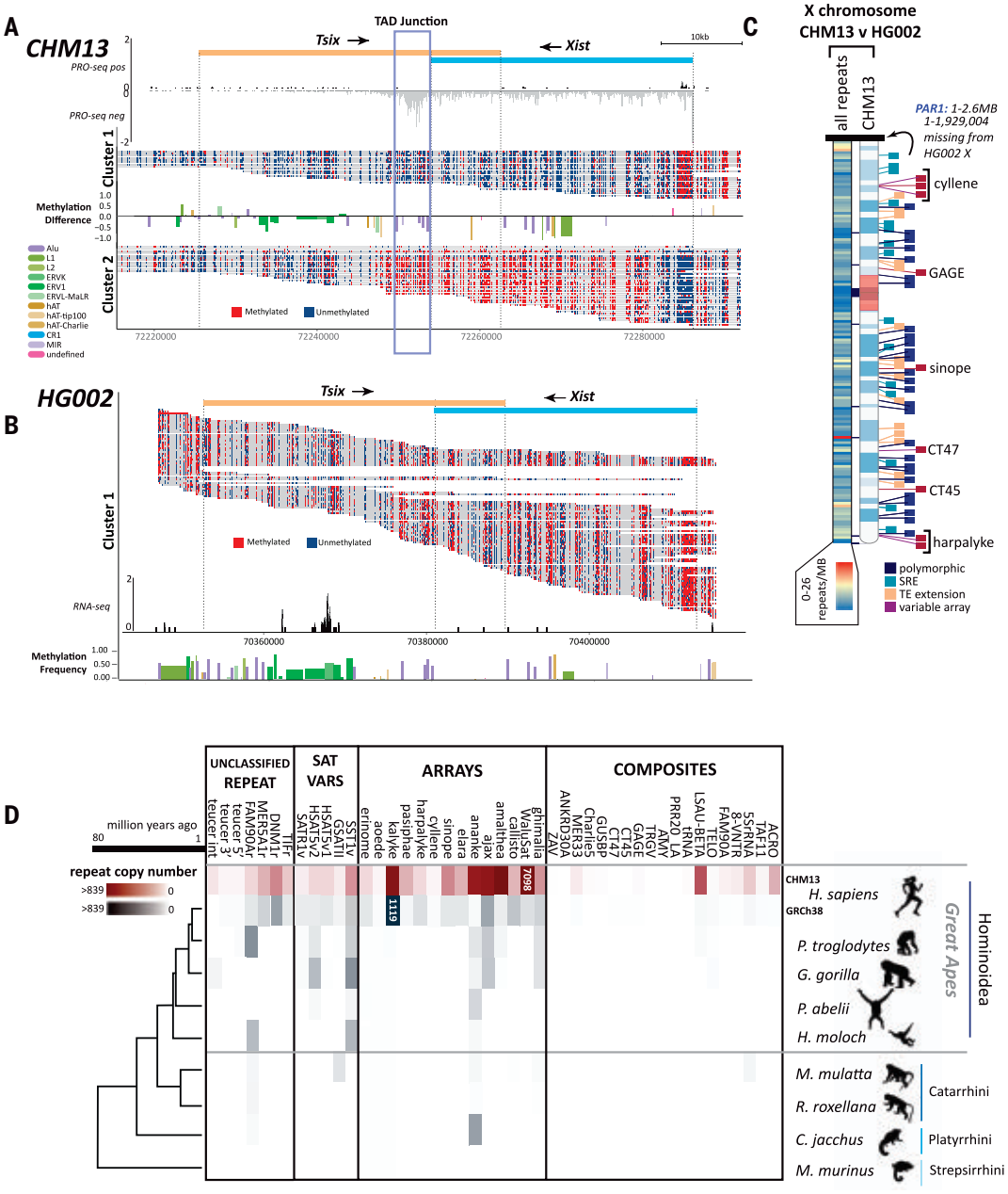


Fig. 6. Repetitive elements define differences between human genomes and nonhuman primates. Single read methylation profiles were extracted, and reads were clustered on the basis of the methylation state of the *Xist* promoter from (**A**) T2T-CHM13 and (**B**) HG002. Differences in repeat methylation were calculated by taking the average methylation per repeat and subtracting cluster 2 repeats from cluster 1 repeats. Directionality of *Xist*/*Tsix* transcript units are indicated (top). Normalized PRO-seq reads show a marked pileup of RNA pol II at the predicted TAD boundary at the 3′ end of the *Xist* transcript [(A), blue box]. (B) Normalized RNA-seq reads across the single cluster for HG002 show no transcriptional signal for *Xist*. (**C**) Heatmap of chromosome X showing the location of all repeat differences between the Xs of HG002 and T2T-CHM13 (left) and the location of the top four categories of repeat differences: polymorphic (insertion/deletion), SRE (short repeat extension), TE extension, and variable array length (right ideogram). Gaps between T2T-CHM13 and GRCh38 are indicated with black blocks between the heatmap and ideogram. (**D**) Copy numbers of previously unknown human repeat annotations identified in T2T-CHM13 grouped by repeats, variants of known satellites, tandemly arrayed sequences, and composite element (inclusive of subunits) for T2T-CHM13 (maroon), GRCh38, and genomes for other primates from the Hominoidea, Catarhini, and Platyrrhini lineages (gray). Heatmap scale denotes number of repeats within the array (0 to 839). Array sizes >839 are indicated within colored blocks. Phylogenetic relationship and millions of years since divergence are indicated on the bottom. Not shown: variants of known centromeric satellites [but see (*12*)] and the repeat annotation for an *AluJb* (*121*) fragment, which could not reliably be delineated in copy number from other closely related full-length *AluJb* elements.

in the T2T-CHM13 assembly, these data supported genome-wide annotation of previously undiscovered repeats and TEs (Fig. 1B). To determine whether these repeat classifications were specific to humans, we searched for orthologous sequences in the human reference GRCh38 and available genome assemblies for primates representing the great apes (*Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii*), Hominoidea (*Hylobates moloch*), Catarrhini (*Macaca mulatta*, *Rhinopithecus roxellana*), Platyrrhini (*Callithrix jacchus*), and Strepsirrhini (*Microcebus murinus*) (*18*).

When comparing copy numbers of repeat annotations between T2T-CHM13 and long-read, high-quality assemblies available for other great apes (chimpanzee, gorilla, and orangutan) (*84*), we still find an increase in copy number across most of the repeats identified herein (Fig. 6D, fig. S2, and table S28). Many repeats appear only as monomers in other primate genomes or are absent in Strepsirrhini, Platyrrhini, Catarrhini, and lesser apes; these reduced counts are largely influenced by the quality of these assemblies and potentially high rates of divergence among repeats, and they highlight the need for telomere-to-telomere assembly approaches for comparative analyses (*85*). Finally, eight of the repeats identified herein are human-specific, with an additional 11 found only as monomers in other species (Fig. 6D and table S28).

Conclusions

The assembly of the complete, telomere-to-telomere human genome reference facilitated development of an atlas of repeats that make up >53% of the human genome. Through this collaborative effort, we have developed a

resource of human repeat annotations and methods to guide future efforts in exploring the complexities of repeat biology in human and other primate genomes. We focused on repeat sequence, CpG methylation, and transcriptional annotation; updated repeat models and implemented repeat modeling tools that supported the identification of previously unknown satellite arrays; expanded the catalog of variants for known repeats and TEs; and developed annotations for complex, composite repeat elements. Deeper exploration of such repeats revealed the complexity of genetic mechanisms that affect repeats during different phases of their life cycle and thus illustrate the myriad mechanisms by which they are major contributors to defining the structure and content of the human genome.

For example, we found that a TE insertion event captured a short sequence, WaluSat, in a primate ancestor. Subsequent segmental duplications of the region carrying this composite TE-sat repeat spread the sequence across several human chromosomes, including four of the acrocentric chromosomes. The satellite portion of the repeat expanded to almost 0.5 Mbp of sequence on the acrocentric chromosomes, resulting in the alteration of the structure of this portion of the chromosome into regions dense with G4s, which are potentially functional elements (*86*). This example highlights the need for future functional studies dissecting the impact of repeats on the local chromosome environment, such as replication timing, local transcription, DNA damage and repair processes, and establishing TAD boundaries. Moreover, this example lays the groundwork for exploring the impact of local environments (such as gene-poor regions as found on the acrocentric arms of human chromosomes) on sequence constraint and mutation rates for emergent repeats.

We provide a high-confidence functional annotation of repeats across the human genome. For example, we find that the tandemly arrayed TE-derived satellite SST1 carries distinctive methylation and transcriptional profiles, including an enhancer embedded in each unit, found only in specific arrays on chromosomes 19 and 4. These arrays are hypervariable in the human population, and alterations in their activity have been linked to cancer (*36*, *48*). However, a full understanding of copy number variation, epigenetic instability, and transcription of SST1 elements has been hampered by a lack of complete annotations of copies of these elements elsewhere in the genome. Our functional annotation revealed transcriptional signatures of both promoters and enhancers within active SST1 elements that may affect local transcription and chromatin structures. Moreover, this enhancer implicates SST1 in defining cellular partitions, such as paraspeckles and phase-separated condensates (*36*, *87*), that could have an impact on other genomic loci.

Combined with defining the linear order and content of centromeric sequences (*12*), we find that engaged RNA polymerase signal is low across centromeric satellites arranged in arrays, irrespective of stages of the cell cycle or development. Rather, active transcription is detected in embedded retroelements coinciding with shifts in methylation states that demarcate active centromere domains. To date, the centromere biology field has been limited by a lack of a linear assembly across human centromeres, challenging the development of models to describe genetic and epigenetic elements that define centromeric chromatin. Our data, in concert with centromere annotations (*12*), reveal that these high-density repeat regions are not static in sequence, epigenetic, or transcriptional activity and that there is a high degree of substructure across the centromeric regions that affect function. Comparing the landscape of the variable centromere forms across domains of life, and in human disease, will reveal the complex life cycle of centromeres (*64*).

Studies of human genetic variation have been relatively blind to repeat variation among individuals, particularly arrayed and complex repeats, as these types of sequences are recalcitrant to short-read sequencing technologies, mapping, and functional annotation methodologies. As a prospective of the utility of complete reference genomes in studying human genetic variation, we compared two T2T X chromosomes. We find 218 kbp of repeat differences between these two chromosomes (0.18% of the chromosome, excluding the 1.9-kbp PAR), including repeat variation in complex arrays that carry exonic material and thus affect gene dosage. Thus, comparative analyses of T2T-level assemblies reveal the potential for discovering an even wider range of repeat variation across the 46 chromosomes that constitute the human genome.

Finally, our work demonstrates the need to increase efforts toward achieving T2T-level assemblies for nonhuman primates to fully understand the complexity and impact of repeat-derived genomic innovations that define primate lineages, including humans. Although we find repeat variants that appear enriched or specific to the human lineage, in the absence of T2T-level assemblies from other primate species, we cannot truly attribute these elements to specific human phenotypes. Thus, the extent of variation described herein highlights the need to expand the effort to create human and nonhuman primate pangenome references to support exploration of repeats that define the true extent of human variation.

Materials and methods summary

Repeat model discovery

RepeatMasker4.1.2-p1 (*88*) with the Dfam3.3 repeat library and RepeatModeler2.0.1 (*8*) were used to define repeats across the genome, further refined using extensive manual curation, as described in (*18*). This database was used to generate a final mask of the T2T-CHM13v1.1 assembly. ULTRA (*9*) was used to improve the accuracy of tandemly repetitive satellite annotations. Gaps of >5 kb in T2T-CHM13v1.1 repeat annotations were identified with BEDtools (*89*) and manually curated. Monomer structure was confirmed using self-alignment plots. Repeat models were further refined to remove any false positives (e.g., fragments of other TEs, pieces of simple repeats), as described in (*18*).

Composite elements

We defined a composite element as a repeating unit consisting of three or more repeated sequences (TEs, simple repeats, subunits, and/or satellites) found as a tandem array in at least one genomic location. A composite subunit is a previously unknown repeat annotation that is found within a composite. Whereas the locations of some composite elements within a family are present as a single copy and thus are likely segmental duplications derived by nonallelic homologous recombination (*90*), a composite family is distinguished by the presence of composite elements in an array in at least one location.

LiftOver/reverse liftOver analyses

LiftOver chains were generated from LASTZ alignments between GRCh38 and T2T-CHM13v1.1 and X chromosomes of T2T-CHM13 and HG002 with considerations as per (*18*). Reverse liftOver was performed from repeat annotations in both assemblies; BEDtools (*89*) was used to intersect the T2T-CHM13 coordinates with regions lacking synteny to GRCh38. Results were parsed into one of five categories: full match (i.e., SINE/*Alu*/*Alu*Sx), class match (i.e., SINE/*Alu*), family match (i.e., SINE), no match, and those set aside and subject to extensive manual curation to identify correct matches.

Methylation metaplots

Nanopore CpG methylation data for T2T-CHM13 and HG002 was processed as in (*21*). CpG methylation frequency was calculated by fraction of methylated reads to total coverage within bins in T2T-CHM13 or HG002 with the BSgenome Bioconductor package (*21*, *91*). Multiples of three bins were further smoothed with the "rollmean" function from the R package Zoo. Methylation clustering was performed by selecting all reads spanning a locus and using the mclust (v5.4.7) R package with the "VII" model to cluster methylation calls across the locus (*92*). CpG density heatmaps were calculated by counting the total number of

CpG sites per position relative to the repeat start and end and dividing by the total number of repeats in each group. Methylation single-read plots were generated in the ggplot2 R package using geom_rect() to plot individual reads with methylated CpGs as red and unmethylated CpGs as blue.

Identification and classification of full-length and truncated TEs

Full-length elements of recently active TE families [*AluY*, L1Hs, HERV-K, and SVA_E/F (*93*)] were retrieved from the RepeatMasker output and cross-referenced with PRO-seq data and CpG methylation data as per (*18*). All retroelement classes were grouped into relative age categories based on divergence and phylogenetic distribution (*6*, *88*, *94*–*99*). LINEs, SINEs, and retroposons were grouped by subfamily; LTRs were grouped by family.

PRO-seq

For each of two PRO-seq replicates, cells were processed as per (*18*, *22*). PRO-seq libraries were prepared as previously described (*22*) with minor modifications (*100*). Permeabilized cells were mixed with permeabilized *Drosophila* S2 nuclei in all 4-biotin-NTP run-on reactions. After amplification, libraries were polyacrylamide gel electrophoresis (PAGE)–purified to remove adapter-dimers and select molecules between 140 and 650 bp. Libraries were sequenced on an Illumina NextSeq 550 (single-end, 75 bp). Raw fastq files were trimmed for quality, length, and adapters using cutadapt (*101*) and reverse complemented using the fastx-toolkit (*102*). Bowtie2 (*103*) alignment to Dm6 was used to remove *Drosophila* spike-in reads; remaining reads were aligned to T2T-CHM13 using default ("best match") parameters (and k-100 for comparison); multimapping alignment files were subjected to single-copy *k*-mer filtering and processed into beds with BEDtools (*89*) for subsequent normalization with nonmitochondrial alignments to obtain counts in reads per million mapped (RPMM) as described in (*18*). Complementary analyses were performed on read data (unmapped) as outlined below and in (*18*).

Statistical analyses and data visualization

BEDtools (*89*) map was used to calculate average methylation and CpG density across all repeats in RepeatMaskerV2 (RMv2) and incorporated into 3D graphs and parallel plots. Genomic data were visualized using RIdeogram (v0.2.2) (*104*), Circos (v0.69-6) (*105*), and Circa (v1.2.2). Genome browser tracks and centromeric satellite (cenSAT) annotations for T2T-CHM13 are as described in (*11*, *12*, *21*, *65*). Heatmaps for PRO-seq profiles were generated using deepTools2 (*106*). Normalized data were binned in 10-bp windows, and repeat elements were anchored to the 3′ end, with the exception of HERV-K, which was divided into subcategories on the basis of length and presence of dual LTRs and scaled as per (*18*). The maximum value per bin and composite profiles were summarized by averaging each bin across all regions in the group; standard error was estimated and is shown in gray in each composite. Methylation heatmaps for HERV-K were generated in R ggplot2 by normalizing repeat size by start and end position and using geom_tile() to plot CpG methylation frequency at each position. For all other elements, methylation heatmaps were anchored at the 3′ and using geom_tile() to plot CpG methylation frequency at each position.

SST1/L1Hs embed analyses

SST1 sequences were extracted from CHM13 annotations via BEDtools (*89*) and aligned with MAFFT (*107*). The evolutionary history was inferred by using RAxML (*108*) and the GTR+G model (*109*) as matched by jModelTest (*110*); 100 bootstrap replicates reported. PRO-seq density for SST1 with <15 and ≥15 reads overlapping were determined by plotting the distribution of read overlaps across all annotated SST1 elements. BEDtools (v2.29.0) (*89*) was used to intersect SST1/L1Hs repeats with genomic locations, methylation (*21*), and transcriptional data. An unpaired *t* test was performed to quantify differences among repeats in each group by repeat length, percent divergence, percent insertions, percent deletions, and average methylation. Violin plots were generated via GraphPad Prism (v9.1.1).

HeLa cell cycle analyses

Given the low rate of cell division and synchronization challenges in CHM13 cells, HeLa-S3 cells were used, noting the caveat that this cell line carries high levels of karyotypic instability (*111*). HeLa-S3 cells were arrested as per (*112*), mitotic cells collected and subsequently grown for the corresponding time or immediately permeabilized (mitotic sample) as described in (*18*). All time points were collected in replicate experiments. Before cellular permeabilization, 10% of each sample was removed, fixed in cold 75% ethanol, and stained with propidium iodide, and DNA content was analyzed using a BD FACSAria II. The flowCore package was used to read FCS files into R. PRO-seq libraries (both replicates) were prepared as previously described (*22*), with minor modifications as for CHM13 (*18*). All data were processed, mapped, and normalized as above for CHM13. Comparative and quantitative analyses are outlined below and described in (*18*).

H9 ChRO-seq data analyses

External chromatin run-on and sequencing (ChRO-seq) data (GSE142316) for four developmental stages in replicate (ES, DE, duodenum, and ileum) (*113*) of H9 cells were used for comparison to CHM13. H9 ChRO-seq data was preprocessed using the proseq2.0 pipeline to generate adapter-trimmed and deduplicated fastq files used for repeat composition analysis as per (*18*).

Preprocessing, mapping, and postprocessing of RNA-seq data (CHM13 and HG002)

Data from two replicates of CHM13 paired-end native RNA sequencing (RNA-seq) using oligoDT (*12*) were processed with the same workflow as the CHM13 PRO-seq data, with minor modifications as per (*18*). External paired-end ribodepleted RNA-seq data for HG002 (GM24385) were used for comparison, preprocessed as per CHM13 RNA-seq and mapped to a combined assembly of T2T-CHM13 autosomes, HG002 chrX, and GRCh38 chrY with Bowtie2.

Comparative analyses of transcript quantification approaches

To complement TE (herein) and centromere satellite repeat annotations (*12*), we implemented a three-pronged approach to define centromere transcription as described in (*18*): a mapping-dependent approach, in which PRO-seq (two replicates) and RNA-seq (two replicates) data were mapped and reads were intersected with single copy *k*-mers derived from the T2T-CHM13 assembly and whole-genome shotgun polymerase chain reaction–free reads (*11*, *114*); a mapping-independent approach in which unmapped PRO-seq and RNA-seq reads were annotated using classification of ambivalent sequences using *k*-mers (CASK) and a T2T-CHM13–dependent *k*-mer database formed via T2T-CHM13 repeat annotations; and a genome-independent approach, in which PRO-seq and RNA-seq reads were processed through RepeatMasker using the human Dfam 3.3 library. RepeatMaskerV2 (RM2) was intersected with cenSAT annotations to identify and label repeats adjacent to alpha satellites designated HOR, dHOR, MON, or "none of the above" regions (RMv2-alpha).

To compare across these three methods, BEDtools (*89*) coverage was used to obtain counts of reads overlapping repeats defined in RMv2 and RMv2-alpha across all mapping methods, requiring at least 50% (~25 to 30 bp, roughly equivalent to the CASK *k*-mer length) of the read to overlap the repeat element [and see (*18*)]. The relative abundance of each repeat was similar across replicates; thus, counts from both replicates were summed. Variable bowtie mapping parameters (default, k-100, and k-100 filtered for single copy *k*-mers with multiple filters) on PRO-seq and RNA-seq datasets were assessed (*18*).

WaluSat analyses

The evolutionary history of WaluSat, *Alu*Sx, and the *Alu*Sx-WaluSat loci were inferred by

using the maximum likelihood method as described (*18*). Dotplots were generated by comparison of 1.5-kb sequences flanking both 5′ and 3′ regions adjacent to WaluSat insertions with FlexiDot as per (*18*). G-quadruplex analysis was performed with G4Hunter (*115*).

Transduction analyses

TE transduction events were analyzed using the modified TSDfinder tool (*67*), filtering for artifacts such as segmental duplications and truncated elements, and refined on the basis of TE age using Kimura-2 distance parameters as described in (*18*).

ChrX liftOver analysis and repeat fasta comparison

Lifted T2T-CHM13 chrX to HG002 coordinates were compared (*18*) using a similarity score as a percentage of the max score (>90% were considered concordant, <50 bp were insufficient; others were considered potentially polymorphic). Sequences of interest were filtered for length differences between the liftOver coordinates. Differences were subject to manual curation depending on repeat type, and the final loci were subjected to RepeatMasker analysis.

Copy number comparison across primates

Copy number comparisons across primate genomes (*18*) were generated with the most recent, available primate genomes for each species: *Pan troglodytes* (accession: GCA_002880755.3) (*84*), *Gorilla gorilla* (accession: GCA_900006655.3) (*116*), *Pongo abelii* (accession: GCA_002880775.3) (*84*), *Hylobates moloch* (accession: GCA_009828535.2), *Macaca mulatta* (accession: GCA_008058575.1) (*117*), *Rhinopithecus roxellana* (accession: GCF_007565055.1) (*118*), *Callithrix jacchus* (accession: GCF_009663435) (*119*), and *Microcebus murinus* (accession: GCF_000165445.2) (*120*). BLAST was used to search each genome for individual instances of the corresponding repeat or composite element, requiring at least an 85% length match to the query repeat/composite monomer and a 100% match requirement across the 85% length for gap tandem arrays.

REFERENCES AND NOTES

1. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017). doi: 10.1038/nrg.2016.139; pmid: 27867194
2. R. Cordaux, S. Udit, M. A. Batzer, C. Feschotte, Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8101–8106 (2006). doi: 10.1073/pnas.0601161103; pmid: 16672366
3. E. V. Koonin, Viruses and mobile elements as drivers of evolutionary transitions. *Philos. Trans. R. Soc. London Ser. B* **371**, 20150442 (2016). doi: 10.1098/rstb.2015.0442; pmid: 27431520
4. A. Koga *et al*., Co-opted megasatellite DNA drives evolution of secondary night vision in Azara's owl monkey. *Genome Biol. Evol.* **9**, 1963–1970 (2017). doi: 10.1093/gbe/evx142; pmid: 28810713
5. D. C. Hancks, H. H. Kazazian Jr., Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016). doi: 10.1186/s13100-016-0065-9; pmid: 27158268
6. J. Storer, R. Hubley, J. Rosen, T. J. Wheeler, A. F. Smit, The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021). doi: 10.1186/s13100-020-00230-y; pmid: 33436076
7. J. D. Fernandes *et al*., The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mob. DNA* **11**, 13 (2020). doi: 10.1186/s13100-020-00208-w; pmid: 32266012
8. J. M. Flynn *et al*., RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020). doi: 10.1073/pnas.1921046117; pmid: 32300014
9. D. Olson, T. Wheeler, ULTRA: A model based tool to detect tandem repeats. *ACM BCB* **2018**, 37–46 (2018). doi: 10.1145/3233547.3233604; pmid: 31080962
10. M. J. P. Chaisson *et al*., Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). doi: 10.1038/nature13907; pmid: 25383537
11. S. Nurk *et al*., The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
12. N. Altemose *et al*., Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
13. Dfam, Transposable Element Classification; www.dfam.org/classification/tree.
14. B. Piégu, S. Bire, P. Arensburger, Y. Bigot, A survey of transposable element classification systems—A call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.* **86**, 90–109 (2015). doi: 10.1016/j.ympev.2015.03.009; pmid: 25797922
15. T. Wicker *et al*., A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007). doi: 10.1038/nrg2165; pmid: 17984973
16. V. V. Kapitonov, J. Jurka, A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* **9**, 411–412 (2008). doi: 10.1038/nrg2165-c1; pmid: 18421312
17. International Wheat Genome Sequencing Consortium (IWGSC), Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018). doi: 10.1126/science.aar7191; pmid: 30115783
18. See supplementary materials and methods.
19. K. M. Carey, G. Patterson, T. J. Wheeler, Transposable element subfamily annotation has a reproducibility problem. *Mob. DNA* **12**, 4 (2021). doi: 10.1186/s13100-021-00232-4; pmid: 33485368
20. P. Pajic *et al*., Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**, e44628 (2019). doi: 10.7554/eLife.44628; pmid: 31084707
21. A. Gershman *et al*., Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
22. D. B. Mahat *et al*., Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016). doi: 10.1038/nprot.2016.086; pmid: 27442863
23. H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013). doi: 10.1126/science.1229386; pmid: 23430654
24. E. M. Wissink, A. Vihervaara, N. D. Tippens, J. T. Lis, Nascent RNA analyses: Tracking transcription and its regulation. *Nat. Rev. Genet.* **20**, 705–723 (2019). doi: 10.1038/s41576-019-0159-6; pmid: 31399713
25. J. O. Kriegs, G. Churakov, J. Jurka, J. Brosius, J. Schmitz, Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet.* **23**, 158–161 (2007). doi: 10.1016/j.tig.2007.02.002; pmid: 17307271
26. P. J. Thompson, T. S. Macfarlan, M. C. Lorincz, Long terminal repeats: From parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol. Cell* **62**, 766–776 (2016). doi: 10.1016/j.molcel.2016.03.029; pmid: 27259207
27. J. Raiz *et al*., The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012). doi: 10.1093/nar/gkr863; pmid: 22053090
28. D. C. Hancks, J. L. Goodier, P. K. Mandal, L. E. Cheung, H. H. Kazazian Jr., Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **20**, 3386–3400 (2011). doi: 10.1093/hmg/ddr245; pmid: 21636526
29. D. C. Hancks, H. H. Kazazian Jr., SVA retrotransposons: Evolution and genetic instability. *Semin. Cancer Biol.* **20**, 234–245 (2010). doi: 10.1016/j.semcancer.2010.04.001; pmid: 20416380
30. J. Pontis *et al*., Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell* **24**, 724–735.e5 (2019). doi: 10.1016/j.stem.2019.03.012; pmid: 31006620
31. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011). doi: 10.1146/annurev-genom-082509-141802; pmid: 21801021
32. G. D. Swergold, Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* **10**, 6718–6729 (1990). pmid: 1701022
33. M. Sokolowski, M. Chynces, D. deHaro, C. M. Christian, V. P. Belancio, Truncated ORF1 proteins can suppress LINE-1 retrotransposition in *trans*. *Nucleic Acids Res.* **45**, 5294–5308 (2017). doi: 10.1093/nar/gkx211; pmid: 28431148
34. B. Brouha *et al*., Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5280–5285 (2003). doi: 10.1073/pnas.0831042100; pmid: 12682288
35. K. Fatyol, K. Illies, A. A. Szalay, D. C. Diamond, C. Janish, Mer22-related sequence elements form pericentric repetitive DNA families in primates. *Mol. Gen. Genet.* **262**, 931–939 (2000). doi: 10.1007/PL00008661; pmid: 10660054
36. R. Nishiyama *et al*., A DNA repeat, NBL2, is hypermethylated in some cancers but hypomethylated in others. *Cancer Biol. Ther.* **4**, 446–454 (2005). doi: 10.4161/cbt.4.4.1622; pmid: 15846090
37. D. Thoraval *et al*., Demethylation of repetitive DNA sequences in neuroblastoma. *Genes Chromosomes Cancer* **17**, 234–244 (1996). doi: 10.1002/(SICI)1098-2264(199612)17:4<234::AID-GCC5>3.0.CO;2-4; pmid: 8946205
38. D. C. Tremblay, G. Alexander Jr., S. Moseley, B. P. Chadwick, Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics* **11**, 632 (2010). doi: 10.1186/1471-2164-11-632; pmid: 21078170
39. J. K. Samuelsson *et al*., Helicase lymphoid-specific enzyme contributes to the maintenance of methylation of SST1 pericentromeric repeats that are frequently demethylated in colon cancer and associate with genomic damage. *Epigenomes* **1**, 2 (2017). doi: 10.3390/epigenomes1010002; pmid: 31867127
40. S. Igarashi *et al*., A novel correlation between LINE-1 hypomethylation and the malignancy of gastrointestinal stromal tumors. *Clin. Cancer Res.* **16**, 5114–5123 (2010). doi: 10.1158/1078-0432.CCR-10-0581; pmid: 20978145
41. K. Suzuki *et al*., Global DNA demethylation in gastrointestinal cancer is age dependent and precedes genomic damage. *Cancer Cell* **9**, 199–207 (2006). doi: 10.1016/j.ccr.2006.02.016; pmid: 16530704
42. H. Nagai *et al*., A novel sperm-specific hypomethylation sequence is a demethylation hotspot in human hepatocellular carcinomas. *Gene* **237**, 15–20 (1999). doi: 10.1016/S0378-1119(99)00322-4; pmid: 10524231
43. G. A. Hartley, M. Okhovat, R. J. O'Neill, L. Carbone, Comparative analyses of gibbon centromeres reveal dynamic genus-specific shifts in repeat composition. *Mol. Biol. Evol.* **38**, 3972–3992 (2021). doi: 10.1093/molbev/msab148; pmid: 33983366
44. A. P. Bird, Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 94–100 (1995). doi: 10.1016/S0168-9525(00)89009-5; pmid: 7732579
45. J. A. Yoder, C. P. Walsh, T. H. Bestor, Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997). doi: 10.1016/S0168-9525(97)01181-5; pmid: 9260521
46. M. P. Ball *et al*., Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009). doi: 10.1038/nbt.1533; pmid: 19329998
47. R. Lister *et al*., Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009). doi: 10.1038/nature08514; pmid: 19829295
48. G. Dumbović *et al*., A novel long non-coding RNA from NBL2 pericentromeric macrosatellite forms a perinucleolar aggregate structure in colon cancer. *Nucleic Acids Res.* **46**,

5504–5524 (2018). doi: 10.1093/nar/gky263; pmid: 29912433
49. J. Carlevaro-Fita *et al*., Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res.* **29**, 208–222 (2019). doi: 10.1101/gr.229922.117; pmid: 30587508
50. B. González *et al*., Somatic hypomethylation of pericentromeric SST1 repeats and tetraploidization in human colorectal cancer cells. *Cancers* **13**, 5353 (2021). doi: 10.3390/cancers13215353; pmid: 34771515
51. G. O. M. Bobkov, N. Gilbert, P. Heun, Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. *J. Cell Biol.* **217**, 1957–1972 (2018). doi: 10.1083/jcb.201611087; pmid: 29626011
52. S. Rošić, F. Köhler, S. Erhardt, Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. *J. Cell Biol.* **207**, 335–349 (2014). doi: 10.1083/jcb.201404097; pmid: 25365994
53. D. M. Carone *et al*., Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading. *Chromosome Res.* **21**, 49–62 (2013). doi: 10.1007/s10577-013-9337-0; pmid: 23392618
54. S. M. McNulty, L. L. Sullivan, B. A. Sullivan, Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C. *Dev. Cell* **42**, 226–240.e6 (2017). doi: 10.1016/j.devcel.2017.07.001; pmid: 28787590
55. S. Catania, A. L. Pidoux, R. C. Allshire, Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin. *PLOS Genet.* **11**, e1004986 (2015). doi: 10.1371/journal.pgen.1004986; pmid: 25738810
56. C. C. Chen *et al*., Establishment of centromeric chromatin by the CENP-A assembly factor CAL1 requires FACT-mediated transcription. *Dev. Cell* **34**, 73–84 (2015). doi: 10.1016/j.devcel.2015.05.012; pmid: 26151904
57. A. C. Chueh, E. L. Northrop, K. H. Brettingham-Moore, K. H. A. Choo, L. H. Wong, LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLOS Genet.* **5**, e1000354 (2009). doi: 10.1371/journal.pgen.1000354; pmid: 19180186
58. G. A. Logsdon *et al*., The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021). doi: 10.1038/s41586-021-03420-7; pmid: 33828295
59. K. H. Miga *et al*., Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi: 10.1038/s41586-020-2547-7; pmid: 32663838
60. L. E. T. Jansen, B. E. Black, D. R. Foltz, D. W. Cleveland, Propagation of centromeric chromatin requires exit from mitosis. *J. Cell Biol.* **176**, 795–805 (2007). doi: 10.1083/jcb.200701066; pmid: 17339380
61. W. L. Johnson *et al*., RNA-dependent stabilization of SUV39H1 at constitutive heterochromatin. *eLife* **6**, e25299 (2017). doi: 10.7554/eLife.25299; pmid: 28760200
62. A. C. Chueh, L. H. Wong, N. Wong, K. H. A. Choo, Variable and hierarchical size distribution of L1-retroelement-enriched CENP-A clusters within a functional human neocentromere. *Hum. Mol. Genet.* **14**, 85–93 (2005). doi: 10.1093/hmg/ddi008; pmid: 15537667
63. R. J. O'Neill, M. J. O'Neill, J. A. Graves, Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**, 68–72 (1998). doi: 10.1038/29985; pmid: 9590690
64. S. J. Klein, R. J. O'Neill, Transposable elements: Genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res.* **26**, 5–23 (2018). doi: 10.1007/s10577-017-9569-5; pmid: 29332159
65. M. R. Vollger *et al*., Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
66. A. Rizzo *et al*., Stabilization of quadruplex DNA perturbs telomere replication leading to the activation of an ATR-dependent ATM signaling pathway. *Nucleic Acids Res.* **37**, 5353–5364 (2009). doi: 10.1093/nar/gkp582; pmid: 19596811
67. S. T. Szak, O. K. Pickeral, D. Landsman, J. D. Boeke, Identifying related L1 retrotransposons by analyzing 3′ transduced sequences. *Genome Biol.* **4**, R30 (2003). doi: 10.1186/gb-2003-4-5-r30; pmid: 12734010
68. J. L. Goodier, E. M. Ostertag, H. H. Kazazian Jr., Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000). doi: 10.1093/hmg/9.4.653; pmid: 10699189
69. O. K. Pickeral, W. Makałowski, M. S. Boguski, J. D. Boeke, Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000). doi: 10.1101/gr.10.4.411; pmid: 10779482
70. A. Damert *et al*., 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res.* **19**, 1992–2008 (2009). doi: 10.1101/gr.093435.109; pmid: 19652014
71. J. Xing *et al*., Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17608–17613 (2006). doi: 10.1073/pnas.0603224103; pmid: 17101974
72. E. J. Gardner *et al*., The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017). doi: 10.1101/gr.218032.116; pmid: 28855259
73. P. Ebert *et al*., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021). doi: 10.1126/science.abf7117; pmid: 33632895
74. J. M. C. Tubio *et al*., Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014). doi: 10.1126/science.1251343; pmid: 25082706
75. B. Pradhan *et al*., Detection of subclonal L1 transductions in colorectal cancer by long-distance inverse-PCR and Nanopore sequencing. *Sci. Rep.* **7**, 14521 (2017). doi: 10.1038/s41598-017-15076-3; pmid: 29109480
76. J. V. Moran, R. J. DeBerardinis, H. H. Kazazian Jr., Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999). doi: 10.1126/science.283.5407.1530; pmid: 10066175
77. J. C. Chow *et al*., LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**, 956–969 (2010). doi: 10.1016/j.cell.2010.04.042 pmid: 20550932
78. M. F. Lyon, X-chromosome inactivation: A repeat hypothesis. *Cytogenet. Cell Genet.* **80**, 133–137 (1998). doi: 10.1159/000014969; pmid: 9678347
79. J. Chaumeil, P. Le Baccon, A. Wutz, E. Heard, A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.* **20**, 2223–2237 (2006). doi: 10.1101/gad.380906; pmid: 16912274
80. N. Stavropoulos, N. Lu, J. T.Lee, A functional role for *Tsix* transcription in blocking *Xist* RNA accumulation but not in X-chromosome choice. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10232–10237 (2001). doi: 10.1073/pnas.171243598; pmid: 11481444
81. 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; pmid: 22495304
82. J. G. van Bemmel *et al*., The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of *Tsix* and *Xist*. *Nat. Genet.* **51**, 1024–1034 (2019). doi: 10.1038/s41588-019-0412-0; pmid: 31133748
83. X. Chen *et al*., Loss of X chromosome inactivation in androgenetic complete hydatidiform moles with 46, XX karyotype. *Int. J. Gynecol. Pathol.* **40**, 333–341 (2021). doi: 10.1097/PGP.0000000000000697; pmid: 33021557
84. Z. N. Kronenberg *et al*., High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018). doi: 10.1126/science.aar6343; pmid: 29880660
85. H. A. Lewin *et al*., Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4325–4333 (2018). doi: 10.1073/pnas.1720115115; pmid: 29686065
86. W. M. Guiblet *et al*., Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. *Genome Res.* **31**, 1136–1149 (2021). doi: 10.1101/gr.269589.120; pmid: 34187812
87. C. Roden, A. S. Gladfelter, RNA contributions to the form and function of biomolecular condensates. *Nat. Rev. Mol. Cell Biol.* **22**, 183–195 (2021). doi: 10.1038/s41580-020-0264-6; pmid: 32632317
88. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0, 2013–2015; www.repeatmasker.org.
89. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
90. L. Zhang, H. H. S. Lu, W.-Y. Chung, J. Yang, W.-H. Li, Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.* **22**, 135–141 (2005). doi: 10.1093/molbev/msh262 pmid: 15371527
91. H. Pagès, BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs, R package version 1.62.0; https://bioconductor.org/packages/BSgenome.
92. L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016). doi: 10.32614/RJ-2016-021; pmid: 27818791
93. J. Xing *et al*., Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res.* **19**, 1516–1526 (2009). doi: 10.1101/gr.091827.109; pmid: 19439515
94. H. Khan, A. Smit, S. Boissinot, Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16**, 78–87 (2006). doi: 10.1101/gr.4001406; pmid: 16344559
95. A. F. Smit, G. Tóth, A. D. Riggs, J. Jurka, Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246**, 401–417 (1995). doi: 10.1006/jmbi.1994.0095 pmid: 7877164
96. A. V. Furano, D. D. Duvernell, S. Boissinot, L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet.* **20**, 9–14 (2004). doi: 10.1016/j.tig.2003.11.006; pmid: 14698614
97. M. K. Konkel, J. A. Walker, M. A. Batzer, LINEs and SINEs of primate evolution. *Evol. Anthropol.* **19**, 236–249 (2010). doi: 10.1002/evan.20283; pmid: 25147443
98. H. Wang *et al*., SVA elements: A hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005). doi: 10.1016/j.jmb.2005.09.085; pmid: 16288912
99. G. L. Freimanis, thesis, University of Wolverhampton (2008).
100. J. Judd *et al*., A rapid, sensitive, scalable method for Precision Run-On sequencing (PRO-seq). bioRxiv 2020.05.18.102277 [Preprint] (2020). https://doi.org/10.1101/2020.05.18.102277.
101. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011). doi: 10.14806/ej.17.1.200
102. Fastx-toolkit, FASTQ/A short-reads preprocessing tools; http://hannonlab. cshl. edu/fastx_toolkit.
103. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: 10.1038/nmeth.1923; pmid: 22388286
104. Z. Hao *et al*., *RIdeogram*: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020). doi: 10.7717/peerj-cs.251; pmid: 33816903
105. M. Krzywinski *et al*., Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009). doi: 10.1101/gr.092759.109; pmid: 19541911
106. F. Ramírez *et al*., deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016). doi: 10.1093/nar/gkw257; pmid: 27079975
107. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002). doi: 10.1093/nar/gkf436; pmid: 12136088
108. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019). doi: 10.1093/bioinformatics/btz305; pmid: 31070718
109. M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, 2000).
110. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012). doi: 10.1038/nmeth.2109; pmid: 22847109
111. A. Frattini *et al*., High variability of genomic instability and gene expression profiling in different HeLa clones. *Sci. Rep.* **5**, 15377 (2015). doi: 10.1038/srep15377; pmid: 26483214
112. M. L. Whitfield *et al*., Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002). doi: 10.1091/mbc.02-02-0030; pmid: 12058064
113. Y. H. Hung *et al*., Chromatin regulatory dynamics of early human small intestinal development using a directed differentiation model. *Nucleic Acids Res.* **49**, 726–744 (2021). doi: 10.1093/nar/gkaa1204; pmid: 33406262
114. A. M. M. Cartney *et al*., Chasing perfection: validation and polishing strategies for telomere-to-telomere genome

assemblies. bioRxiv 2021.07.02.450803 [Preprint] (2021). https://doi.org/10.1101/2021.07.02.450803.
115. V. Brázda *et al*., G4Hunter web application: A web server for G-quadruplex prediction. *Bioinformatics* **35**, 3493–3495 (2019). doi: 10.1093/bioinformatics/btz087; pmid: 30721922
116. D. Gordon *et al*., Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016). doi: 10.1126/science.aae0344; pmid: 27034376
117. Y. He *et al*., Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat. Commun.* **10**, 4233 (2019). doi: 10.1038/s41467-019-12174-w; pmid: 31530812
118. L. Wang *et al*., A high-quality genome assembly for the endangered golden snub-nosed monkey (*Rhinopithecus roxellana*). *Gigascience* **8**, giz098 (2019). doi: 10.1093/gigascience/giz098; pmid: 31437279
119. V. Jayakumar *et al*., An improved de novo genome assembly of the common marmoset genome yields improved contiguity and increased mapping rates of sequence data. *BMC Genomics* **21** (suppl. 3), 243 (2020). doi: 10.1186/s12864-020-6657-2; pmid: 32241258
120. P. A. Larsen *et al*., Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (*Microcebus murinus*). *BMC Biol.* **15**, 110 (2017). doi: 10.1186/s12915-017-0439-6; pmid: 29145861
121. R. Bandyopadhyay, C. McQuillan, S. L. Page, K. H. Choo, L. G. Shaffer, Identification and characterization of satellite III subfamilies to the acrocentric chromosomes. *Chromosome Res.* **9**, 223–233 (2001). doi: 10.1023/A:1016648404388; pmid: 11330397
122. S. J. Hoyt *et al*., From telomere to telomere: the transcriptional and epigenetic state of human repeat elements analysis code: T2T-CHM13, Zenodo (2022); https://zenodo.org/record/5537106.

ACKNOWLEDGMENTS

We thank the UConn Computational Biology Core for computational support. We thank the NIH Intramural Sequencing Center and the UConn Center for Genome Innovation for sequencing. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov),the UConn HPC Xanadu Cluster (https://bioinformatics.uconn.edu; https://health.uconn.edu/high-performance-computing), the Stanford Research Computing Center, the Stanford Sherlock HPC cluster, the PALMA-II HPC cluster of the University of Münster (https://confluence.uni-muenster.de/display/HPC), and the University of Montana's Griz Shared Computing Cluster (GSCC). Special thanks to M. Diekhans for assistance with UCSC browser tracks and liftOver chain development. **Funding:** This work was supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (AMP). Grants from the US National Institutes of Health: R01GM123312-02 and R21CA240199 to S.J.H., G.A.H., P.G.S.G., and R.J.O.; R01HG002939 and U24HG010136 to J.M.S., J.R., A.F.A.S.; R01HG009190 to A.G. and W.T.; R01HG00990905 to C.L. and A.F.S.; R35GM128857 to L.W. and L.J.C.; R01GM132600, P20GM103546, and U24HG010136 to D.O. and T.J.W.; U24HG010263, U24HG006620, U01CA253481, and R24DK106766-01A1 to M.C.S.; R01HG002385, R01HG010169, and U01HG010971 to M.R.V. and E.E.E.; 1ZIAHG200398R01 to A.R. and A.M.P.; 1R01HG011274-01, R21HG010548-01, and U01HG010971 to K.H.M. Grants from the National Science Foundation: NSF 1613806 and NSF 1643825 to R.J.O.; NSF DBI-1627442, NSF IOS-1732253, and NSF IOS-1758800 to M.C.S. Connecticut Innovations grant 20190200 to R.J.O. Mark Foundation for Cancer Research grant 19-033-ASP to M.C.S. Stowers Institute for Medical Research grant to L.G.d.L. and J.L.G. Howard Hughes Medical Institute Hanna H. Gray Fellowship to N.A. E.E.E. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** Repeat annotation pipeline implementation: S.J.H., J.M.S., J.R., T.J.W., A.F.A.S., and R.J.O. Repeat manual curation: S.J.H., J.M.S., G.A.H., P.G.S.G., J.R., A.F.A.S., D.O., T.J.W., and R.J.O. Dfam database update: J.M.S., J.R., and A.F.A.S. ULTRA data analyses: D.O. and T.J.W. Transduction analyses: R.H., M.R., and W.M. PRO-seq analyses: S.J.H., C.L., L.W., A.R., A.F.S., L.J.C., and R.J.O. PRO-seq data production: S.J.H. (CHM13 and RPE-1) and L.W. (HeLa) CASK pipeline: C.L. and A.F.S. WaluSat analyses: L.G.d.L., J.L.G., P.G.S.G., S.J.H., G.A.H., and R.J.O. Methylation analyses: A.G. and W.T. Centromere Sat Team data integration: N.A. and K.H.M. Segmental duplication and dotplot analyses: M.R.V. and E.E.E. Assembly team lead: A.M.P. Centromere annotation team lead: K.H.M. Variant team lead: M.C.S. Methylation analysis team lead: W.T. Repeat analysis team lead: R.J.O. Segmental duplication team lead: E.E.E. Data visualization: R.J.O., G.A.H., S.J.H., P.G.S.G., and A.G. Figures: R.J.O. Project administration: S.J.H. and R.J.O. Supervision: L.J.C., J.L.G., N.A., W.M., A.F.A.S., A.F.S., T.J.W., M.C.S., E.E.E., A.M.P., W.T., K.H.M., and R.J.O. Manuscript draft: R.J.O. and S.J.H. Supplement draft: S.J.H., G.A.H., P.G.S.G., J.M.S., and R.J.O. Editing: S.J.H., R.J.O., J.M.S., T.J.W., A.F.S., W.M., J.L.G., G.A.H., and P.G.S.G., with the assistance of all authors. **Competing interests:** K.H.M. has received travel funds to speak at symposia organized by Oxford Nanopore. K.H.M. is a scientific advisory board (SAB) member of Centaura, Inc. E.E.E. is a SAB member of Variant Bio, Inc. W.T. has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore Technologies. All other authors declare that they have no competing interests. **Data and materials availability:** Sequencing data and assemblies, PRO-seq CHM13/RPE-1, RNA-seq CHM13 (NCBI BioProject PRJNA559484; www.ncbi.nlm.nih.gov/bioproject/559484); sequencing data, assemblies, and other supporting data on Amazon Web Services (*11*); repeat library for previously unknown repeat entries, UCSC assembly hub browser, RepeatMasterv2 Track CHM13v1.1, RepeatMaskerv2 Track GRCh38 + chrY, RepeatMaskerv2 Track HG002 chrX, RepeatMasterv2 Track Composites CHM13v1.1, RepeatMasterv2 Track previously unknown satellites and arrays CHM13v1.1, sequence alignment and nexus files for phylogenetic analyses, and all scripts and codes used herein (*122*); CHM13v1.1 Meryl 21-mers and 51-mers (*114*); PRO-seq HeLa (GSE179576); updated human repeat database (www.dfam.org). CHM13hTERT cells were obtained for research use via a material transfer agreement with U. Surti and the University of Pittsburgh.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abk3112
Materials and Methods
Figs. S1 to S53
Tables S1 to S28
References (*123*–*155*)
MDAR Reproducibility Checklist

8 July 2021; accepted 8 February 2022
10.1126/science.abk3112

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

Epigenetic patterns in a complete human genome

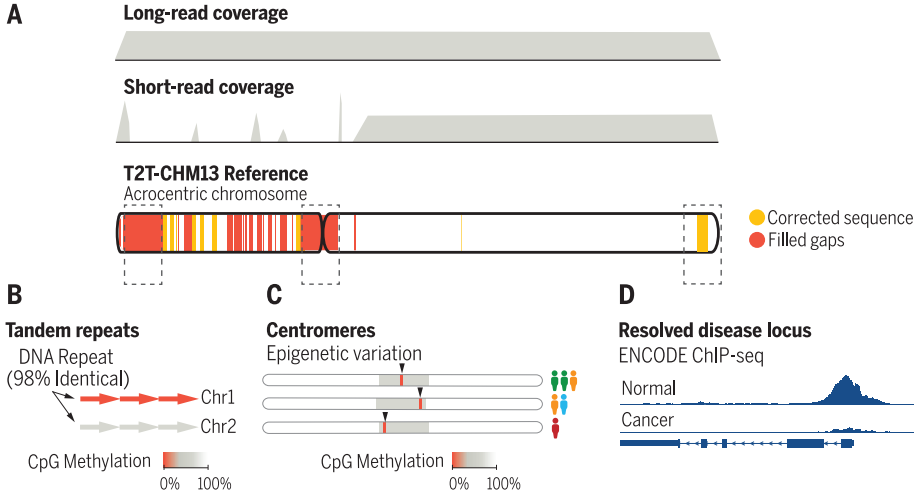
Ariel Gershman, Michael E. G. Sauria, Xavi Guitart, Mitchell R. Vollger, Paul W. Hook, Savannah J. Hoyt, Miten Jain, Alaina Shumate, Roham Razaghi, Sergey Koren, Nicolas Altemose, Gina V. Caldas, Glennis A. Logsdon, Arang Rhie, Evan E. Eichler, Michael C. Schatz, Rachel J. O'Neill, Adam M. Phillippy, Karen H. Miga*, Winston Timp*

INTRODUCTION: The human reference genome has served as the foundation for many large-scale initiatives, including the collective effort to catalog the epigenome, the set of marks and protein interactions that act to control gene activity and cellular function. However, for more than two decades, efforts to construct a complete epigenome have been hampered by an incomplete reference genome. With recent technological advances, we can now study genome structure and function comprehensively across a complete telomere-to-telomere human genome assembly, T2T-CHM13. As a result, we can now broaden the human epigenome to include 225 million base pairs (Mbp) of additional sequence.

RATIONALE: The epigenome refers to DNA modifications (e.g., CpG methylation), protein-DNA interactions, histone modifications, and chromatin organization that collectively influence gene expression, genome regulation, and genome stability. These epigenetic features are heritable upon cell division but dynamic during development, generating profiles that are unique to different tissues and cell types. Here, we present an epigenetic annotation of the human genome in which we explore previously unresolved regions, including acrocentric chromosome short arms, segmentally duplicated genes, and a diverse collection of repeat classes, including human centromeres. Generating a complete epigenetic annotation of the previously missing 8% of the human genome provides a foundation for elucidating the functional roles of these genomic elements that are critical to our understanding of genome regulation, function, and evolution.

RESULTS: Completion of the human epigenome required that we develop approaches to profiling the previously unresolved regions. Using the T2T-CHM13 reference with existing short-read epigenetic data, we identified 3 to 19% more enrichment sites for epigenetic markers. However, even with the complete reference, these short-read epigenetic methods cannot correctly resolve regions of the genome of high similarity, including segmental duplications, gene paralogs, or large repeat arrays. On the other hand, long-read epigenetic methods can resolve single-molecule epigenetic patterns within these regions by anchoring to flanking or infrequent unique regions, providing a foundational assessment of these areas. Long-read methylation calls using the T2T-CHM13 assembly increased the number of probeable CpG sites by 10% (3.2 M), revealing epigenetic patterning of genomic regions that were previously intractable. We generated long-read methylomes of distinct developmental time points and surveyed >99% of the genome's CpGs. We probed highly homologous gene families and observed paralog-specific differences in regulation between disease and nondisease states. In tandem repeats, we identified differences in epigenetic regulation between genetically identical sequences present across different genomic locations, observing locus- and single-molecule-level differences in methylation. Our analysis revealed that these regions vary in epigenetic and transcriptional activity despite high sequence identity, highlighting the importance of the local chromosome environment as a modulator of epigenetics. Finally, the T2T-CHM13 genome assembly has opened exploration of the human centromere, enabling us to probe the epigenetic elements that define centromeric chromatin. The centromere is the site of assembly of the kinetochore complex, an essential complex for eukaryotic cell division. We generated complete epigenetic maps of human centromeres, revealing epigenetic markers of centromere activity that denote active human kinetochores. We predicted kinetochore site localization within active centromeres and report variability of kinetochore localization across individuals representing diverse ancestry.

CONCLUSION: The improvements in epigenetic profiling using T2T-CHM13 set the foundation for complete assemblies and long-read epigenetics for major biological advancements. Using technological advances in genome resequencing and alignment, we present a comprehensive functional assessment of previously unresolved genomic regions. This study marks the start of exploration into duplicated and repetitive portions of the epigenome, pioneering the exploration of epigenetics in a complete human genome. ■



Epigenetic characterization across a complete human genome. (**A**) The T2T-CHM13 reference contains filled gaps and corrected sequences. Using short- and long-read sequencing data, we functionally annotated these added regions. (**B**) Tandem repeats, which are nearly identical, vary in epigenetic state depending on genomic location. (**C**) The epigenetic basis of centromere identity is variable among diverse individuals. (**D**) In genes associated with disease, short reads mapped to T2T-CHM13 elucidate epigenetic dysregulation in human disease states.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: khmiga@ucsc.edu (K.H.M.); wtimp@jhu.edu (W.T.)
Cite this article as A. Gershman *et al.*, *Science* **376**, eabj5089 (2022). DOI: 10.1126/science.abj5089

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abj5089

RESEARCH ARTICLE

HUMAN GENOMICS

Epigenetic patterns in a complete human genome

Ariel Gershman[1], Michael E. G. Sauria[2], Xavi Guitart[3], Mitchell R. Vollger[3], Paul W. Hook[4], Savannah J. Hoyt[5,6], Miten Jain[7], Alaina Shumate[4], Roham Razaghi[4], Sergey Koren[8], Nicolas Altemose[9], Gina V. Caldas[10], Glennis A. Logsdon[3], Arang Rhie[8], Evan E. Eichler[3,11], Michael C. Schatz[2], Rachel J. O'Neill[5,6], Adam M. Phillippy[8], Karen H. Miga[7]*, Winston Timp[1,4]*

The completion of a telomere-to-telomere human reference genome, T2T-CHM13, has resolved complex regions of the genome, including repetitive and homologous regions. Here, we present a high-resolution epigenetic study of previously unresolved sequences, representing entire acrocentric chromosome short arms, gene family expansions, and a diverse collection of repeat classes. This resource precisely maps CpG methylation (32.28 million CpGs), DNA accessibility, and short-read datasets (166,058 previously unresolved chromatin immunoprecipitation sequencing peaks) to provide evidence of activity across previously unidentified or corrected genes and reveals clinically relevant paralog-specific regulation. Probing CpG methylation across human centromeres from six diverse individuals generated an estimate of variability in kinetochore localization. This analysis provides a framework with which to investigate the most elusive regions of the human genome, granting insights into epigenetic regulation.

The human reference genome has served as the foundation for many large-scale epigenetic initiatives (*1*–*3*) that aimed to catalog regulatory elements involved in gene activity and cellular function. However, efforts to construct a complete annotation of functional elements have been hampered by an incomplete reference genome. With recent technological advances, we are now able to study genome structure and function comprehensively across the finished, telomere-to-telomere human genome assembly, T2T-CHM13, which is based on the CHM13 cell line derived from a complete hydatidiform mole (*4*). As a result, we can now broaden the human epigenome to include 225 million base pairs (Mbp) of sequence, representing entire acrocentric chromosome short arms, gene family expansions, and a diverse collection of repeat classes.

The epigenome is influenced both by the specific genetic sequence and the sequence context, i.e., the flanking regions and placement of the loci within the complex structure and organization within the nucleus (*5*). The same genetic sequence can perform different functions or be regulated differently depending on the location of the sequence and its epigenetic state. This is especially relevant given possible evolutionary advantages that may be conferred by gene duplication, such as selectively silencing or activating different paralogous gene copies. These processes are hypothesized to diversify gene activity across developmental time and different tissues (*6*). Beyond evolutionary questions, epigenetic dysregulation of repetitive sequences can play a key role in development and human disease. A diverse set of repeat sequences that are difficult to probe in the human reference genome GRCh38 have been implicated in facioscapulohumeral muscular dystrophy (FSHD) (associated with deletions in *D4Z4*) (*7*); schizophrenia (associated with an expanded repeat in *TAF1I*) (*8*); neuroblastoma (associated with somatic hypomethylation of *SST1*) (*9*); lung cancer (associated with *CT47* expression) (*10*); pancreatic ductal adenocarcinomas (associated with HSat2 expression) (*11*); and immunodeficiency, centromeric region instability, and facial anomalies syndrome (ICF) (associated with heterochromatin abnormalities in HSat2,3) (*12*).

Within the improved T2T-CHM13 reference, the previously unresolved areas are highly repetitive, containing only infrequent sites of unique, mappable regions. This presents a limitation to short-read sequence mapping strategies, even with a more accurate reference and unique k-mer anchored alignments (*13*, *14*). Emerging long-read technologies (*15*) offer sequence lengths capable of spanning infrequent unique markers and provide a direct measurement of the base sequence and epigenetic state on single molecules (*16*, *17*).

[1]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Biology and Computer Science, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [5]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA. [6]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [7]UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [8]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [9]Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA. [10]Department of Molecular and Cell Biology, University of California Berkeley, Berkeley CA, USA. [11]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.
***Corresponding author. Email: khmiga@ucsc.edu (K.H.M.); wtimp@jhu.edu (W.T.)**

RESULTS

Epigenetic profiles from a T2T genome in disease-relevant loci

The T2T-CHM13 assembly resolves gaps and corrects misassembled or patched regions in GRCh38, leading to the introduction of nearly 225 Mbp (*4*). Using existing short-read epigenetic data from the ENCODE project (*1*), we probed previously unidentified areas of the genome. To ensure accurate mapping to these regions, we intersected ENCODE chromatin immunoprecipitation–sequencing (ChIP-seq) alignments with unique k-mers of varying size of k (range, k = 50 to 100; fig. S1 and tables S1 and S2) (*1*, *18*). On average, 2.35% more reads mapped to T2T-CHM13 than GRCh38 across six different histone marks and CTCF, an important regulator of chromatin architecture (fig. S2). Reads filtered out of GRCh38 due to non-unique mapping were largely confined to the satellite DNA and segmental duplications (SDs) (fig. S3). Although the total number of peaks called per sample was variable because of differences in cell type, all samples had an increase in the number of peaks called when comparing T2T-CHM13 with GRCh38 (Fig. 1A). As expected, we saw the most substantial increase in H3K9me3 (19.4%) and H3K27me3 (15.2%) enrichment compared with GRCh38 (Table 1), consistent with the introduced peri/centromeric satellites (CenSat), SDs, and other repetitive sequences in T2T-CHM13 (Fig. 1A) that are associated with constitutive heterochromatin (*19*). The number of called peaks in activating marks increased as well; most notably, there was a 4.9% increase in H3K36me3, a mark present across active gene bodies. Previously unresolved activating histone peaks (H3K27ac, H3K4me1, H3K36me3, and H3K4me3) and CTCF were primarily enriched in unique genic regions and in SDs (Fig. 1A).

T2T-CHM13 increased the number of annotated genes by 5.7% (*4*), revealing 2680 genes exclusive to T2T-CHM13 with no assigned ortholog in GRCh38 (*18*). These gene predictions require detailed study for functionality and validation. Here, we generated a functional annotation of the previously unresolved genes using activating peaks (H3K4me3 or H3K27ac) from ENCODE cell lines. We annotated activating peaks from at least two ENCODE cell lines at the transcriptional start site at 57 of these previously unresolved genes (table S3). Of these loci, most (n = 20) were long noncoding RNAs (lncRNAs), including *LINC01666*, which is known for its associations with gastric cancer (*20*). Many (n = 19) were pseudogenes, including FSHD region gene 1 (*FRG1*), which is a poorly understood candidate gene for FSHD

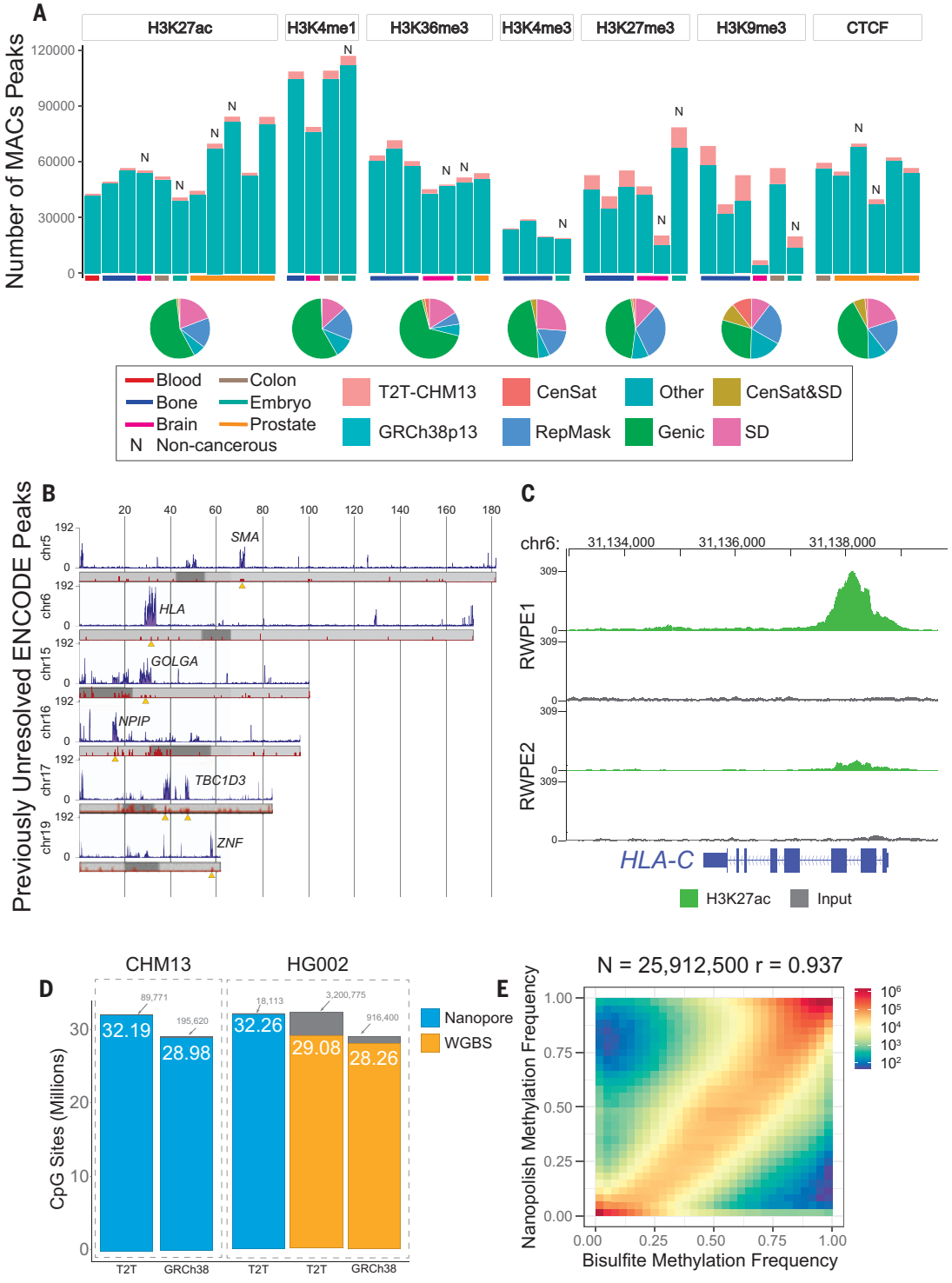


Fig. 1. Epigenetics in previously unresolved genome regions. (**A**) Top: bar plots of the number of peaks called per ENCODE sample using dynamic k-mer mapping to GRCh38 (blue) or T2T-CHM13 (salmon). Bottom: pie charts indicating the genomic localization of peaks found only in T2T-CHM13. (**B**) Number of T2T-CHM13 unique ENCODE peaks across chromosomes 5, 6, 15, 16, 17, and 19 in 50-kb bins (purple). Chromosome ideograms show the density of previously unannotated genes (red) with the centromere annotated as dark gray. Orange triangles denote regions of interest with a high density of previously uncalled peaks. (**C**) ENCODE ChIP-seq read coverage at the HLA-C gene locus on chromosome 6. (**D**) Number of CpGs with methylation profiled comparing sequencing method and reference assembly. (**E**) Correlation of HG002 WGBS and Nanopolish methylation calls aligned to T2T-CHM13.

(*21*). Three were protein-coding genes, including *BOLBA2B*, one of the most common genes associated with autism (*22*).

Our analysis of previously unresolved ENCODE peaks revealed enrichment of peaks for high-copy-number gene families (e.g., *GOLGA*, *NPIP*, *ZNF*, and *TBC1D3*) (Fig. 1B). Large structural variants resolved in T2T-CHM13 explain the additional ChIP-seq mapping events (fig. S4). Epigenetic annotation at these genetic loci may lead to insights of paralog-specific function in evolution (e.g., human-specific neural genes) and disease (*23*, *24*). For instance, *SMN1/2* is associated with spinal muscular atrophy (SMA) and was historically one of the most difficult regions to assemble (*25*). At the *SMN2* gene, we observed peaks of the activating H3K4me3 mark at the promoter in all four ENCODE cell lines analyzed (fig. S5), indicating high transcriptional activity of the gene across tissues. SMA is a leading cause of childhood death (*26*) and has the potential to be treated by regulating expression through histone deacetylase inhibitors, but understanding the disease-specific epigenetic differences between paralogs has been challenging (*27*).

Table 1. Peaks called using ENCODE datasets. Shown is a summary of ENCODE peak analysis including the mark profiled, summed peak calls per mark across all datasets, difference in peak number between references, and number of datasets.

Mark	Peaks called in GRCh38	Peaks called in CHM13	Difference in no. of peaks	Increase in peaks	No. of datasets
H3K9me3	194,681	241,497	46,816	19.4%	6
H3K27me3	249,945	294,819	44,874	15.2%	6
H3K36me3	373,933	393,224	19,291	4.9%	7
CTCF	327,713	342,284	14,571	4.3%	6
H3K4me1	396,332	412,907	16,575	4.0%	4
H3K27ac	611,645	632,837	21,192	3.3%	11
H3K4me3	88,985	91,724	2739	3.0%	4

Another previously intractable region of the genome, the HLA locus, is critical for understanding a wide range of biology from immunity to neuropsychiatric disorders (*28*, *29*). Our results reveal enrichment of ENCODE peaks across a variety of histone marks at the HLA locus (Fig. 1B and fig. S6A). Decreasing expression of HLA genes is associated with soft tissue cancers, particularly prostate cancer, and can even be indicative of chemotherapy resistance (*30*). Comparing non-neoplastic adult human prostate epithelial cells (RWPE-1) and the c-Ki-ras–transformed prostate cancer model cells from the same donor (RWPE-2) (*31*), we observed a decline in H3K27ac, an activating mark, at HLA gene promoters, concomitant with an increase in CTCF binding in RWPE-2 (Fig. 1C and fig. S6B). The differences in histone marks in this region indicate epigenetic dysregulation of the HLA locus in prostate cancer that may warrant further studies and inform upon potential therapies (*32*).

Long-read sequencing to derive complete human methylomes

Methylation profiling has traditionally faced difficulties in mapping success rates to repetitive regions of the genome, and such mapping inefficiencies are exaggerated by the bisulfite conversion of unmethylated cytosine to uracil, sequenced as thymine (*33*). Methylation profiles in T2T-CHM13 using long-read nanopore data demonstrate an increase in the genome coverage (32.8 M compared with 29.17 M in GRCh38, omitting chromosome Y) and surveyed more CpGs (10%, 3.18 M) compared with short-read whole-genome bisulfite sequencing (WGBS) (Fig. 1D). We called nanopore methylation data with Nanopolish (*34*), finding a high correlation (R = 0.937) both to WGBS results in regions mappable by both data types (Fig. 1E) and to the alternative nanopore methylation caller, Megalodon (R = 0.952) (fig. S7). Examining the difference between mapping of WGBS and nanopore methylation data, we generated short-read mappability scores in 200-bp windows, with a score of 0 being unmappable and 200 being highly mappable (*18*). We found that the 165 Mbp of sequence with a score of 0 (highly unmappable) is enriched in SDs and satellite DNA. Stratifying the nanopore data by read length, we found reads longer than 50 kilo–base pairs (kb) were capable of accurately determining methylation in these regions (figs. S8 and S9).

We sequenced the CHM13 cell line, representing an early developmental state, and HG002, a terminally differentiated lymphoblast cell line. The sequenced cell line CHM13 and HG002 nanopore datasets surveyed 32.19 M CpGs (99.7% of total) and 32.26 M CpGs (99.9% of total). As expected for differentiated cell lines, most of the HG002 genome is methylated (75% median methylation), with a secondary peak of unmethylated CpGs largely reflecting unmethylated CpG islands (CGIs) (fig. S10). By contrast, CHM13 is markedly hypomethylated (36.8% median methylation), as expected from a trophoblastic cell line (*35*). Comparing CHM13's methylation state with existing DNA reduced representation bisulfite-sequencing data on early human embryos (fig. S11 and table S4) (*35*), we observed that CHM13 clusters closely with cleavage and blastocyst-stage embryos as well as trophectoderm tissue.

To probe chromatin state in repetitive DNA, we generated long-read nanoNOMe data on HG002 using M.CviPI methyltransferase to decorate accessible chromatin with exogenous GpC methylation (*16*) and called CpG and GpC methylation with Nanopolish to measure chromatin accessibility (figs. S12 and S13). With the combination of long-read epigenetic data and the complete human reference, we now describe a complete human epigenome, providing a foundation for further study.

Paralog-specific epigenetic regulation

The neuroblastoma breakpoint family (*NBPF*) family of genes has been implicated in the expansion of the human prefrontal cortex since our lineage diverged from apes (*36*). One of its copies, *NBPF1*, has been reported to act as a tumor suppressor in neuroblastoma, in which hypomethylation of CGIs has been associated with astrocytoma formation (*37*). Understanding the regulation of this gene family, however, has been particularly challenging because the *NBPF* genes correspond to large, high-identity duplications (>98%) that are copy number polymorphic among humans and map to gaps in the existing reference sequences (*38*). The fully resolved nature of T2T-CHM13 allowed us to remap ENCODE data to discover regulatory elements associated with this gene family (Fig. 2A). When comparing the balance between H3K36me3, a mark of active exons and gene bodies, and H3K27me3, a repressive mark, in samples including the BE2C cell line (neuroblastoma) and primary brain microvascular tissue (normal brain), we found that BE2C shows a higher proportion of H3K27me3 peaks (BE2C 38, brain 8) and a lower proportion of H3K36me3 peaks (BE2C 36, brain 89) at *NPBF* loci (fig. S14 and table S5). Taking advantage of the increased resolution and more accurate *NBPF* copy number provided by T2T-CHM13 (*39*), we assayed paralog-specific epigenetic changes occurring in neuroblastoma (Fig. 2B). Among the different *NBPF* gene copies, the largest shifts in epigenetic regulation occurred at *NBPF26* and *NBPF10*, moving from active marks in primary brain microvascular tissue to repressive marks in BE2C. These specific *NBPF* copies are noteworthy because they associate with the human-specific duplicate genes *NOTCH2NLA* and *NOTCH2NLR*, determinants of the size and complexity of the human neocortex (*40*). This association identifies the functional *NBPF* copies, emphasizing the importance of studying paralog-specific epigenetics for the discovery of potential drug targets.

Regulatory regions are excluded because of low short-read mappability scores among high identity paralogs as in the *NBPF* gene family (Fig. 2C and fig. S15) (*18*). We found that genome-wide methylation, H3K4me2 (a mark of active promoters), and H3K27me3 (a repressive mark) correlate with Iso-Seq coverage (transcription) and together can be used to systematically evaluate the functional activity of this gene family (fig. S16). We correlated this activity with the evolutionary age of the paralogs, estimated using *NBPF* gene paralogs

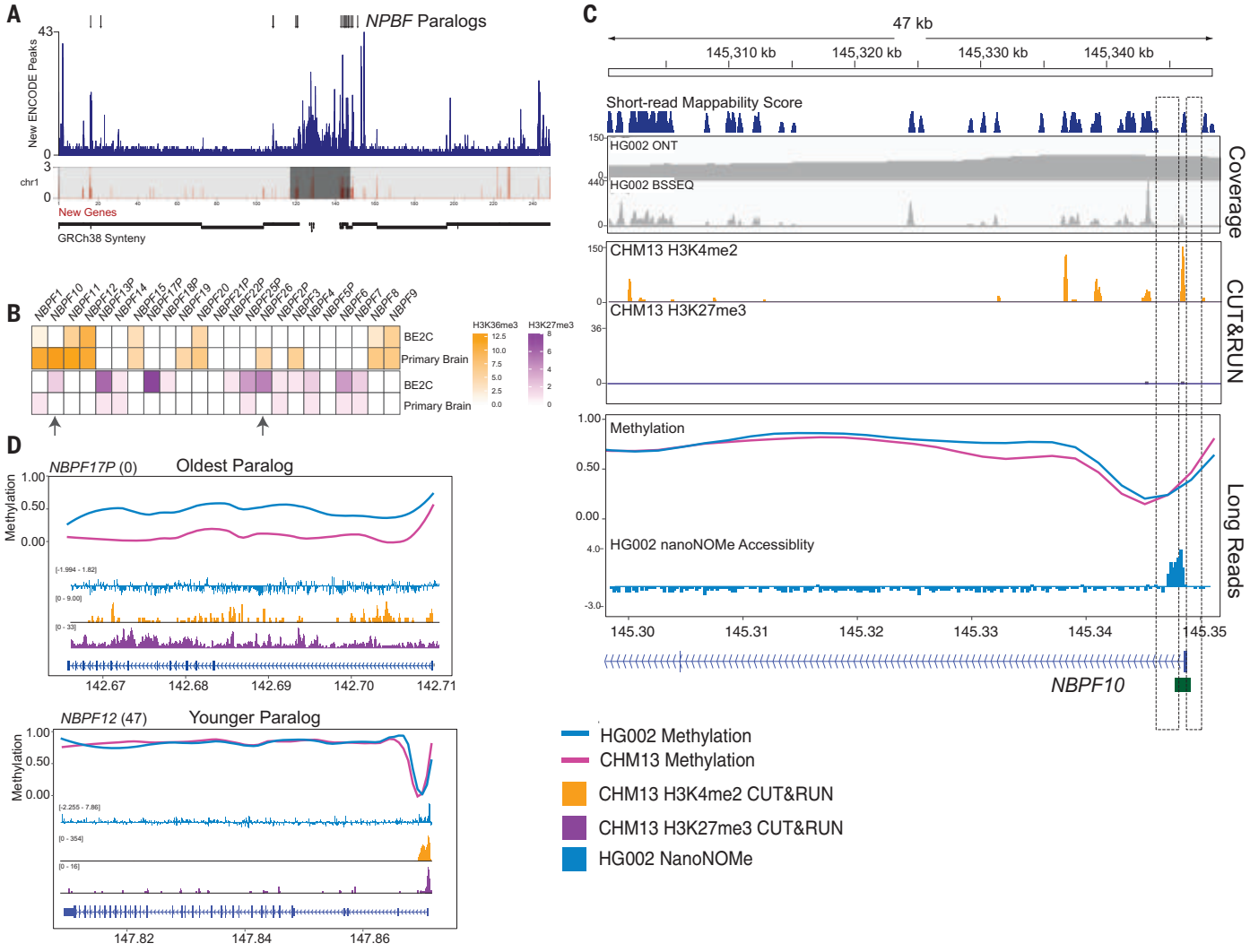


Fig. 2. Paralog-specific epigenetic regulation of the NBPF gene family. (**A**) Location of T2T-CHM13 previously uncalled ENCODE peaks across chromosome 1 in 50-kb bins (purple). Chromosome ideograms contain the density of previously unannotated genes (red) and centromere annotations (dark gray). NBPF paralogs are indicated by black arrows (top). (**B**) Heatmap illustrating the number of peaks for H3K36me3 (orange) and H3K27me3 (purple) per NBPF paralog in the ENCODE cell line BE2C (neuroblastoma) and brain tissue (primary brain microvascular tissue). Arrows indicate NBPF10 and NBPF26. (**C**) Epigenetic data at the NBPF10 promoter and first intron (chromosome 1: 145,300,425 to 145,348,763). Short-read mappability score from 0 to 200 calculated as a 200-bp region, with a score of 200 being the most mappable and 0 being the least mappable. Coverage tracks (Illumina WGBS and ONT) and CUT&RUN tracks display read pileups. Long-read methylation tracks show base-level methylation frequency, with 0 as unmethylated and 1 as fully methylated. The long-read HG002 accessibility track is a 200-bp binned *Z*-score of nanoNOMe GpC methylation frequency. Dashed boxes highlight the promoter region that is largely unmappable with short-reads. (**D**) Bottom: younger NBPF12 gene paralog displaying CHM13 and HG002 nanopore methylation, CHM13 H3K4me2 and H3K27me3 CUT&RUN coverage, and HG002 nanoNOMe. Top: older NBPF17P gene paralog displaying CHM13 and HG002 nanopore methylation, CHM13 H3K4me2 and H3K27me3 CUT&RUN, and HG002 nanoNOMe. Numbers in parentheses refer to the number of PacBio Iso-seq transcripts mapped to this paralog.

from six nonhuman primates from local genome assembly of the *NBPF* gene family from each primate (*39*) (fig. S17). The oldest paralog, *NBPF17P*, has low Iso-Seq coverage correlated with an epigenetic signature consistent with a repressive state, including promoter hypermethylation and inaccessibility, enrichment of H3K27me3, and decline of H3K4me2 (Fig. 2D). By contrast, the younger paralogs, including human-specific copies, have higher Iso-Seq coverage and epigenetic signatures consistent with active genes, including hypomethylated and accessible promoters and enrichment of H3K4me2. Activity in the younger paralogs is more variable, with *NBPF10* and *NBPF20* displaying high functional activity and sharing promoters with *NOTCH2NLA* and *NOTCH2NLB*. Taken together, our results illustrate the role of epigenetics in the regulation of gene paralogs, silencing evolutionarily older paralogs while activating newer copies. This provides mechanistic insight into potentially functional genes related to human-specific cortical expansion and dysregulation in neoplasia.

Array-specific epigenetic regulation of tandem repeats

Using k-mer–directed ENCODE alignments to the T2T-CHM13 reference, we report epigenetic features from human centromeric regions, subtelomeres, and acrocentric short arms, which represent previously unresolved regions of the genome that are dominated by CenSat DNAs (fig. S18). Five different ENCODE lines had an enrichment of H3K9me3 in CenSat DNA, notably observed in short-read mappable regions of the acrocentric short arms (fig. S19).

SJCRH30 (a rhabdomyosarcoma-derived line) had lower H3K9me3 enrichment in CenSat compared with the rest of the chromosome, suggesting satellite epigenetic dysregulation as a clinically relevant pathology in rhabdomyosarcoma (figs. S19 and S20A and B). This trend can be observed with more detail in an HSat3 repeat on the acrocentric arm of chromosome 15, where H3K9me3 in SJCRH30 is clearly depleted compared with HAP-1 (fig. S20C).

In contrast to these heterochromatic marks, we found enrichment of activating marks, including H3K27ac, H3K4me3, and CTCF in the telomere-associated repeat (TAR) region, typically located 2 kb upstream from the canonical telomeric repeat. A CTCF site in the TAR loci drives transcription of the TERRA lncRNA (*41*), a negative regulator of telomerase-mediated telomere elongation. We observed enrichment of CTCF in all ENCODE cell lines at the TAR loci (fig. S21A), but the subtelomeric regions were rich in SDs, resulting in the TAR sequence being dispersed throughout the genome (*42*). When comparing telomeric TAR sequences with nontelomeric TAR sequences, we did not observe statistically significant differences (Kruskal-Wallis, $P = 0.12$) in sequence divergence (fig. S21B). Although both telomeric and nontelomeric TAR sequences are enriched for CTCF, the nontelomeric TAR sequences are more enriched for the activating chromatin marks H3K27ac and H3K4me3, suggesting differences in TERRA activity.

Examining nanopore CpG methylation in tandemly repeated satellite DNA elements in CHM13 and HG002 revealed hypomethylation in CHM13 compared with HG002 (Fig. 3A) (*43*). To assess the chromatin profile of satellite repeats, we called accessibility peaks from the HG002 nanoNOMe data (*18*). We found that, when corrected for the size of the region, repeats have lower peak density than the genome as a whole. The number of nanoNOMe peaks per megabase of sequence was lower in satellite DNA (1.5), LINEs (8), SINEs (15), and LTRs (13.4) compared with the whole genome (31.8) (Fig. 3B and table S6) (*44*). The human satellites (HSat 2,3) and monomeric alpha satellites (MON) were largely devoid of accessibility peaks. Repetitive DNA is typically associated with densely packed heterochromatin (*45*), and our findings are consistent with this association and transcriptional profiles from (*44*). However, our data allow us to investigate accessibility profiles within previously unmappable satellite repeats.

Contrary to the expectation of compact chromatin and satellite DNA, we discovered enrichment of accessibility peaks in the SST1 satellite both inside the CenSat (41.4 peaks/Mbp) and in the chromosome arms (198.1 peaks/Mbp). Our peak annotations in HG002 were consistent with (*44*), which showed higher activity in CHM13 at noncentromeric arrays on chromosomes 4 and 19 compared with other SST1 arrays (table S7). After the SST1, the satellite repeat with the second highest peak enrichment was the ACRO_Composite, a 7-kb repeat found across 12 chromosomes, including as tandemly arrayed sequences across the five acrocentrics with high sequence identity across composite units (*44*). The tandemly arrayed promoter elements in the ACRO_composite give rise to a periodic bimodal methylation structure across the array (Fig. 3C). This epigenetic pattern has been proposed to be important for both the efficient transcription of noncoding RNAs and the maintenance of the nearly perfect tandem arrays (*46*). The array has regions of increased CpG methylation that were associated with nanoNOMe peaks and transcription (CHM13 PRO-seq) (Fig. 3C). We quantified nanoNOMe peak densities across the ACRO_Composite between chromosomes and found that chromosome 21 has the highest (4.5 peaks/100 kb) and chromosomes 13 and 15 have the lowest (0 peaks/100 kb) (Fig. 3D). The absence of nanoNOMe peaks in chromosomes 13 and 15 is correlated with low transcriptional activity (fig. S22). This high-resolution look within the acrocentric repeats indicates chromosome-specific activity of the ACRO_Composite across both CHM13 and HG002, suggesting a persistent functional role for the ACRO noncoding RNA throughout early- and late-stage development.

By contrast, we also observed methylation periodicity in untranscribed satellite repeats such as the HSat2, and these regions were largely inaccessible as measured by nanoNOMe (Fig. 3E) (*44*). This periodicity in methylation corresponds to the underlying chromatin structure and echoes the genetic repeat size, suggesting the presence of functional genomic elements. Our initial epigenetic assessments of these assembled satellite sequences indicate a complicated regulatory structure stretching beyond the accepted notion that the repetitive fraction of mammalian genomes is entirely methylated and repressed by a highly condensed chromatin state (*47*).

Single-read-level analysis in satellite arrays reveals array heterogeneity

Long reads, coupled to a complete reference assembly, confer the ability to explore methylation patterns of single molecules, each of which represents the methylation pattern of a single allele from a single cell. The X chromosome provides a unique opportunity to study these patterns because of the role of allele-specific methylation in X chromosome inactivation (XCI). Female somatic tissues have a mixture of paternal or maternal X expression because the same X chromosome is not always repressed; therefore, the active X (Xa) and inactive X (Xi) cannot be distinguished with heterozygous single-nucleotide polymorphisms alone. Examining methylation state at CGIs, we clustered reads on the CHM13 X chromosome as hyper- or hypomethylated (*18*). To determine whether the clusters represent the Xa and Xi, we first focused on genes known to be subject to XCI (XCI genes) or known to escape inactivation (escape genes) and compared our results with the clonal female lymphoblast cell line GM12878, in which the Xi is always the paternal allele (fig. S23A and B) (*48*). There, we found the Xa to have hypomethylated promoters and hypermethylated gene bodies compared with the Xi (*49*). However, in CHM13, we discovered that not all genes (e.g., *TAF9B* and *PRKX)* were properly regulated, with *TAF9B* escaping XCI and *PRKX* being subject to XCI, contrary to expectation. This is likely due to failure of X chromosome inactivation in androgenetic CHMs (fig. S23C and D and table S8) (*50*).

Moving this analysis into repetitive regions, we analyzed DXZ4, a satellite that acts as a major epigenetic regulator of XCI (*51*). This 165-kb macrosatellite repeat contains 3-kb monomeric units, each with a bidirectional CGI promoter and a CTCF site that is hypomethylated on the Xi and hypermethylated on the Xa in healthy cells (*52*, *53*). Single-read clustering revealed two distinct clusters of reads, one with higher methylation across the repeat and the other with lower methylation (Fig. 3F). This analysis revealed an unexpected level of heterogeneity in methylation of monomers within the array. We hypothesize that this variation is a result of the aberrant XCI state of CHM13, because intra-array variation was not observed in the Xa at DXZ4 in HG002 (fig. S24). Observing epigenetic differences between monomers of satellite repeats could grant insights into human disease, providing a detailed mechanistic understanding of satellite dysregulation. From this analysis, we demonstrate that we can cluster reads using methylation alone to identify heterogeneous populations and intra-array epigenetic variation even in the absence of heterozygous genetic variants.

Methylation maps of human centromeres reveal complex epigenetic patterns

Human centromeres are composed of alpha satellite DNA, with an AT-rich, ~171-bp repeat unit or "monomer." The largest arrays of alpha satellites in the human genome are further organized in chromosome-specific, higher-order repeats (HORs) or larger, multimonomeric repeat units (*54*). Centromeres can contain multiple distinct alpha satellite HOR arrays that can be classified into active and inactive HORs (*55*, *56*). The HORs within active arrays have specialized epigenetic regulation that are important in establishing and maintaining centromere identity (*56*, *57*).

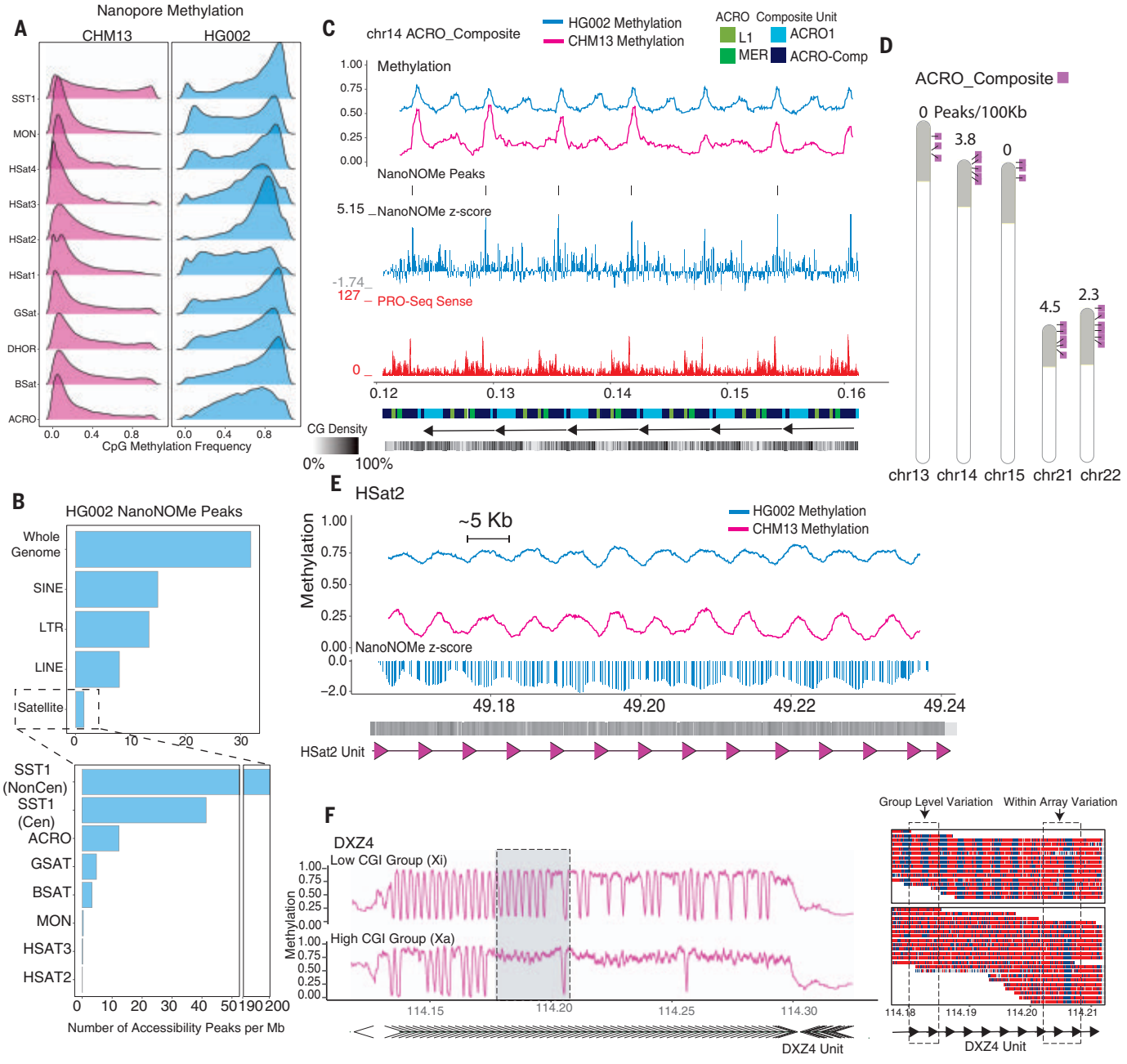


Fig. 3. Context-specific epigenetics in high-identity tandem repeats. (**A**) Nanopore methylation frequency of satellite repeat classes in CHM13 and HG002. (**B**) HG002 NanoNOMe statistically significant peak calls per 1 Mb of sequence in all major repeat classes compared with the whole genome (top) and within different satellite repeats (bottom). (**C**) Nanopore CpG methylation profiles, HG002 NanoNOMe accessibility peaks and *Z*-score (negative is inaccessible, positive is accessible), and non–kmer-filtered (multimapping) PRO-Seq coverage at the ACRO_Composite repeat (chromosome 14: 121,193 to 162,142). Annotation tracks at the bottom are the RepeatMasker V2 annotation from (*44*), monomeric annotations of the ACRO_Composites, and a GC density track. (**D**) Ideogram showing the arrayed locations of the ACRO_Composite across the acrocentric chromosomes (purple) within the acrocentric short arms (gray shaded). Listed above each chromosome is the nanoNOMe ACRO_composite peak density in peaks/100 kb. (**E**) Nanopore CpG methylation profiles and HG002 NanoNOMe accessibility *Z*-score of the HSat2 repeat (chromosome 16: 49,163,529 to 49,239,753). Annotation bars below represent CpG density and HSat2 repeat units on the bottom. (**F**) The DXZ4 locus on CHM13 clustered into two haplotypes (low CGI methylation and high CGI methylation) based solely on promoter methylation state. Left: methylation frequency plot of each haplotype. Right: single reads from the gray highlighted region on the left, with boxes highlighting CGI cluster group-level epigenetic variability and intra-array-level epigenetic variability between neighboring monomeric units.

Centromere protein A (CENP-A) is an H3 variant that is enriched in centromeric nucleosomes and marks sites of kinetochore assembly (*58*). In HOR arrays, notable hypomethylation colocalizing with CENP-A enrichment at chromosomes X and 8 has been described (*13*, *14*). We extended this finding to all CHM13 centromeres, referring to this hypomethylation as the centromeric dip region (CDR) (Fig. 4A and table S9). We found that CDRs were present only in active HORs (fig. S25), and that active HORs were larger in size and had higher mean methylation frequency than inactive HORs, as exemplified by the chromosome 5 centromere (Fig. 4B).

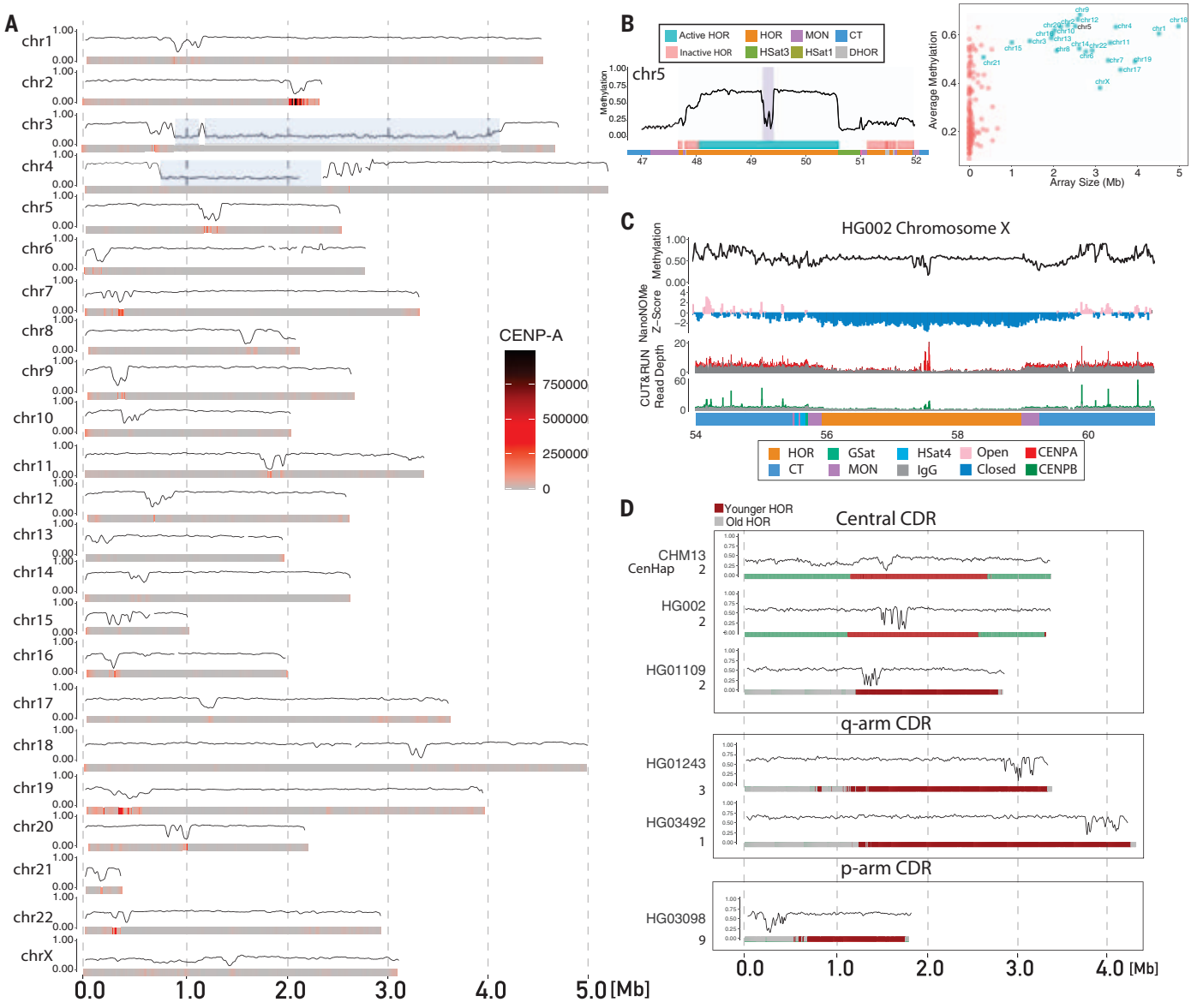


Fig. 4. Epigenetic maps within human centromeres. (**A**) Smoothed methylation frequency in 10-kb bins of the active HOR array for all CHM13 chromosomes. CENP-A enrichment from CUT&RUN data is shown as a heatmap under each plot. Chromosomes 3 and 4 have an HSat1 repeat (blue highlight) that breaks up the live HOR array. (**B**) Left: CHM13 methylation in the centromeric region of chromosome 5. Smoothed methylation frequency is plotted in 10-kb bins. HOR arrays are annotated as blue ("active") and pink ("inactive"). Right: scatter plot of average methylation within each HOR array versus size in megabase pairs. (**C**) Methylation, nanoNOMe accessibility, and CENP-A and CENP-B CUT&RUN data across the chromosome X centromeric array on HG002. Smoothed methylation and accessibility are plotted in 15-kb bins, and CUT&RUN is plotted as raw read counts with input shaded gray. Bottom bar annotates satellite regions indicating the location of the HOR, MON, GSat, HSat4, and CT regions. (**D**) Methylation in the active HOR array across diverse individuals. Coriell cell line sample ID and cenhap group are annotated to left. HORs are annotated as red (younger) and gray (older) computed on the basis of sequence divergence.

These results underscore the importance of methylation in proper centromere regulation and kinetochore assembly.

To investigate whether CDRs were confined only to early developmental samples, we examined HG002 nanopore-sequencing data to probe centromere methylation in an adult differentiated cell line. However, the high level of HOR array variability and the resulting inability to confidently phase and map reads from diploid chromosomes prevented us from using the T2T-CHM13 HOR reference for HG002 reads, as evidenced by the anomalous coverage that we observed for HG002 alignments in the HOR arrays (fig. S26) (*59*). Instead, we took advantage of the haploid nature of the HG002 X chromosome and used an HG002-specific X centromere reference (*4*, *56*), and were able to clearly observe a CDR (Fig. 4C). Furthermore, using nanoNOMe, the CDR was coordinated in this sample with a highly inaccessible region. When we examined the size of the inaccessible regions in the HOR versus the surrounding pericentromeric and centromeric transition (CT) regions, we found that the HORs were enriched in dinucleosomes compared with these other regions (fig. S27). Finally, looking at CUT&RUN CENP-A and centromere protein B (CENP-B) data, we observed a peak of CENP-A and CENP-B binding at the CDR. This is coordinated with a marked hypomethylation of the CENP-B motif within CDRs as opposed to outside the CDRs (fig. S28), and methylation is known to reduce CENP-B binding (*60*). This finding highlights the potential functional

importance of the CDR for kinetochore formation.

Taking this a step further, using Human Pangenome Reference Consortium (HPRC) data, we leveraged the assembled X chromosomes of four additional diverse male samples representing individuals included in the 1000 Genomes Project (Fig. 4D) (*56*, *61*). All arrays showed a distinct CDR in the X chromosome, with positional variability in the CDR location across individuals. Furthermore, CDR position was shared between individuals with more closely related centromere-spanning haplotypes (cenhap) assignments.

Cenhaps are long haplotypes that include centromere arrays due to reduced recombination in CenSat regions (*56*, *62*). Three of the samples, CHM13 (European), HG002 (European), and HG01109 (Puerto Rican), are within cenhap group 2, and all contain a centrally positioned CDR within an evolutionarily "younger" region of the HOR, as defined in (*56*). Two of the samples, HG01243 (Puerto Rican) and HG03492 (Pakistani), are within cenhap groups 3 and 1, which have been shown to be phylogenetically related (i.e., they share a clade with cenhaps 1 to 4) (*56*), and have a CDR positioned more toward the q-arm side of the centromere within the evolutionarily younger region of the HOR array. Finally, one of the samples, HG03098 (African), from the more distantly related cenhap group 9, has a CDR positioned toward the p-arm of the centromere in an older (more diverged) region of the HOR array (supporting the previous observation of an epi-allele in the region using available short-read datasets) (*56*). Therefore, we demonstrate the use of CDRs to identify epigenetic variability within human centromeres, variations that may influence centromere function during cell division. These variations show the critical importance of epigenetic profiling in the centromere, finding variation between individuals in a discrete, epigenetically defined region of the centromere.

DISCUSSION

This work provides a comprehensive view of the epigenetic organization of a complete human genome, uncovering complex epigenetic patterns in the previously unresolved 8% of the human genome. Functional annotation of these intractable regions has not been overlooked because of their lack of importance but rather because of technological limitations. Our study opens these regions to explore their epigenome, leaving no region of the genome unreachable. Here, with the combination of a complete genome assembly and technological advances in epigenetic profiling, we make substantial strides in functional genome assessment, expanding ENCODE (*1*) to include 3 to 19% more peak calls and increasing the number of CpG methylation calls by 10%. Long-read epigenetic methods, here focusing on nanopore methylation and chromatin accessibility, can resolve single-molecule epigenetic patterns within these regions, providing a foundational assessment of these areas. Long-read methylomes of distinctive developmental time points surveyed >99% of CpGs, establishing the CHM13 and HG002 methylomes as the most complete human methylomes to date (*3*). With these datasets, we profiled the additional 225 Mbp of sequence and 2680 gene annotations.

Of the previously unresolved genes, we found 57 with evidence of active promoters, including H3K4me3 or H3K27ac marks, in more than one cell type. We found 82 genes with a single cell type supporting active promoters, providing evidence that these previously unresolved gene annotations are functionally active across tissues. With more data from different tissue types, we may identify even more functional genes. More generally, we found that evolutionarily older gene paralogs were epigenetically repressed (similar to the epigenetic silencing of transposons), conferring genome stability and thus influencing genome evolution (*63*, *64*).

Examining satellite DNA, we integrated short- and long-read datasets to investigate complete satellite arrays, revealing that these regions vary in epigenetic and transcriptional activity despite high sequence identity and highlighting the importance of the local chromosome environment as a modulator of epigenetics. Repetitive DNA on the acrocentric short arms is known to play a role in nucleolar formation; however, the previous absence of these regions from the human reference has hampered research (*65*). Our findings suggest that, rather than acting in unison, the repeat families on these individual acrocentric chromosomes all have their own epigenetic identity, likely contributing to their unique functional roles in genome integrity and organization.

One of the features of our single-molecule epigenetic data is the ability to investigate single-molecule patterns of epigenetics. We used methylation alone to cluster reads in repetitive areas devoid of heterozygous polymorphisms. This includes the DXZ4 array, in which the methylation signature is critical to X chromosome inactivation (*66*, *67*). With the increase in resolution, our results show methylation variability between the clustered populations and intra-array epigenetic variation within adjacent monomers in the same array. Because satellite arrays are known to be hypervariable in the human population and linked to several human diseases, these results highlight the importance of long-read single-molecule epigenetic studies for understanding disease pathology.

Finally, the T2T-CHM13 genome assembly has opened exploration of the human centromere, enabling us to probe the epigenetic elements that define centromeric chromatin. We extended our original discovery of the CDR in chromosome 8 and chromosome X to all chromosomes, and found that CDRs denote the position of centromere-associated proteins (CENP-A and CENP-B in the HG002 genome) in differentiated cells (HG002, a lymphoblast). This provides evidence of CDRs outside of early developmental CHMs and emphasizes their importance in kinetochore positioning and epigenetic regulation of chromosome segregation. Expanding our CDR analysis to male X chromosomes representing diverse haplotypes, we uncovered variability in the localization of the CDR within the X HOR array. Such variability in active centromeric arrays has been explored through the presence of epi-alleles (*68*); however, we have been able to demonstrate the use of CDRs to precisely predict kinetochore site localization within an active array and report across individuals representing diverse ancestry. When combined with findings in other organisms, e.g., maize (*69*) and medaka (*70*), this suggests that the CDR is a conserved, functionally important feature of complex centromeres across vertebrate and plant lineages. Proper kinetochore formation is an essential process for eukaryotic cell division, a process that occurs in humans 330 billion times per day to sustain life. Our results lead to two major conclusions about the CDR: (i) CDR location on a given array is fixed in early development and maintained upon differentiation and (ii) there is a single stable CDR in each centromere. Our initial profile provides a multitude of avenues for future research, including how CDR position influences meiotic and mitotic stability, disease, and aneuploidy.

Our results act as a foundational study, expanding studies of the human genome through the use of the complete reference. There remain significant challenges to further exploring the epigenome in a larger and more diverse sample set to achieve optimal sequence alignment, especially among structurally variable repetitive regions, e.g., HORs. Efforts by the HPRC (*71*) to generate fully phased diploid genome assemblies will enable population-scale exploration of these areas. Limitations of short-read sequencing in unique regions can be supplemented by long-read epigenetic methods currently under rapid development (*16*, *17*). We are on the precipice of exploration into duplicated and repetitive portions of the genome. Further development of long-read epigenetic profiling across different populations and disease states will reveal more about regulation within the genome's most elusive regions.

METHODS SUMMARY

Methylation processing

Nanopore reads were obtained from (*13*, *14*, *72*). Ultra-long nanopore reads were aligned to the CHM13 reference (*4*) with Winnowmap version 2.0 (*73*) with a k-mer size of 15. BAM

files were filtered for primary alignments with SAMtools (version 1.9); analysis of centromeric regions was done on reads >50 kb. To measure CpG methylation in nanopore data, we used Nanopolish (version 0.13.2) with a log-likelihood ratio (LLR) cutoff of −1.5/1.5 (*34*). HG002 bisulfite FASTQs were collected from the Oxford Nanopore Technologies (ONT) open data repository https://labs.epi2me.io/gm24385-5mc. Paired-end FASTQs were aligned with Bismark (version 0.22.2) (*74*). For Nanopolish to Megalodon comparisons, Megalodon was run with the r9.4.1_450bps 5mC model with thresholding set as default.

NanoNOMe

HG002 cells were grown in culture and treated according to methods outlined in (*16*). Purified genomic DNA was prepared for nanopore sequencing following the protocol in the genomic sequencing by ligation kit LSK-SQK109 (ONT). To measure CpG and GpC methylation in nanopore data, we used Nanopolish (version 0.13.2) on the nanonome branch https://github.com/jts/nanopolish/tree/nanonome (*34*). We set an LLR threshold of −1/1 for GpC methylation calls and −1.5/1.5 for CpG methylation calls.

Methylation clustering

Methylation clustering was performed across the CHM13 X chromosome on all CGIs that overlap an annotated promoter of a protein-coding gene. Within the CGI, reads with an average methylation >0.2 were considered methylated, and reads with an average methylation <0.2 were considered unmethylated. Reads were only considered if they spanned the entirety of the CG islands and were longer than 5 kb. Clustered reads were then intersected with known escape and XCI genes from (*51*). The same clustering procedure was performed at the DXZ4 locus.

CUT&RUN

CUT&RUN was performed as detailed in (*75*) with some variations. For library preparation, NEBNext Ultra II End repair/A-tailing and ligation kits were used as indicated by the manufacturer, with 1.5 pg of Spike-in Yeast DNA added (obtained from the Henikoff laboratory at the Fred Hutchinson Cancer Research Center). Marker-assisted mapping of CUT&RUN data (CHM13 CENP-A, CHM13 H3K4me2, CHM13 H3K27me3, HG002 CENP-A, and HG002 CENP-B) to a sample-specific reference [CHM13 to T2T-CHM13 or HG002 to CHM13 autosomes (chromosomes 1 to 22), HG002 T2T (chromosome X), and GRCh38 (chromosome Y)] was performed according to the methods outlined in (*56*).

ENCODE dynamic k-mer–assisted mapping

We selected several ChIP-seq datasets generated as part of the ENCODE project (*1*), choosing those with at least 100-bp paired-end sequencing data and at least one matching input control. These criteria yielded 96 total sequencing libraries (table S9). Reads were mapped with Bowtie2 [version 2.4.1 (*76*)], alignments were filtered using SAMtools [version 1.10 (*77*)], and polymerase chain reaction duplicates were identified and removed with Picard tools [version 2.22.1 (http://broadinstitute.github.io/picard)]. Alignments were then filtered for unique k-mers. Specifically, for each alignment, reference sequences aligned with template ends were compared with a database of k-mers unique in the whole genome. For each end of the paired-end sequencing reads, the k-mer length was determined by finding the largest multiple of 5 less than or equal to the aligned reference sequence length. Peak calls were made using MACS2 (version 2.2.7.1) (*78*) with default parameters and estimated genome sizes of 3.03×10^9 and 2.79×10^9 for chm13v1 and GRCh38p13, respectively.

REFERENCES AND NOTES

1. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
2. J. Dekker *et al*., The 4D nucleome project. *Nature* **549**, 219–226 (2017). doi: 10.1038/nature23884; pmid: 28905911
3. A. Kundaje *et al*., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi: 10.1038/nature14248; pmid: 25693563
4. S. Nurk *et al*., The complete sequence of a human genome. *Science* **376**, 44–53 (2022). doi: 10.1126/science.abj6987
5. D. Jost, C. Vaillant, Epigenomics in 3D: Importance of long-range spreading and specific interactions in epigenomic maintenance. *Nucleic Acids Res.* **46**, 2252–2264 (2018). doi: 10.1093/nar/gky009; pmid: 29365171
6. N. V. Fedoroff, Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758–767 (2012). doi: 10.1126/science.338.6108.758; pmid: 23145453
7. D. Gabellini, M. R. Green, R. Tupler, Inappropriate gene activation in FSHD: A repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell* **110**, 339–348 (2002). doi: 10.1016/S0092-8674(02)00826-7; pmid: 12176321
8. H. A. Bruce *et al*., Long tandem repeats as a form of genomic copy number variation: Structure and length polymorphism of a chromosome 5p repeat in control and schizophrenia populations. *Psychiatr. Genet.* **19**, 64–71 (2009). doi: 10.1097/YPG.0b013e3283207ff6; pmid: 19672138
9. D. Thoraval *et al*., Demethylation of repetitive DNA sequences in neuroblastoma. *Genes Chromosomes Cancer* **17**, 234–244 (1996). doi: 10.1002/(SICI)1098-2264(199612)17:4<234::AID-GCC5>3.0.CO;2-4; pmid: 8946205
10. Y.-T. Chen *et al*., Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. *Genes Chromosomes Cancer* **45**, 392–400 (2006). doi: 10.1002/gcc.20298; pmid: 16382448
11. D. T. Ting *et al*., Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **331**, 593–596 (2011). doi: 10.1126/science.1200801; pmid: 21233348
12. K. M. Hassan, T. Norwood, G. Gimelli, S. M. Gartler, R. S. Hansen, Satellite 2 methylation patterns in normal and ICF syndrome cells and association of hypomethylation with advanced replication. *Hum. Genet.* **109**, 452–462 (2001). doi: 10.1007/s004390100590; pmid: 11702227
13. G. A. Logsdon *et al*., The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021). doi: 10.1038/s41586-021-03420-7; pmid: 33828295
14. K. H. Miga *et al*., Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi: 10.1038/s41586-020-2547-7; pmid: 32663838
15. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020). doi: 10.1038/s41576-020-0236-x; pmid: 32504078
16. I. Lee *et al*., Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020). doi: 10.1038/s41592-020-01000-7; pmid: 33230324
17. A. B. Stergachis, B. M. Debo, E. Haugen, L. S. Churchman, J. A. Stamatoyannopoulos, Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020). doi: 10.1126/science.aaz1646; pmid: 32587015
18. Materials and methods are available as supplementary materials.
19. A. Janssen, S. U. Colmenares, G. H. Karpen, Heterochromatin: Guardian of the Genome. *Annu. Rev. Cell Dev. Biol.* **34**, 265–288 (2018). doi: 10.1146/annurev-cellbio-100617-062653; pmid: 30044650
20. J. Chen, Z.-H. Yuan, X.-H. Hou, M.-H. Shi, R. Jiang, LINC01116 promotes the proliferation and inhibits the apoptosis of gastric cancer cells. *Eur. Rev. Med. Pharmacol. Sci.* **24**, 1807–1814 (2020). pmid: 32141549
21. D. Gabellini *et al*., Facioscapulohumeral muscular dystrophy in mice overexpressing FRG1. *Nature* **439**, 973–977 (2006). doi: 10.1038/nature04422; pmid: 16341202
22. G. Giannuzzi *et al*., The human-specific BOLA2 duplication modifies iron homeostasis and anemia predisposition in chromosome 16p11.2 autism individuals. *Am. J. Hum. Genet.* **105**, 947–958 (2019). doi: 10.1016/j.ajhg.2019.09.023; pmid: 31668704
23. Z. Jiang *et al*., Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007). doi: 10.1038/ng.2007.9; pmid: 17922013
24. F. A. M. Maggiolini *et al*., Genomic inversions and GOLGA core duplicons underlie disease instability at the 15q25 locus. *PLOS Genet.* **15**, e1008075 (2019). doi: 10.1371/journal.pgen.1008075; pmid: 30917130
25. J. Schmutz *et al*., The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, 268–274 (2004). doi: 10.1038/nature02919; pmid: 15372022
26. T. W. Prior, Perspectives and diagnostic considerations in spinal muscular atrophy. *Genet. Med.* **12**, 145–152 (2010). doi: 10.1097/GIM.0b013e3181c5e713; pmid: 20057317
27. J. Hauke *et al*., Survival motor neuron gene 2 silencing by DNA methylation correlates with spinal muscular atrophy disease severity and can be bypassed by histone deacetylase inhibition. *Hum. Mol. Genet.* **18**, 304–317 (2009). doi: 10.1093/hmg/ddn357; pmid: 18971205
28. P. Cruz-Tapias, J. Castiblanco, J.-M. Anaya, "Major histocompatibility complex: Antigen processing and presentation," in *Autoimmunity: From Bench to Bedside* (El Rosario Univ. Press, 2013), pp. 271–284.
29. A. Sekar *et al*., Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016). doi: 10.1038/nature16549; pmid: 26814963
30. T. Tsukahara *et al*., Prognostic significance of HLA class I expression in osteosarcoma defined by anti-pan HLA class I monoclonal antibody, EMR8-5. *Cancer Sci.* **97**, 1374–1380 (2006). doi: 10.1111/j.1349-7006.2006.00317.x; pmid: 16995877
31. D. Bello, M. M. Webber, H. K. Kleinman, D. D. Wartinger, J. S. Rhim, Androgen responsive adult human prostatic epithelial cell lines immortalized by human papillomavirus 18. *Carcinogenesis* **18**, 1215–1223 (1997). doi: 10.1093/carcin/18.6.1215; pmid: 9214605
32. Z. Souri *et al*., HDAC inhibition increases HLA class I expression in uveal melanoma. *Cancers* **12**, 3690 (2020). doi: 10.3390/cancers12123690; pmid: 33316946
33. M. Karimzadeh, C. Ernst, A. Kundaje, M. M. Hoffman, Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018). doi: 10.1093/nar/gky677; pmid: 30169659
34. J. T. Simpson *et al*., Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017). doi: 10.1038/nmeth.4184; pmid: 28218898
35. H. Guo *et al*., The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014). doi: 10.1038/nature13544; pmid: 25079557
36. I. K. Suzuki *et al*., Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation. *Cell* **173**, 1370–1384.e16 (2018). doi: 10.1016/j.cell.2018.03.067; pmid: 29856955
37. X. Wu *et al*., CpG island hypermethylation in human astrocytomas. *Cancer Res.* **70**, 2718–2727 (2010). doi: 10.1158/0008-5472.CAN-09-3631; pmid: 20233874

38. P. H. Sudmant *et al*., Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010). doi: 10.1126/science.1197005; pmid: 21030649
39. M. R. Vollger*et al*., Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022). doi: 10.1126/science.abj6965
40. I. T. Fiddes *et al*., Human-specific NOTCH2NL genes affect Notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369.e22 (2018). doi: 10.1016/j.cell.2018.03.051; pmid: 29856954
41. Z. Deng *et al*., A role for CTCF and cohesin in subtelomere chromatin organization, TERRA transcription, and telomere end protection. *EMBO J.* **31**, 4165–4178 (2012). doi: 10.1038/emboj.2012.266; pmid: 23010778
42. A. Ambrosini, S. Paul, S. Hu, H. Riethman, Human subtelomeric duplicon structure and organization. *Genome Biol.* **8**, R151 (2007). doi: 10.1186/gb-2007-8-7-r151; pmid: 17663781
43. C. Li *et al*., DNA methylation reprogramming of functional elements during mammalian embryonic development. *Cell Discov.* **4**, 41 (2018). doi: 10.1038/s41421-018-0039-9; pmid: 30109120
44. S. J. Hoyt *et al*., From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022). doi: 10.1126/science.abk3112
45. J. J. Yunis, W. G. Yasmineh, Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. *Science* **174**, 1200–1209 (1971). doi: 10.1126/science.174.4015.1200; pmid: 4943851
46. C. Jiang, D. Liao, Striking bimodal methylation of the repeat unit of the tandem array encoding human U2 snRNA (the RNU2 locus). *Genomics* **62**, 508–518 (1999). doi: 10.1006/geno.1999.6052; pmid: 10644450
47. G. Nishibuchi, J. Déjardin, The molecular basis of the organization of repetitive DNA-containing constitutive heterochromatin in mammals. *Chromosome Res.* **25**, 77–87 (2017). doi: 10.1007/s10577-016-9547-3; pmid: 28078514
48. A. M. Cotton *et al*., Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* **24**, 1528–1539 (2015). doi: 10.1093/hmg/ddu564; pmid: 25381334
49. A. Hellman, A. Chess, Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141–1143 (2007). doi: 10.1126/science.1136352; pmid: 17322062
50. X. Chen *et al*., Loss of X chromosome inactivation in androgenetic complete hydatidiform moles with 46, XX karyotype. *Int. J. Gynecol. Pathol.* **40**, 333–341 (2021). doi: 10.1097/PGP.0000000000000697; pmid: 33021557
51. P. Bansal, Y. Kondaveeti, S. F. Pinter, Forged by DXZ4, FIRRE, and ICCE: How tandem repeats shape the active and inactive X chromosome. *Front. Cell Dev. Biol.* **7**, 328 (2020). doi: 10.3389/fcell.2019.00328; pmid: 32076600
52. B. P. Chadwick, DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. *Genome Res.* **18**, 1259–1269 (2008). doi: 10.1101/gr.075713.107; pmid: 18456864
53. J. Giacalone, J. Friedes, U. Francke, A novel GC-rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes. *Nat. Genet.* **1**, 137–143 (1992). doi: 10.1038/ng0592-137; pmid: 1302007
54. H. F. Willard, J. S. Waye, Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987). doi: 10.1016/0168-9525(87)90232-0
55. V. A. Shepelev *et al*., Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genom. Data* **5**, 139–146 (2015). doi: 10.1016/j.gdata.2015.05.035; pmid: 26167452
56. N. Altemose *et al*., Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022). doi: 10.1126/science.abl4178
57. R. C. Allshire, G. H. Karpen, Epigenetic regulation of centromeric chromatin: Old dogs, new tricks? *Nat. Rev. Genet.* **9**, 923–937 (2008). doi: 10.1038/nrg2466; pmid: 19002142
58. A. A. Van Hooser *et al*., Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. *J. Cell Sci.* **114**, 3529–3542 (2001). doi: 10.1242/jcs.114.19.3529; pmid: 11682612
59. K. H. Miga, Centromeric satellite DNAs: Hidden sequence variation in the human population. *Genes* **10**, 352 (2019). doi: 10.3390/genes10050352; pmid: 31072070
60. Y. Tanaka, H. Kurumizaka, S. Yokoyama, CpG methylation of the CENP-B box reduces human CENP-B binding. *FEBS J.* **272**, 282–289 (2005). doi: 10.1111/j.1432-1033.2004.04406.x; pmid: 15634350
61. The 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245
62. S. A. Langley, K. H. Miga, G. H. Karpen, C. H. Langley, Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *eLife* **8**, e42989 (2019). doi: 10.7554/eLife.42989; pmid: 31237235
63. Z. Lippman *et al*., Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004). doi: 10.1038/nature02651; pmid: 15269773
64. A. V. Badyaev, Epigenetic resolution of the 'curse of complexity' in adaptive evolution of complex traits. *J. Physiol.* **592**, 2251–2260 (2014). doi: 10.1113/jphysiol.2014.272625; pmid: 24882810
65. M. van Sluis *et al*., Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. *Genes Dev.* **33**, 1688–1701 (2019). doi: 10.1101/gad.331892.119; pmid: 31727772
66. E. M. Darrow *et al*., Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4504–E4512 (2016). doi: 10.1073/pnas.1609643113; pmid: 27432957
67. R. J. L. F. Lemmers *et al*., Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum. Mol. Genet.* **27**, 3488–3497 (2018). doi: 10.1093/hmg/ddy236; pmid: 30281091
68. M. E. Aldrup-MacDonald, M. E. Kuo, L. L. Sullivan, K. Chew, B. A. Sullivan, Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* **26**, 1301–1311 (2016). doi: 10.1101/gr.206706.116; pmid: 27510565
69. D.-H. Koo, F. Han, J. A. Birchler, J. Jiang, Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome. *Genome Res.* **21**, 908–914 (2011). doi: 10.1101/gr.116202.110; pmid: 21518739
70. K. Ichikawa *et al*., Centromere evolution and CpG methylation during vertebrate speciation. *Nat. Commun.* **8**, 1833 (2017). doi: 10.1038/s41467-017-01982-7; pmid: 29184138
71. K. H. Miga, T. Wang, The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021). doi: 10.1146/annurev-genom-120120-081921; pmid: 33929893
72. K. Shafin *et al*., Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020). doi: 10.1038/s41587-020-0503-6; pmid: 32686750
73. C. Jain, A. Rhie, N. Hansen, S. Koren, A. M. Phillippy, A long read mapping method for highly repetitive reference sequences. *bioRxiv* (2020), p. 2020.11.01.363887. doi: 10.1101/2020.11.01.363887
74. F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011). doi: 10.1093/bioinformatics/btr167; pmid: 21493656
75. J. Thakur, S. Henikoff, Unexpected conformational variations of the human centromeric chromatin complex. *Genes Dev.* **32**, 20–25 (2018). doi: 10.1101/gad.307736.117; pmid: 29386331
76. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: 10.1038/nmeth.1923; pmid: 22388286
77. P. Danecek *et al*., Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021). doi: 10.1093/gigascience/giab008; pmid: 33590861
78. Y. Zhang *et al*., Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008). doi: 10.1186/gb-2008-9-9-r137; pmid: 18798982
79. A. Gershman *et al*., Code repositories used for T2T epigenetics for: Epigenetic patterns in a complete human genome, *Zenodo* (2022); https://doi.org/10.5281/zenodo.6046354.
80. M. R. Vollger, A. Lo, Version of SafFire used in T2T figures for: Epigenetic patterns in a complete human genome, Zenodo (2022); https://doi.org/10.5281/zenodo.5911863.
81. A. Zeileis, G. Grothendieck, zoo: S3 infrastructure for regular and irregular time series. *J. Stat. Softw.* **14**, 1–27 (2005). doi: 10.18637/jss.v014.i06
82. M. Sobecki *et al*., MadID, a versatile approach to map protein-DNA interactions, highlights telomere-nuclear envelope contact sites in human cells. *Cell Rep.* **25**, 2891–2903.e5 (2018). doi: 10.1016/j.celrep.2018.11.027; pmid: 30517874
83. K. D. Hansen, B. Langmead, R. A. Irizarry, BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012). doi: 10.1186/gb-2012-13-10-r83; pmid: 23034175
84. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi: 10.1093/bioinformatics/btp324; pmid: 19451168
85. J. R. Miller *et al*., Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008). doi: 10.1093/bioinformatics/btn548; pmid: 18952627
86. C. A. Davis *et al*., The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46** (D1), D794–D801 (2018). doi: 10.1093/nar/gkx1081; pmid: 29126249
87. M. Kokot, M. Dlugosz, S. Deorowicz, KMC 3: Counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017). doi: 10.1093/bioinformatics/btx304; pmid: 28472236
88. F. Ramírez *et al*., deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44** (W1), W160–W165 (2016). doi: 10.1093/nar/gkw257; pmid: 27079975
89. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
90. P. Rice, I. Longden, A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000). doi: 10.1016/S0168-9525(00)02024-2; pmid: 10827456
91. M. Kirsche, A. Das, M. C. Schatz, Sapling: Accelerating suffix array queries with learned data models. *Bioinformatics* **37**, 744–749 (2021). doi: 10.1093/bioinformatics/btaa911; pmid: 33107913
92. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi: 10.1093/molbev/mst010; pmid: 23329690
93. R. Bouckaert *et al*., BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019). doi: 10.1371/journal.pcbi.1006650; pmid: 30958812
94. T. Marques-Bonet, O. A. Ryder, E. E. Eichler, Sequencing primate genomes: What have we learned? *Annu. Rev. Genomics Hum. Genet.* **10**, 355–386 (2009). doi: 10.1146/annurev.genom.9.081307.164420; pmid: 19630567
95. M. de Manuel *et al*., Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016). doi: 10.1126/science.aag2602; pmid: 27789843
96. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019). doi: 10.1038/s41587-019-0201-4; pmid: 31375807
97. S. Kovaka *et al*., Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019). doi: 10.1186/s13059-019-1910-1; pmid: 31842956

ACKNOWLEDGMENTS

We thank R. Workman, S. Sholes, and A. Royles for reading and editing the manuscript; I. Lee for engaging in discussion about epigenetics and chromatin state; and Circulomics for help in ultra-high-molecular-weight DNA extraction. **Funding:** This study was supported by the National Institutes of Health (grant R01HG009190 to W.T., grant F32 GM134558 to G.A.L., grants R24 DK106766-01A1 and U24HG010263 to M.C.S., grants 1R01HG011274-01 and 1U01HG010971 to K.H.M., and an Intramural Research Program grant from the National Human Genome Research Institute to S.K., A.R., and A.M.P.); a Damon Runyon Postdoctoral Fellowship and PEW Latin American Fellowship to G.V.C.; and a Howard Hughes Medical Institute Hanna Gray Fellowship to N.A. **Author contributions:** K.H.M. and W.T. conceived the study. W.T., A.M.P., K.H.M., and A.G. coordinated the collaboration. K.H.M. and N.A. performed repeat characterization and satellite DNA assembly. A.G., M.E.G.S., and M.C.S. performed ENCODE and mappability analyses. G.V.C. and N.A. performed CUT&RUN. N.A., A.R., and S.K. performed ONT mapping and methylation calling. S.J.H. and R.J.O. performed TE and noncentromeric repeat annotations. A.G. and A.S. performed gene annotation liftover. M.R.V., G.A.L., and E.E.E. performed SegDup annotations. A.R. performed marker-assisted mapping of CUT&RUN data. A.G., P.W.H, and R.R. performed HG002 cell culture and nanoNOMe sequencing and analysis. A.G., M.R.V., X.G., and E.E.E. performed phylogenetic and aging analysis of *NBPF* genes. M.J. performed megalodon methylation calling. A.G., W.T., and K.H.M. developed figures. A.G. and W.T. drafted the manuscript. All authors provided critical feedback and read and approved the final manuscript. **Competing interests:** W.T. has two patents (8,748,091 and 8,394,584) licensed to ONT. K.H.M. and W.T. have received travel funds to speak at symposia

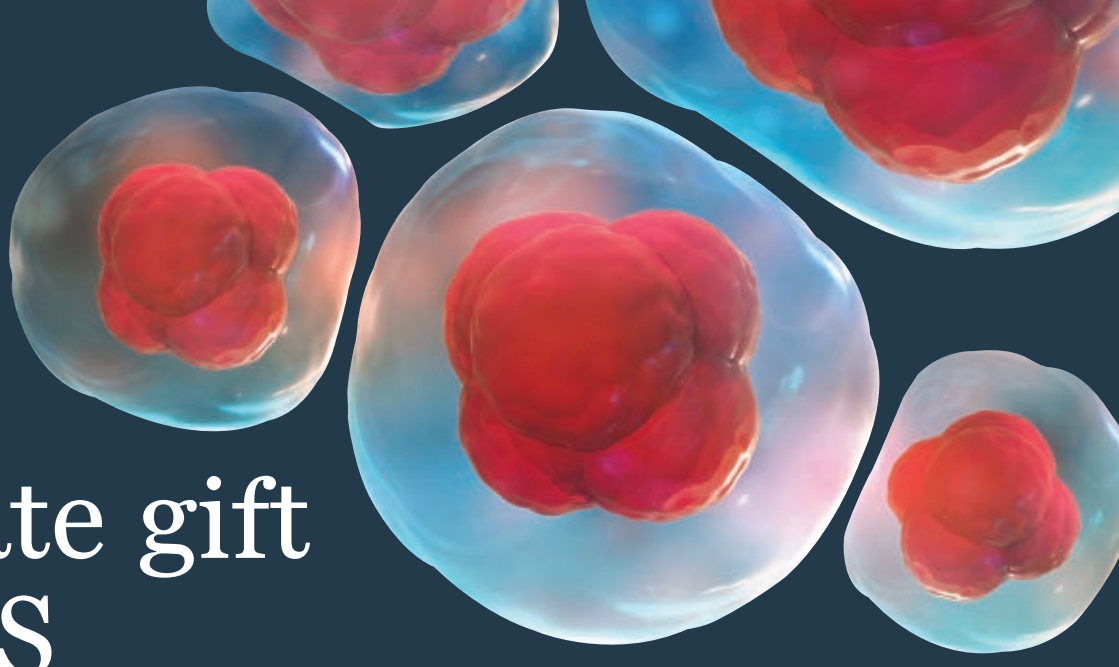
organized by ONT. K.H.M. is a scientific advisory board member of Centaura, Inc. **Data and materials availability:** Nanopolish methylation calls are available on Zenodo (*79*). HG002 nanoNOMe data can be accessed on the Sequence Read Archive BioProject with accession number PRJNA725525. CUT&RUN data on CHM13 and HG002 can be accessed on the Sequence Read Archive with BioProject accession numbers PRJNA559484 and PRJNA752795. All other datasets used in this study are properly cited with accessions referenced in the methods and materials. CHM13hTERT cells were obtained for research use through a material transfer agreement with U. Surti and the University of Pittsburgh. Code for all CHM13 and HG002 CpG and GpC methylation, the ENCODE analysis pipeline, mappability analysis (MUK), and NBPF timing analysis code and alignments are available on Zenodo (*79*). SatFire figures are also available at Zenodo (*80*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj5089
Materials and Methods
Figs. S1 to S28
References (81–97)
Tables S1 to S9
MDAR Reproducibility Checklist

26 May 2021; resubmitted 29 November 2021
Accepted 14 February 2022
10.1126/science.abj5089

1848 SOCIETY
AAAS



An estate gift to AAAS

Going all the way back to 1848, our founding year, the American Association for the Advancement of Science (AAAS) has been deeply committed to advancing science, engineering and innovation around the world for the benefit of all people.

Today, we are dedicated to advocating for science and scientific evidence to be fully and positively integrated into public policy and for the community to speak with one voice to advance science and engineering in the United States and around the world.

By making AAAS a beneficiary of your will, trust, retirement plan or life insurance policy, you will become a member of our 1848 Society and will help fuel our work on behalf of science and society – including publishing the world's most promising, innovative research in the *Science* family of journals and engaging in the issues that matter locally, nationally and around the world.

"As a teacher and instructor, I bear responsibility for the younger generations. If you have extra resources, concentrate them on organizations, like AAAS, that are doing work for all."

—Prof. Elisabeth Ervin-Blankenheim, 1848 Society member

If you intend to include AAAS in your estate plans, provide this information to your lawyer or financial adviser:

Legal Name: American Association for the Advancement of Science
Federal Tax ID Number: 53-0196568
Address: 1200 New York Avenue, NW, Washington, DC 20005

If you would like more information on making an estate gift to AAAS, cut out and return the form below or send an email to philanthropy@aaas.org. Additional details are also available online at **www.aaas.org/1848Society**.

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

cut here

Yes, I would like more information about joining the AAAS 1848 Society.

1848 SOCIETY
AAAS

PLEASE CONTACT ME AT:

Name: _____

Address: _____

City: _____ State: _____ Zip code: _____ Country: _____

Email: _____ Phone: _____

RETURN THIS FORM TO:
AAAS Office of Philanthropy and Strategic Partnerships • 1200 New York Avenue, NW • Washington, DC 20005 USA

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**



PHOTO: ROBERT NICHOLLS

An artist's conception of an early mammal with a relatively large body relative to brain size

PALEONTOLOGY

Body first

Mammals have the largest ratio of brain to body size (encephalization) among vertebrates. It has been believed that this relationship emerged early on in mammalian evolution, with enlarging brains leading the way into new and diverse forms. However, Bertrand *et al*. looked at encephalization rates across mammals beginning in the Paleocene and found instead that body sizes were the first to increase, allowing for niche filling after the extinction of the dinosaurs (see the Perspective by Smith). It was only later, in the Eocene, that brain size began to increase, likely driven by a need for greater cognition in increasingly complex environments. This led to the highly encephalized brains of today, including those of humans. —SNV

Science, abl5584, this issue p. 80; see also abo1985, p. 27

GAMMA-RAY ASTRONOMY

Proton acceleration in a recurrent nova

If a white dwarf strips material from a companion star, then enough hydrogen can build up on the surface to trigger a thermonuclear explosion, ejecting material without destroying the white dwarf. This is observed as a nova. Some novae have been seen to emit high-energy gamma rays, but the origin of that emission has been unclear. The H.E.S.S. Collaboration observed the 2021 outburst of RS Ophiuchi, a recurrent nova, determining its spectral and temporal evolution at giga–electron volt and tera–electron volt energies. Modeling of the emission physics shows that the expanding nova shock wave efficiently accelerated protons, providing a source of gamma rays and cosmic rays. —KTS

Science, abn0567, this issue p. 77

CHRONIC PAIN

Pain, pain, go away...

Damage to the nervous system pathologically alters the somatosensory system, which can result in neuropathic pain. Although pain development has been well studied, the mechanisms orchestrating pain recovery remain unclear. Working in nerve-injured mice, Kohno *et al*. identified a CD11c$^+$ spinal microglia population that appears after pain development (see the Perspective by Sideris-Lampretsas and Malcangio) and is essential for recovery from neuropathic pain. The ability of these cells to promote pain recovery relied on their high levels of expression of insulin-like growth factor-1. The CD11c$^+$ microglia remained even after pain recovery, and pain hypersensitivity returned if they were depleted. These findings elucidate mechanisms underlying remitting and relapsing neuropathic pain and could potentially help in the development of therapeutic strategies. —SMH

Science, abf6805, this issue p. 80; see also abo5592, p. 33

HUMAN EVOLUTION

5000 years of Xinjiang genetics

The Xinjiang region of China is bordered by mountains and represents an important historical region. Sampling ancient genomes, Kumar *et al*. investigated the changes in populations of this region over time from the Bronze Age, ~5000 to 3000 years before present (BP), covering the Iron Age, ~3000 to 2000 years BP,

and into the Historical Era, ~2000 years BP. This analysis found that older individuals represented ancestries from Steppe cultures, and that a later inflow of East and Central Asian ancestry entered the region around the end of the Bronze Age toward the beginning of the Iron Age. During the Historical Era, mixing continued but retained a core Steppe component such that populations form a genetic continuum. This retention of genetic continuity in a central population is surprising because it represents patterns more typically observed in isolated populations. Furthermore, these genetic links identify a previously unknown lineage that could potentially explain the spread of the Indo-European languages. —LMZ

Science, abk1534, this issue p. 62

ENGINEERING

Nanoneedle-based ocular drug delivery

Traditional ocular drug delivery commonly requires topical administration, which has low delivery efficiency because of the complex eye environment. Large drug doses or frequent topical applications could lead to a high risk of side effects. Park *et al.* report a tear-soluble contact lens platform integrated with silicon nanoneedles for ocular drug delivery. The contact lens quickly dissolves in tear fluid and releases anti-inflammatory drugs, and the silicon nanoneedles can penetrate into the cornea painlessly and degrade gradually to realize long-term sustained therapeutic drug delivery. This platform was demonstrated in a rabbit model to treat a chronic ocular disease with reduced side effects compared with current standard approaches. —WG

Sci. Adv. 10.1126/sciadv.abn1772 (2022).

RESPIRATORY DISORDERS

Young lung inflammation

Chorioamnionitis is a potentially fatal inflammatory condition of the fetus caused by inflammation of the chorion and/or amnion that is linked to a large fraction of preterm births and an elevated risk for respiratory disease in childhood and adulthood. Toth *et al*. describe the development of a prenatal rhesus macaque model that used intra-amniotic lipopolysaccharide (LPS) challenge to induce experimental chorioamnionitis. This model mirrored structural and temporal changes consistent with prenatal human lung development. LPS challenge resulted in extensive damage to alveolar structures that resembled chronic lung disease in newborns, and single-cell RNA sequencing analysis showed dramatic disruption of alveologenesis signals. Intrauterine blockade of interleukin-1β and tumor necrosis factor-α ameliorated inflammatory responses and restored lung integrity. —CNF

Sci. Transl. Med. **14**, eabl8574 (2022).

MEMBRANES

Pulling carbon dioxide out of the air

A challenge in the design of polymeric membranes for gas separation is the trade-off between permeability, or how fast gases can flow through the membrane, and selectivity, the ability to separate one gas from another. In general, the more selective the membrane, the more slowly gases can flow through it. Sandru *et al*. overcame this trade-off through a layered design. They used a bottom layer of porous polyacrylonitrile that acts as a physical support for the middle layer of either elastomer-like polydimethylsiloxane or glassy-type polytetrafluoroethylene. The authors then grafted a patchy layer of polyvinylamine, which selectively attracts carbon dioxide, thus pulling it into the membrane and leading to much higher separation from nitrogen. —MSL

Science, abj9351, this issue p. 90

IN OTHER JOURNALS

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

PLANT ECOLOGY

Precision pollination

Plants that rely on pollinators for fertilization have evolved ways of making pollination more efficient. Stewart *et al.* tested whether flower symmetry, petal fusion, and stamen number affect the placement of pollen on pollinators' bodies. They observed generalist and specialist pollinators (including native honeybees and nectar bats) visiting flowers in Thailand, and then measured pollen deposited on different areas of the body. Flowers with bilateral symmetry, fused petals, and fewer stamens deposited pollen on specific areas. Thus, the bilateral flower structure helped the precision and efficiency of pollination across pollinator taxa. —BEL

New Phytol. 10.1111/nph.18050 (2022).



The bat *Eonycteris spelaea* visiting bilaterally symmetrical *Oroxylum indicum* flowers, which have fused corollas, allowing precise deposition of pollen on the bat's head.

PHOTO: MERLIN TUTTLE.ORG/SCIENCE SOURCE

VIROLOGY

Baby steps toward protection

Respiratory syncytial virus (RSV) is a major cause of lower respiratory tract infections in young children, presenting a particular danger to premature babies. The only specific intervention is palivizumab, an antibody that is given monthly to high-risk infants, but it only offers partial protection from RSV. Hammitt *et al*. report on the results of clinical testing of nirsevimab, another monoclonal antibody. Unfortunately, the authors did not compare nirsevimab directly with palivizumab, but its once per season dosing and efficacy compared with placebo make it a promising new contender for RSV prevention. —YN

N. Engl. J. Med. **386**, 837 (2022).

ELECTROCHEMISTRY

Predicting ion mobility in solids

There is considerable interest in detecting correlations between physicochemical properties of materials tailored for certain applications. Such correlations, if determined to be relevant to the underlying functionality and to be applicable to a broad range of materials, can become a major tool in computational screening for optimized practical solutions. Using density functional theory calculations, Sotoudeh and Groß derived a universal and reliable descriptor for ion mobility in crystalline solids, a critical performance parameter for electrochemical energy storage devices. The proposed descriptor exhibits high predictive power and is based on scaling relations between the materials' features



that are readily available and capture the key factors influencing ion mobility in solids. —YS

JACS Au **2**, 463 (2022).

QUANTUM LOGIC

Solving a puzzle with quantum logic

Swiss mathematician Leonard Euler devised a sudoku-style mathematical puzzle in 1779 that has been mathematically proven to be unsolvable. The puzzle consists of a 6 × 6 board and 36 "army officers," with each officer belonging to one of six different regiments (e.g., A, B, C...), and also having one of six different ranks (e.g., 1, 2, 3...). Rather *et al.* created and solved a "quantum" version of this puzzle in which each officer can assume a superposition of multiple regiments and ranks, e.g., partially A and C and partially 2 and 4. The quantum version of this puzzle entangles the properties of the 36 elements and provides an elegant way to explore the nascent field of quantum combinatorics. It will also be interesting to find out if there is something special about the size of the 6 × 6 board. —YY

Phys. Rev. Lett. **128**, 080507 (2022).

METAL NANOPARTICLES

Surface ligands and electronic density

Metal nanoparticles are normally stabilized by organic surface ligands, but how these ligands affect the electronic density of states of nanoparticles in solution has been difficult to determine experimentally. Litak *et al*. assessed changes in the Fermi level (E_F) of alkanethiol-coated gold nanoparticles with a differential nuclear magnetic resonance using different solutions in coaxial sample tubes. Chemical shifts in hexane peaks in the outer tube that occurred when the nanoparticles were present relative to solvent in the inner tube measured the induced Pauli paramagnetic susceptibility and in turn E_F. Magnetic susceptibility was lower with longer alkyl chain ligands that donated more charge across the gold–sulfur bond and in turn pushed E_F higher into the conduction band. —PDS

ACS Nano **16**, 4479 (2022).

INFECTION

Bacterial spanner in the works

How the mammalian immune system coordinates responses to all of the microbes that it encounters simultaneously remains an open question. In mice, Biram *et al.* report that innate immune responses to some pathogens can disrupt ongoing adaptive immune responses to others. Infection with bacteria such as *Salmonella* Typhimurium or *Listeria monocytogenes* triggers the recruitment of Sca1$^+$ monocytes to lymph nodes, where their antimicrobial activity disrupts cell respiration. B cells in germinal centers, which are crucial for antibody production, then die off because their specialized metabolisms make them vulnerable. Bacterial lipopolysaccharide alone was sufficient to trigger the production of monocytes and collapse germinal centers, indicating that the use of bacterial adjuvants in vaccines may in many cases do more harm than good. —STS

Immunity **55**, 442 (2022).

HIV

Traveling with HIV

Why do test-and-treat HIV prevention trials have such disappointing effects on overall HIV prevalence? Regions in eastern and southern Africa still have very high levels of HIV infections despite significant public health efforts. Magosi *et al.* assessed test-and-treat control measures in 30 communities in Botswana. Deep-sequencing and phylogenetic analysis showed that most transmission occurred among people roughly the same age within a community or from an adjacent community. More HIV was imported into treatment communities than was exported. Therefore, for HIV, as for other pathogens, movement of infected individuals can play a pivotal role in long-term persistence. The work also shows that although focused public health efforts may prevail within a community, these efforts may be rather easily undermined. —CA

eLife **11**, e72657 (2022).

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

SOLAR CELLS

Stabilizing inverted solar cells

Although inverted (p-i-n) perovskite solar cells (PSCs) have advantages in fabrication and scaling compared with n-i-p cells, their power conversion efficiencies (PCEs) are usually lower. Azmi *et al*. show that by tailoring the number of octahedral inorganic sheets in two-dimensional perovskite (2DP) passivation layers for three-dimensional perovskite active layers, PCEs of more than 24% could be achieved (see the Perspective by Luther and Schelhas). The 2DP layers formed with oleylammonium iodide molecules at the electron-selective interface passivated trap states and suppressed ion migration. These PSCs retained more than 95% of their initial efficiency after 1000 hours of damp-heat testing (85°C and 85% relative humidity), which passes a key industrial stability standard. —PDS

Science, abm5784, this issue p. 73; see also abo3368, p. 28

COGNITIVE SCIENCE

Humans succeed through social learning

Our capacity to accumulate complex algorithms over generations allows human beings to adapt to diverse environments and solve challenges that go beyond our individual limitations. However, cultural accumulation of innovative algorithms is difficult to explain. Thompson *et al*. studied a large number of participants to explore the evolution of algorithms under different learning conditions (see the Perspective by Henrich). Selective social learning that involved knowledge of the success level of different strategies or of different models preserved difficult-to-invent, efficient algorithms more than random social learning or one-attempt asocial learning. Two efficient algorithms were used by many people, but the most efficient one only spread under selective social learning. —PRS

Science, abn0915, this issue p. 95; see also abo0713, p. 31

T CELLS

Checks and balances

Regulatory T cells (T_{regs}) are critical for maintaining immune tolerance, but their suppressive functions must be temporarily overcome to generate productive immune responses against pathogens and tumors. Teh *et al*. demonstrate that caspase-8 regulates the homeostasis of T_{regs} by tuning their susceptibility to inflammatory cell death. Caspase-8–deficient T_{regs} were protected against apoptotic cell death during homeostasis but were more susceptible to necroptotic cell death during chronic viral infection. Human T_{regs} were also more sensitive to necroptosis induced by caspase inhibition than conventional T cells. Although caspase-8 in T_{regs} functioned to prevent lethal inflammation, it also antagonized antiviral and antiparasitic immunity, highlighting a key regulatory mechanism in controlling the balance between immunity and pathologic inflammation. —CO

Sci. Immunol. **7**, eabn8041 (2022).

NEUROSCIENCE

Palmitoylation for axonal degeneration

In response to nerve injury, activation of the kinases DLK and JNK3 in axons induces retrograde signaling to the soma, where gene expression changes promote neuronal degeneration. Niu *et al*. found that palmitoylation of both DLK and JNK3 mediated their colocalization onto axon vesicles, where a positive-feedback phosphorylation loop between the kinases maintained their activation state as the vesicles trafficked to the soma. Moreover, JNK3 palmitoylation was essential for axonal retrograde signaling in response to optic nerve crush injury in mice. —LKF

Sci. Signal. **15**, eabh2674 (2022).

CHEMISTRY

Computationally designed pesticides

Computational modeling of the properties of small-molecule pesticides promises to optimize performance while minimizing ecotoxicity. Kostal *et al.* developed an in silico strategy to design safe pesticides for which parameters for aquatic toxicity and photodegradation rate are considered. The safest pesticide is less bioavailable and more readily degraded photochemically. Focusing on phenols and anilines and analyzing 700 known compounds, they concluded that phenol derivatives are the safest pesticides and propose further structural optimizations based on their computed properties. The authors make a compelling case that their computational approach should be used in the rational design of potent and safer pesticides. —SRC

Sci. Adv. 10.1126/sciadv.abn2058 (2022).

ECOLOGY

Genes to ecology

In the past few decades, the identification of keystone species, that is, those with essential roles in structuring a community or ecosystem, has increased across systems. Barbour *et al*. extended this concept to genes, showing that a single allele of a particular plant defense gene facilitates species coexistence across a small experimental trophic system (see the Perspective by Nosil and Gompert). Specifically, plants with this allele grew faster, supporting larger populations of two species of herbivores and their predators. This finding suggests that genotype variation can play a role in the structure and function of organismal systems. —SNV

Science, abf2232, this issue p. 70; see also abo3575, p. 30

Science
JOURNALS AAAS



Publish your research in the *Science* family of journals

The *Science* family of journals (*Science, Science Advances, Science Immunology, Science Robotics, Science Signaling*, and *Science Translational Medicine)* are among the most highly-regarded journals in the world for quality and selectivity. Our peer-reviewed journals are committed to publishing cutting-edge research, incisive scientific commentary, and insights on what's important to the scientific world at the highest standards.

Submit your research today!

Learn more at **Science.org/journals**

RESEARCH ARTICLE

HUMAN EVOLUTION

Bronze and Iron Age population movements underlie Xinjiang population history

Vikas Kumar[1,2], Wenjun Wang[1,3], Jie Zhang[4], Yongqiang Wang[4], Qiurong Ruan[4], Jianjun Yu[4], Xiaohong Wu[5], Xingjun Hu[4], Xinhua Wu[6], Wu Guo[6], Bo Wang[7], Alipujiang Niyazi[4], Enguo Lv[4], Zihua Tang[8], Peng Cao[1], Feng Liu[1], Qingyan Dai[1], Ruowei Yang[1], Xiaotian Feng[1], Wanjing Ping[1], Lizhao Zhang[1], Ming Zhang[1], Weihong Hou[1], Yichen Liu[1,2]*, E. Andrew Bennett[1]*, Qiaomei Fu[1,2,9]*

The Xinjiang region in northwest China is a historically important geographical passage between East and West Eurasia. By sequencing 201 ancient genomes from 39 archaeological sites, we clarify the complex demographic history of this region. Bronze Age Xinjiang populations are characterized by four major ancestries related to Early Bronze Age cultures from the central and eastern Steppe, Central Asian, and Tarim Basin regions. Admixtures between Middle and Late Bronze Age Steppe cultures continued during the Late Bronze and Iron Ages, along with an inflow of East and Central Asian ancestry. Historical era populations show similar admixed and diverse ancestries as those of present-day Xinjiang populations. These results document the influence that East and West Eurasian populations have had over time in the different regions of Xinjiang.

Xinjiang in northwest China has played a central role in the exchange of material culture, agriculture, and technology between the West and East Eurasian peoples (*1*–*4*). The Xinjiang region is surrounded by the Altai Mountains to the north and the Kunlun and Pamir mountain ranges to the south. The Tianshan Mountains in the central west region separate Xinjiang into the Zungharian and Tarim Basins, consisting mostly of arid semidesert with habitable regions around the rivers providing fertile land for farming (Fig. 1) (*2*, *3*, *5*, *6*). In the Bronze Age (BA), metallurgical technologies traveled through Xinjiang to East Asia, agriculturally important plants such as wheat and barley [~5000 before the present (B.P.)] traversed from the west through the Inner Asian Mountain Corridor (IAMC), and millet (~4000 B.P.) is presumed to have entered Xinjiang from the east through the Hexi Corridor (*3*–*5*, *7*). An understanding of the origins of the BA inhabitants of Xinjiang is necessary to follow the changes that these cultural and technological transfers have had on the population structure over the ensuing millennia.

Several hypotheses have been proposed regarding the origins of the BA populations of Xinjiang, which have been linked to migrations from the influential cultures surrounding the region. To the northwest were the Steppe cultures of the Afanasievo, Chemurchek, Okunevo, and Botai; to the south was the Central Asian Oxus or Bactria–Margiana Archaeological Complex (BMAC) civilization; and to the east, the Siba culture was present around the Hexi Corridor (*2*, *8*–*12*). Within this context, archaeological and mitochondrial studies have suggested that the BA inhabitants and cultures of Xinjiang were not derived from any indigenous Neolithic substrate but rather from a mix of West and East Eurasian people (*5*, *13*, *14*), whereas BA burial traditions suggest links with both North Eurasian Steppe cultures and the Central Asian BMAC civilization (*15*). Linguistically, the presence of the now-extinct Indo-European Tocharian language, attested to in ~5th to 10th century CE texts from the Tarim Basin, also raised questions regarding the origin and extent of Indo-European–speaking people in Xinjiang (*7*, *15*). To date, one mitochondrial study of the BA Xiaohe site in the Tarim Basin found evidence of both Steppe- and central Siberian–derived haplogroups and only a later influence from the BMAC (*13*), whereas a broader mitochondrial survey indicated that these results may not apply to the region as a whole (*14*). A recent genomic analysis of the BA Tarim Basin mummies found evidence of a local ancestry derived in part from ancient North Eurasians (ANEs) (*16*). Thus, to gain a more complete picture of the BA settlement of Xinjiang, in-depth exploration using the more precise analysis of nuclear DNA across a geographically comprehensive collection of BA sites will be needed.

The Iron Age (IA) of East and West Asia is marked by extensive population movements, genetic admixtures, and cultural changes (*1*, *11*, *14*, *17*). In Xinjiang, the first iron materials, which date from before the first millennium (~3150 B.P.), were linked to Steppe nomads (*5*), such as the Sakas, or to the Scythians, a broad and important IA nomadic culture whose presence has been reported in many archaeological sites of the Yili River Valley in northwest Xinjiang and the southern Tarim Basin (*5*). Many nomadic confederations, such as the Saka, Hun, Pazyryk, Xiongnu, and Tagar, arose in regions surrounding Xinjiang during the IA (*1*, *3*). The people of these nomadic cultures maintained a high diversity in IA Xinjiang (*13*, *14*). Of these, the Sakas were the descendants of Late Bronze Age (LBA) herders (such as the Andronovo, Srubnaya, and Sintashta) with additional ancestries derived from Lake Baikal (Shamanka_EBA) (EBA, Early Bronze Age) and BMAC populations (*1*, *17*, *18*). Sakas have been associated with the Indo-Iranian Khotanese language, which was spoken in southern Xinjiang before spreading to other parts of the region (*19*). A previous genomic study from a single IA site has suggested the presence of Steppe-related ancestry in IA Xinjiang individuals (*20*); however, the high level of genetic variability and mobility reported for the broader region highlights the necessity of a broader investigation for a comprehensive understanding of the IA Xinjiang populations. The periods after 2200 B.P. oversaw several noteworthy struggles for control of the region by neighboring powers, such as the Yuezhi, Xiongnu, Han, and Turks (*1*, *3*, *17*). Thus, Xinjiang represents a key area for studying the past confluence and coexistence of populations with dynamic cultural, linguistic, and genetic backgrounds. To track the demographic changes of the Xinjiang population over the cultural and technological transitions during the past 5000 years, we generated and analyzed genome-wide data from 201 archaeological samples, collected from 39 sites across Xinjiang and dating from the BA and IA to the historical era (HE).

Results

We collected archaeological samples from 39 sites and generated genome-wide data from 201 ancient individuals; 104 of these individuals were directly radiocarbon dated, whereas dates were inferred using archaeological information for the remaining 97 (Fig. 1 and table S1) (*21*). DNA was extracted from bones, and libraries were constructed using previously described protocols (*22*, *23*). Samples that passed ancient DNA filters for DNA damage and low contamination rates (<5%) were used for analyses (table S1; details are given in the

[1]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Beijing 100044, China. [2]Shanghai Qi Zhi Institute, Shanghai 200232, China. [3]National Centre for Archaeology, Beijing 100013, China. [4]Institute of Cultural Relics and Archaeology in Xinjiang, Urumqi 830011, China. [5]School of Archaeology and Museology, Peking University, Beijing 100871, China. [6]Institute of Archaeology, Chinese Academy of Social Science, Beijing 100710, China. [7]Xinjiang Uygur Autonomous Region Museum, Urumqi 830002, China. [8]Institute of Geology and Geophysics, Chinese Academy of Science, Beijing 100020, China. [9]University of Chinese Academy of Sciences, Beijing 100049, China.
*Corresponding author. Email: fuqiaomei@ivpp.ac.cn (Q.F.); eabennett@gmail.com (E.A.B.); yichen.liu@ivpp.ac.cn (Y.L.)

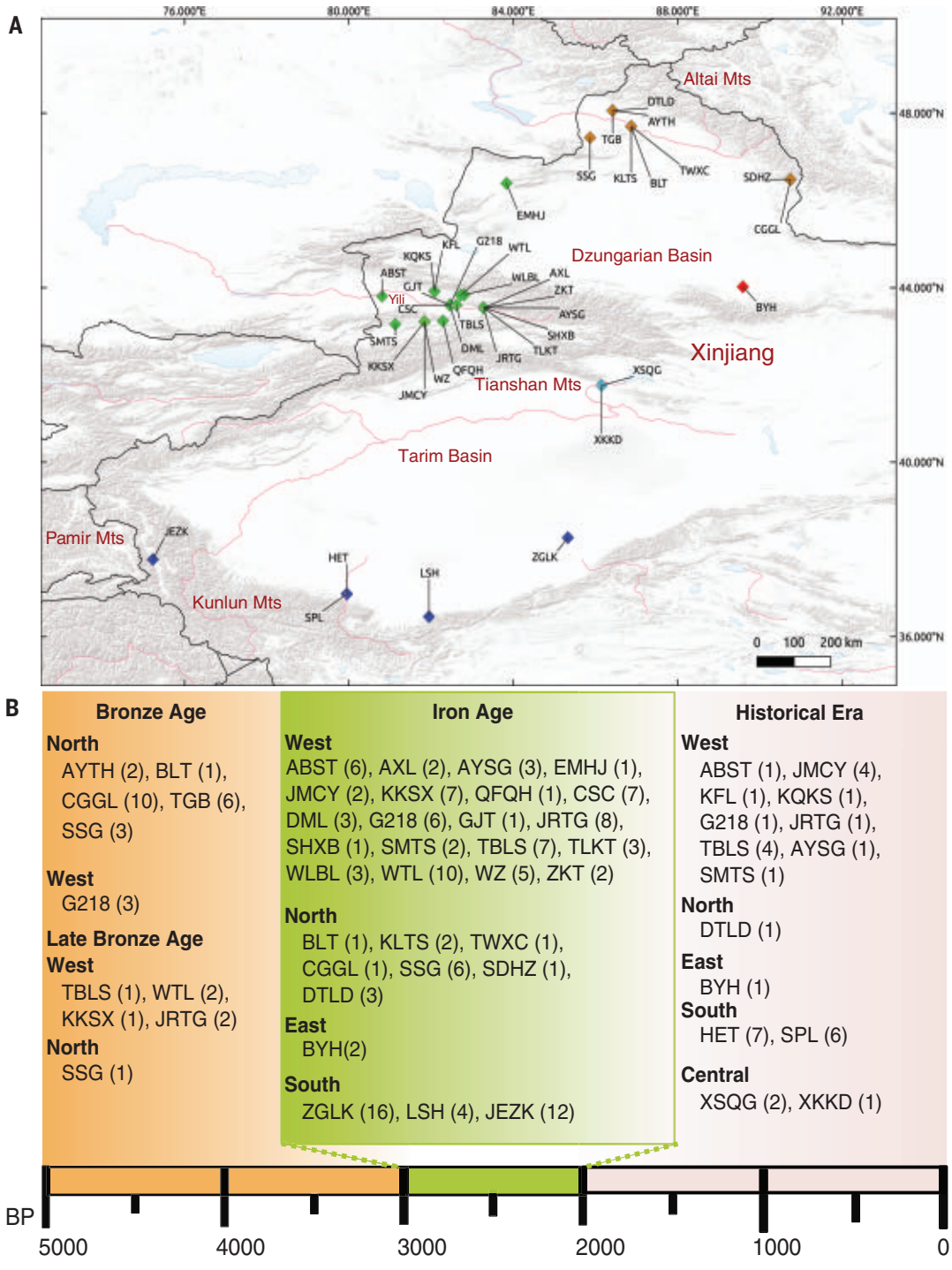


Fig. 1. Xinjiang sampling locations along with their groupings used in this study. (**A**) Map showing the geographic sampling locations of the archaeological sites included in this study. (**B**) BA, IA, and HE time periods are depicted on a timeline in years before the present (BP). The number of individuals from each site and time period is given in parentheses. Abbreviations are defined in table S1 and the supplementary materials.

supplementary materials) (*21*, *24*). Genomic libraries were enriched to target ~1.2 million single-nucleotide polymorphims (SNPs) for human DNA (*25*), generating a mean coverage ranging from 0.01× to 8.60×. Later, pseudo-haploid genotypes were called on the targeted SNPs, resulting in ~5000 to 1,141,000 SNPs (table S1) (*21*). Additionally, among the 201 individuals, 27 with adequate endogenous DNA content were shotgun sequenced to a coverage of 0.22× to 3.60×, resulting in ~232,000 to 1,077,000 SNPs (table S1) (*21*). We sampled across Xinjiang from north (n = 40), west (n = 105), south (n = 49), central (n = 3), east (n = 3), and unknown subregions (n = 1). The sequenced individuals were from either the BA (Xinj_BA), ~5000 to 3500 B.P.; the LBA (Xinj_LBA), ~3500 to 3000 B.P.; the IA (Xinj_IA), ~3000 to 2000 B.P.; or the HE (Xinj_HE), <2000 B.P. (table S1) (*21*). After removing five individuals for high contamination, kinship testing identified 87 pairs of related individuals among all samples (table S2) (*21*). For the related individuals, only those with the higher number of SNPs (n = 43) were used for further analysis, which resulted in 152 unrelated individuals after discarding two with <30,000 SNPs (table S1) (*21*). After analyzing the newly sequenced individuals along with published ancient and present-day populations using principal components analysis (PCA), ADMIXTURE (*26*), f-statistics (*27*), qpAdm (*28*), and DATES (*11*), we observed that these individuals were highly admixed, and many contained unique ancestries. On the basis of these observations, we also regrouped individuals from the 39 archaeological sites into 64

subgroups for the follow-up analyses, primarily using PCA and ADMIXTURE to identify genetically homogeneous individuals, including those found across neighboring sites and similar time periods (table S1) (*21*).

BA Xinjiang shows a mix of eastern and western Steppe ancestry

Attempts to explain the origins of BA Xinjiang populations have focused on similarities with surrounding Steppe and Central Asian populations. Two major competing hypotheses have been proposed to describe the first BA settlers of the region. The Steppe hypothesis states that the Tarim Basin was settled by the Afanasievo Steppe culture to the north and finds support from similarities between cultural artifacts, burial customs, and skeletal characteristics (*5*, *13*, *14*). The Bactrian oasis hypothesis highlights the similarities in the desert basin environment and sustenance practices with the BMAC culture to the west of Xinjiang across the Pamir and Tianshan mountain ranges, connected through the IAMC (*15*). Additional archaeological evidence also suggests East Asian connections from the Gansu and Qinghai (GanQing) regions of northern China through the Hexi Corridor (*3*). We investigate the origins of BA Xinjiang populations through genome-wide data of 20 BA individuals from six sites merged with published BA individuals (*16*) and an additional seven LBA individuals (<3500 B.P.) from five sites, all in northern and western Xinjiang (Fig. 1 and table S1) (*21*). We observe a high affinity between BA and LBA Xinjiang populations and those with ANE ancestry

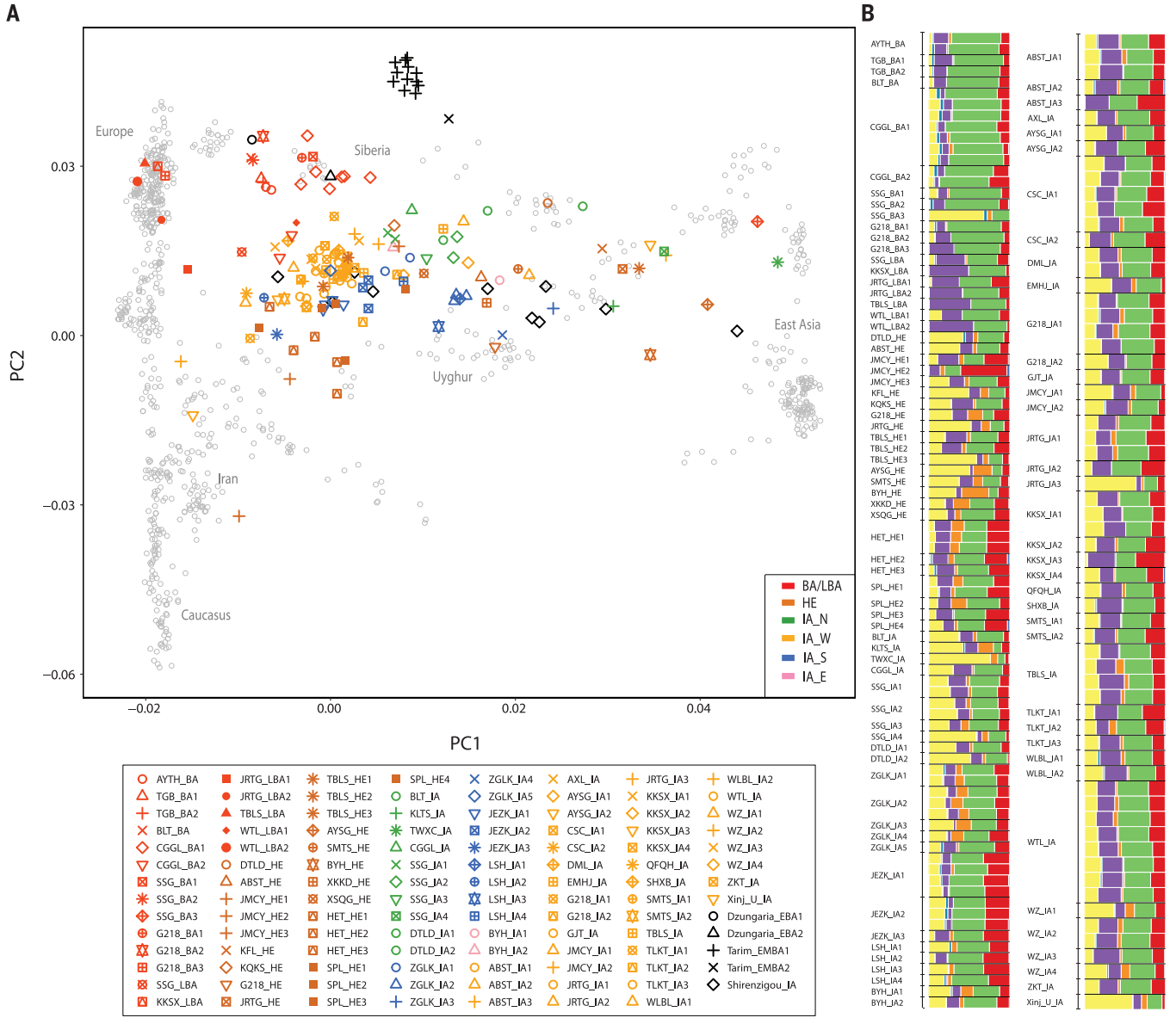


Fig. 2. PCA and ADMIXTURE analyses of Xinjiang populations. (**A**) Ancient Xinjiang and other ancient populations are shown as different colors and shapes. Present-day populations are shown as gray circles, and only major groups are included. Most of the ancient Xinjiang populations lie on the cline extending from European and Siberian to East Asian populations. The published populations of Dzungaria_EBA, Tarim_EMBA, and Shirenzigou_IA are depicted in various black shapes. IA north (IA_N), south (IA_S), west (IA_W), and east (IA_E) are the geographical locations of the IA individuals. (**B**) ADMIXTURE analysis of all the newly reported ancient individuals at $K = 7$. The four major components are maximized in the following populations: ANE (green), Iranian farmer (red), Anatolian farmer (violet), and East Asian hunter gatherer (yellow). The other three are maximized in Han (orange), Mixe (cyan), and Papuan (dark blue). Supplementary figures of PCA and ADMIXTURE show all the present-day and ancient populations (*21*).

in both PCA and outgroup *f3* analysis, as represented by the 24,000-year-old Mal'ta and the 17,000-year-old Afontovogora3 (AG3) individuals, whose ancestry is also present in the Steppe Early to Middle Bronze Age (EMBA) (namely Yamnaya and Afanasievo) and western Siberian hunter gatherers (WSHGs) (Fig. 2A and figs. S1, S2, S4, and S9) (*21*).

In addition to the ANE ancestry, we further identified three other major ancestry components—East Asian hunter gatherers and Iranian and Anatolian farmers—using ADMIXTURE (K = 7, where K indicates the number of subpopulations making up the total population) models of BA Xinjiang individuals (Fig. 2B and figs. S4, S7, and S8) (*21*). To quantify the proportions of different ancestral sources, we used a qpAdm modeling analysis (*28*). Similar to the admixture results, the BA Xinjiang populations are composed of four major sources related to ANEs (27 to 91%), Anatolian (8 to 25%) and Iranian farmers (14 to 26%), and East Asians (9 to 73%) (fig. S29 and tables S3 and S4) (*21*). Additionally, models with more contemporaneous potential ancestral sources portray BA populations from northern and western Xinjiang as a mixture of four major ancestry sources derived from ancestry present in the BA Tarim Basin mummies at the Xiaohe site Xinj_BA1_TMBA1 (8 to 85%) (*16*), Afanasievo (57 to 100%), Shamanka_EBA (10 to 92%), and Gonur1_BA (BMAC) (~43%) (Fig. 3A and tables S5 and S6) (*21*). The genetic connection with Steppe EMBA populations is also validated by 10 individuals, 67% carrying Afanasievo-associated R1b1 Y-Haplogroups (table S1 and fig. S26) (*11*, *21*).

The presence of deep ANE ancestry observed in BA Xinjiang populations (Fig. 3A, table S5, and fig. S14) (*21*) may be traced to Tarim_EMBA1 ancestry, ~72% of which could be derived from the Upper Palaeolithic individual AG3 (*16*). ANE ancestry in present-day populations from South America (e.g., Surui and Karitiana), Europe, and Siberia may explain the affinity for BA Xinjiang with these populations, as shown in the PCA and outgroup-*f3* results (figs. S3, S10, and S13) (*14*, *21*). Although most of the BA Xinjiang populations showed a high affinity with each other—demonstrating a degree of homogeneity in the region at this time, as shown using outgroup-*f3* statistics (fig. S12) (*21*)—several individuals exhibited more diverse ancestries. These individuals acted as outliers in our dataset, and they were likely to have represented less-common ancestries in the sampled region or a high degree of mobility of individuals or smaller groups.

One such individual from the Songshugou site in northern Xinjiang Xinj_BA7_oEA, radiocarbon dated to 5043 to 4861 B.P., revealed an East Asian affinity in PCA and ADMIXTURE, supported by *f4*-statistics (Fig. 2 and figs. S4 and S16) (*21*). In qpAdm models, Xinj_BA7_oEA could be modeled as a two-source admixture of Xinj_BA1_TMBA1 (~8%) and Shamanka_EBA (~92%) (Fig. 3A and tables S5 and S6) (*21*), but it cannot be modeled using the Neolithic Mongolian ancestry (Mongolia_N_East) present in the Chemurchek_northAltai individuals (table S3) (*17*, *21*). Estimating the timing of the admixture event using DATES, which calculates the timing of admixture events, resulted in dates between ~6754 and 6041 B.P. (fig. S15) (*21*), which may place the Northeast Asian ancestry prevalent in the Baikal region in Xinjiang before the arrival of Steppe ancestry. Additionally, we identified an Afanasievo-related individual in the western Xinjiang Yili River Valley (Xinj_BA5_oSte) that overlaps directly with Steppe_EMBA Afanasievo and Yamnaya populations in the PCA (Fig. 2A and figs. S1, S2, and S4) (*21*) and shares more genetic drift with present-day European populations than with present-day Siberian populations (fig. S13) (*21*). In fact, we show this Steppe-related ancestry to be derived from Afanasievo and not from other western Steppe populations, as observed in *f3*-affinity tests (fig. S11) and *f4*-statistics (fig. S17) (*21*). When we estimated the admixture proportions with qpAdm, we observed that Xinj_BA5_oSte could be modeled as having an unadmixed Afanasievo ancestry (Fig. 3A and tables S5 and S6) (*21*). In addition to the outliers with East Asian and Steppe EMBA affinities, two individuals from the Chemurchek Culture of Chananguole (Xinj_BA6_aBMAC), radiocarbon dated to 4352 to 4096 B.P., showed high affinity with BMAC populations in PCA and ADMIXTURE (figs. S1 and S4) as well as in a qpAdm model requiring ~43% BMAC ancestry (tables S5 and S6) (*21*). This BMAC ancestry content is similar to that reported for two contemporaneous Chermurchek-associated individuals at the nearby site of Yagshiin Huduu in modern-day Mongolia (*17*), which demonstrates a continuity with southern Altai Chemurcheck populations and establishes BMAC ancestry in BA Xinjiang. DATES analysis estimates the Botai and BMAC admixture to have occurred ~5281 to 4575 B.P. The mean dates for the Afanasievo-related admixture of the BA Xinjiang samples are ~4877 to 4642 B.P. (fig. S15). This estimated timing of admixture would place Steppe-related people in northern and western Xinjiang at a similar time as when the Afanaseivo and Chemurchek cultures were developing near the Altai region. These results would be in line with the linguistic phylogenetic Tocharian split time of ~5000 B.P. (*29*) and may support the arrival of the Indo-European Proto-Tocharian language in Xinjiang earlier than the appearance of Indo-European languages in Europe (*28*)

After the identification of the four major ancestry components of BA Xinjiang populations (ANE, East Asian hunter gatherer, and Iranian and Anatolian farmer), we tracked the proportional changes of these ancestries over time. Steppe Middle to Late Bronze Age (MLBA) ancestry is characterized by the Andronovo and Sintashta pastoral cultures, which reached the eastern Steppe by at least ~3000 B.P., and archaeological evidence has also suggested its presence in Xinjiang (*3*, *30*). The Andronovo- and Sintashta-related ancestries differ from that of EBA herders by an increase in Anatolian farmer–related components coming from western Eurasia (*1*). The appearance of these ancestries in Xinjiang at this time would place this region within the MLBA expansion of these Steppe populations. According to PCA, ADMIXTURE, and *f3*- and *f4*-statistical results (Fig. 2 and figs. S18 and S19) (*21*), the seven LBA individuals from northern and western Xinjiang showed an increase of the Anatolian as well as Iranian farmer–related ancestries, similar to the Steppe_MLBA populations (*11*). These seven individuals are separated into four subgroups (table S1) (*21*) according to their genetic affinities. A high degree of Steppe_MLBA (81 to 100%) is further observed in the working qpAdm models, with the rest of the ancestry being derived from Xinj_BA7_oEA or Shamanka-related people (7 to 12%) and BMAC (~12%) (Fig. 3A and table S7) (*21*). Additionally, a single Xinj_LBA3 population from west Xinjiang could be modeled using Steppe_EMBA Afanasievo (~88%) and East Asian ancestry (~12%) (Fig. 3A and tables S7 and S8) (*21*), which suggests continuity between BA and MLBA populations, despite the influx of Steppe_MLBA ancestry. Notably, one individual from the west Yili Jirentaigoukou site, Xinj_LBA4, worked as a single-source qpAdm model with Sintashta, indicating greater affinity with Sintashta than Andronovo and the possible presence of both Andronovo and Sintashta ancestries in Xinjiang (tables S7 and S8) (*21*).

These results, based on a broad sampling across the region, point to BA populations in the Tarim Basin containing a deep ANE-related ancestry (*16*), with Northeast Asian ancestry similar to that from the EBA Lake Baikal region appearing at least in the north and west. We also report additional movements and admixtures along the IAMC with Afanasievo- and BMAC-related populations, validating aspects of both the Steppe and Bactrian oasis hypotheses for the settlement of Xinjiang. In the LBA, we find a continuation of existing genetic profiles with an additional influx of Steppe_MLBA ancestry.

IA admixture with nomadic Steppe cultures

The transition to the IA, which took place early in the first millennium BCE, is marked by the establishment of different nomadic groups around Xinjiang. This period also witnessed increased mobility leading to a greater connectivity between East and West Eurasian

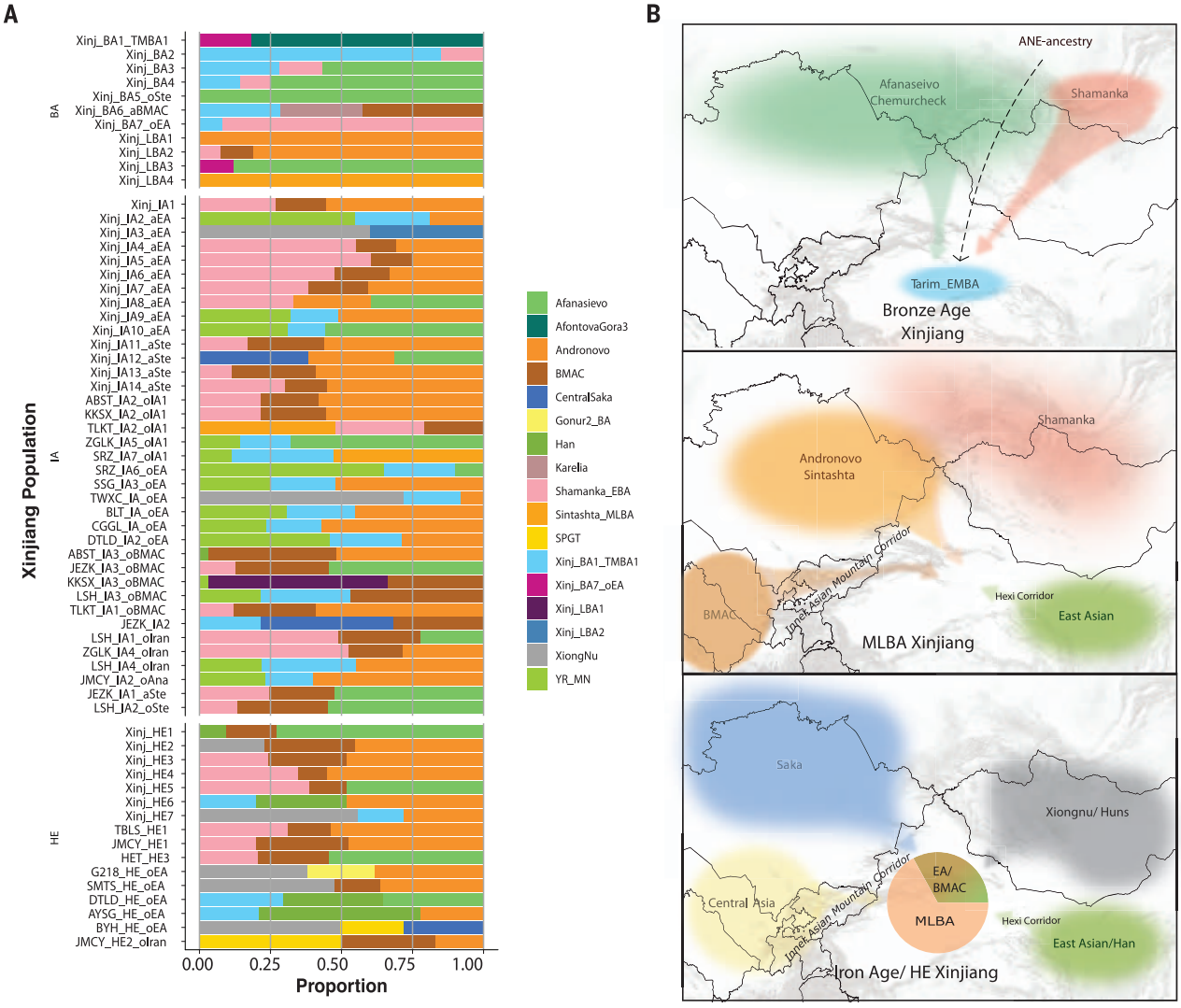


Fig. 3. Inferred qpAdm models and summary of ancient Xinjiang with population movements. (**A**) The proximal qpAdm admixture proportions for all Xinjiang populations. Each bar represents admixture proportion of the listed subgroups for BA, LBA, IA, and HE populations. Subgroup details are provided in table S1 and the supplementary text, and qpAdm modeling results are provided in detail in tables S5 to S13 (*21*). The Swat Valley Protohistoric Grave Type IA populations are in the SPGT group, and the Yellow River basin Middle Neolithic population is in the YR_MN group. (**B**) Inferred scenarios of admixtures in BA, LBA, IA, and HE Xinjiang with possible population movements shown as arrows. Xinjiang BA populations can be mostly characterized with Steppe_EMBA and Xinj_BA1_TMBA1 (Tarim Basin EMBA) ancestries with additional ancestries of Central Asia (BMAC)—as observed in Chemurcheck culture (Steppe_EMA)—and Northeast Asia (Shamanka). MLBA Xinjiang populations contain additional Andronovo Steppe, Central Asian (BMAC), and East Asian ancestries, whereas the IA and HE populations show the major Xinjiang and Steppe MLBA ancestries with additional components from BMAC and East Asian (EA) sources, shown as a pie chart summarizing the qpAdm modeling of IA and HE populations using the Xinj_LBA population (table S13 and fig. S28) (*21*). Where possible, coloring corresponds to ancestry in (A).

populations (*5*, *14*, *31*), which might have been facilitated by an increasingly widespread use of horses as well as the existence of several natural passes connecting Xinjiang to neighboring regions. The IAMC and Pamir region to the west of Xinjiang have been proposed as a BA route connecting Central Asia with the Tarim Basin in Xinjiang (*15*). Characterizing the extent of Central Asian BMAC-related ancestry in Xinjiang over time would provide information about the development of the social and cultural connections between these regions. We explored the impact of these proposed interactions on IA Xinjiang populations by analyzing 98 IA individuals throughout Xinjiang (Fig. 1 and table S1) (*21*).

We initially examined the relationships between Xinj_IA and Xinj_BA populations. In the PCA analyses, Xinj_IA individuals are not well-differentiated geographically, and they group closely with the Xinj_BA cluster lying along a west-to-east cline and surrounded by Steppe_MLBA and Steppe_EMBA population clusters (Fig. 2A and figs. S1, S2, and S5) (*21*). The IA Xinjiang populations show a coexistence of diverse ancestries in the region, having affinities for Steppe, East Asian, or Central Asian populations, and are classified into 37 subgroups (Fig. 2A, fig. S5, and table S1) (*21*). Building on trends established during the MLBA, IA Xinjiang populations can be additionally characterized by increasing proportions of Anatolian and Iranian farmer–related ancestries with an additional increase of East Asian–related ancestry evident in ADMIXTURE and qpAdm results (Fig. 2A; figs. S5, S8, and S27; and tables S9 and S10) (*21*). In outgroup-*f3* tests, most of these individuals shared the most alleles with ancient Steppe-related populations, present-day

Europeans, and Siberians (figs. S9, S10, S21, and S22) (*21*). However, several individuals showed an increased East Asian affinity in ADMIXTURE and *f3* statistics, supporting a substantial influx during the IA of East Asian ancestry into Xinjiang (figs. S5, S9, S10, S21, and S22) (*21*). To better define the genetic relationships between LBA and IA populations, we used Xinj_BA and Xinj_LBA as proximal sources in qpAdm models and observed that IA Xinjiang individuals could be modeled using Xinj_LBA1 populations with additional ancestry from Xinj_BA_TMBA1, East Asian, or BMAC and Indus periphery sources, demonstrating a genetic continuity of Xinj_LBA ancestry into the IA, with additional genetic contributions from East and Central Asia (fig. S28 and table S13) (*21*). As in Xinj_LBA, the predominant ancestry observed in the IA is derived from three sources: Steppe MLBA (~55%), BMAC (~18%), and Shamanka EBA (~27%) represented by Xinj_IA1 (43 individuals from 18 archaeological sites) (Fig. 3A, fig. S5, and table S9). Other subgroups with two or more individuals exhibit an affinity (denoted by "_a") for either East Asian (32 to 60%) or Steppe MLBA (24 to 59%) populations, whereas in some working models, BMAC ancestry is replaced by Xinj_BA1_TMBA1 ancestry (13 to 26%) (Fig. 3A; fig. S28; and tables S9, S10, and S13) (*21*). Additionally, 21 outlier individuals (denoted by "_o") with unique ancestries as shown in PCA and ADMIXTURE plots either show high affinity with one or more of the three basic sources listed above or contain unique ancestry represented by Steppe EMBA or Saka populations, as shown in qpAdm models (Fig. 3A; fig. S28; and tables S9, S10, and S13) (*21*).

Because we have identified an influx of Steppe_MLBA sources into Xinj_LBA populations, we further studied the impacts of this core Steppe ancestry on Xinj_IA populations. To differentiate between Steppe_EMBA and Steppe_MLBA affinity of IA Xinjiang populations, we used the known absence of Anatolian farmer–like ancestry in the EMBA Steppe populations. We performed this test using outgroup-*f3* statistics of the form, *f3(Anatolia_Neolithic, X; Mbuti)* and *f3 (Kostenki14, X; Mbuti)*, taking advantage of the lack of Anatolian farmer ancestry in the Upper Paleolithic Kostenki14 individual from Europe. The results supported a greater affinity of Xinj_IA to Anatolian farmers, linking the more recent Steppe_MLBA ancestry to the greater part of the Xinj_IA populations (fig. S18) (*21*). We also observed a cline of increasing Anatolian and Iranian farmer–related ancestry accompanied by a decrease in WSHG affinity in Xinj_IA compared with the Xinj_BA population using an *f4*-statistics–based approach (fig. S19) (*21*). The qpAdm models also validated an increase in the Anatolian farmer ancestry component reaching up to ~43% in some of the Xinjiang populations and an Iranian farmer component of ~38% (fig. S27 and tables S3 and S4) (*21*). The Steppe_MLBA ancestry found in the IA is further supported by the presence of the distinct Steppe_MLBA Y-haplogroup, R1a1 (n = 30), and greater affinity in outgroup-*f3* analysis to Steppe_MLBA than to BA populations (figs. S9, S21, and S28) (*21*). We identified seven IA populations with Afanasievo ancestry in qpAdm models and three individuals overall carrying the Steppe_EMBA (R1b1) haplogroups (tables S1, S9, and S10). We also found models with Afanasievo and Andronovo ancestry in four IA populations (ABST_IA2_oIA1, Xinj_IA8_aEA, Xinj_IA10_aEA, and Xinj_IA12_aSte) with additional Shamanka_EBA ancestry (P > 0.01). The DATES results suggested that the admixture event in Xinj_IA10_aEA occurred between ~4351 and 3177 B.P., a broad range that will require sampling additional IA individuals to refine (fig. S15) (*21*). This continuity of core Steppe_EMBA ancestry may lend support to the persistence of BA Indo-European languages into IA Xinjiang (*7*).

The IA also shows an increase in the frequency of BMAC ancestry in the *f4*-statistics comparisons (figs. S19 and S20) (*21*). Seven IA populations were found to contain BMAC ancestry (30 to 47%), and we observed four IA populations that could also be modeled using Indus periphery ancestry sources SPGT and two with Gonur_2BA (18 to 37%) (Fig. 3A and tables S9 and S10). The increase in the appearance of BMAC ancestry suggests a substantial movement of people from either BMAC- or Indus periphery–derived populations into the Xinjiang region during the IA (Fig. 3A), most likely through the IAMC route over the Pamir and Tianshan Mountains. Notably, we found nine IA individuals to be genetically similar to previously identified Saka populations, and they could be modeled as a single source with TianShan and Central Sakas using qpAdm (Fig. 3A and tables S9 and S10) (*21*). Together with the results discussed above, we show an increase of external ancestries appearing in Xinjiang during the IA originating from multiple Central, South, and East Asian sources, demonstrating a broadening of demographic contacts and exchanges between groups from the surrounding areas, with IA nomads, such as the Sakas and the Xiongnu, playing a substantial role. Further, although the spread of languages is not always congruent with population histories (*32*), the presence of Saka ancestry in Xinj_IA populations supports an IA introduction of the Indo-Iranian Khotanese language, which was spoken by the Saka and later attested to in this region (*19*).

East Asian ancestry increased from the BA to the IA

Archaeological discoveries of similar copper and bronze objects in the neighboring regions of Xinjiang and Gansu-Qinghai suggest an important link with East Asia and the likely introduction of metallurgy into northwest China (*33*, *34*). Along with such objects, the presence and extent of East Asian ancestry in the region is key to defining the movement and contacts between BA and IA Xinjiang populations with different parts of East Asia. Northeast Asian ancestry from the Siberian Lake Baikal region (Shamanka) is present in BA Xinjiang populations, where, apart from the major Steppe ancestry, some admixture models also require Northeast Asian sources (Fig. 3A, fig. S27, and table S3) (*21*). In Songshuguo in northern Xinjiang, ~92% of the genome of one BA individual (Xinj_BA7_oEA) could be sourced from Shamanka_EBA (Fig. 3A and table S5) (*21*). During the IA, an increase in the East Asian ancestry component is frequently observed among Xinjiang samples, and the distribution of Xinj_IA individuals on the PCA follows a west-to-east cline of increasing East Asian ancestry (Fig. 2A and figs. S1 to S3) (*21*). This increase in East Asian ancestry compared with that in the BA is also observed in ADMIXTURE analysis (Fig. 2B and fig. S5) (*21*) and is supported by outgroup *f3*-statistics (fig. S9) (*21*). The prevalence of both East Asian– and Steppe-related ancestry is further supported by the comparisons using *f4 (WSH, Shamanka_EN; Xinj_Pop, Mbuti)* and *f4 (WSH, BMAC/Indus Periphery; Xinj_Pop, Mbuti)*, with two separate groups among IA populations—one having more affinity for East Asians and the other with more affinity for Steppe-related populations (fig. S20) (*21*).

Among the ancient East Asian populations, most of the Xinjiang BA and IA populations show a higher affinity in outgroup-*f3* results with ancient Northeast Asian Neolithic populations from Siberia, such as Shamanka_EN, Lokomotiv_N, and DevilsCave_N, and eastern Steppe IA populations, such as the Xiongnu, than with the northern (Boshan) or southern (Man_Bac) East Asian populations (figs. S9 and S25) (*21*). Among the present-day East Asian populations, ancient Xinjiang populations in general show the highest affinity for Uyghur-, Austronesian (Ami)–, Hmong-Mien (She)–, and Sino-Tibetan (Han)–speaking populations rather than for Austroasiatic (Cambodian)– and Tai-Kadai (Dai)–speaking people (figs. S10 and S25) (*21*). The uniparental markers with mitochondrial haplogroups C, D, M, and N in 68 individuals from the north, south, and west, and the presence of Y-chromosomal haplogroup O3a2c in four individuals from the south and east also reflect the high degree of East Asian ancestry in Xinjiang (*35*) (table S1 and fig. S26) (*21*). QpAdm models revealed the primary source of East Asian ancestry found in BA Xinjiang to be similar to the Northeast Asian ancestry common among Neolithic Mongolia (Mongolia_N), Lake Baikal (Shamanka_EN), and Yellow River (YR_MN) populations (tables S3

and S4) (*16*). By contrast, the East Asian components of IA populations are more diverse. Although models of some populations support ancestry from either Xiongnu or Shamanka_EBA (e.g., Xinj_IA14_aSte), other IA populations work with only Xiongnu (and Han-related populations) (e.g., Xinj_IA3_aEA) or Shamanka_EBA (e.g., Xinj_IA4_aEA and Xinj_IA7_aEA) (tables S9 and S10 and fig. S28) (*21*). The appearance of Xiongnu-related ancestry in IA Xinjiang aligns with the westward expansion of the Xiongnu in ~2200 B.P., after the defeat of the Yuezhi in Gansu and the Tarim Basin (*1*). A general increase in broad East Asian ancestry beginning in the IA, in addition to the continued presence of the Northeast Asian ancestry of the BA, is observed in outgroup-*f3* statistics (fig. S25) (*21*). Although this may be the result of Han-related ancestry components previously documented in the Xiongnu (*17*), our results were unable to trace these East Asian streams to specific populations and cannot rule out the possible movement of additional East Asian groups into the region beginning in this period.

Genetic continuity of IA mixed ancestry in HE populations

Our analysis of IA Xinjiang populations revealed intense mobility and genetic exchange across the region, with growing contributions of ancestry from East and South Asia. Very few genomic data exist from the post-IA populations in the HE, but this information is crucial to trace the contributions that post-IA demographic changes had on historical populations and the evolution of modern Xinjiang genomic diversity. Political control of the region passed between the Xiongnu and the Han Dynasty until the third century CE, and in the 6th century CE, the Turkic Khaganate established control over Xinjiang, later to be driven out by the Tang Dynasty. In the ensuing centuries, Xinjiang fell under the various influences of Tibetans, Uyghurs, and Mongols (*5*). To document the effects that these events may have had on Xinjiang populations, we sequenced and analyzed 27 HE (Xinj_HE) individuals from 15 sites (table S1) (*21*). These Xinj_HE individuals clustered with the Xinj_IA individuals and similarly overlapped with Steppe nomadic groups, such as Tianshan Sakas, Huns, and Central Sakas, on a PCA (Fig. 2A and fig. S6) (*21*).

Similar to the IA individuals, the Xinj_HE populations shared the most alleles with Steppe_MLBA populations and present-day Europeans in outgroup-*f3* analysis, with outlier individuals showing more affinity for either East Asian or Central Asian populations (figs. S9, S10, S23, and S24) (*21*). Comparisons of South and Central Asian ancestry using *f4(Gonur1_BA, Gonur2_BA; Xinj_HE, Mbuti)* showed two individuals [AYSG_HE_oEA and BYH_HE_oEA ($Z > -2.3$, where Z is the number of standard deviations from zero)] with a greater affinity to Indus periphery (Gonur2_BA) than to BMAC (Gonur1_BA) ancestry (table S14) (*21*). Similar to the ancestry present in the IA, we observed three major ancestry sources of Steppe_EMBA or Steppe_MLBA (17 to 73%), BMAC (10 to 33%), and East Asian (9 to 57%), with several exceptions (Fig. 3A and tables S11 and S12) (*21*). We grouped the 27 individuals of the HE into 16 subgroups, analyzing HE individuals and populations showing unique ancestries separately. Two HE individuals were found to share a considerable amount of ancestry derived from the SPGT population—JMCY_HE2_oIran and BYH_HE_oEA from Yili and the northeastern part of Xinjiang, respectively—which may indicate a continued inflow of South and Central Asian ancestry into Xinjiang (Fig. 3A and table S11) (*21*). In addition to the Steppe MLBA influence, we found several qpAdm working models that included Steppe_EMBA or Xinj_BA1_TMBA1 ancestry as one of their components, two of them also containing the R1b1 haplogroup (tables S11 and S12) (*21*). We also identified one individual from the HE Yili region (TBLS_HE1) with a working qpAdm model having both Afanasievo or Xinj_BA1_TMBA1 (18 to 39%) and Xinj_LBA1 (Andronovo) or Sintashata_LBA (29 to 63%) ancestry sources but having high affinity to Andronovo, because the qpAdm model works with Afanasievo as an outgroup (tables S11 and S12) (*21*). We also identified seven populations with Tarim_EMBA (16 to 34%) and Andronovo ancestry (22 to 63%) (tables S11 and S12) (*21*). Taken together, these findings document examples of first-wave Steppe EMBA ancestry persisting into the HE.

To study the possible phenotypic effects of these ancestry shifts, we investigated predicted eye, hair, and skin pigmentation using the HIrisPlex-S system (*36*) on the higher-coverage individuals (n = 38 to 64; table S15) spanning all time periods. We observed the appearance of lighter hair color and skin tone in north and west Xinjiang beginning in the LBA and continuing into the IA. Blue eye alleles also appeared in these regions at least during the IA, with one to two early IA samples having blue eyes in the Yili region (figs. S29 to S31 and table S15). Light eye and hair pigmentation has previously been identified in people associated with the Steppe_MLBA Andronovo culture (*37*), with which several of these individuals are associated. Although we have phenotype results from only five HE samples from all regions, we note that all five have darker hair and eye pigmentation, corresponding to a period of increased ancestry from East, South, and Central Asia, although we interpret this with caution because of the extremely small sample size.

Notably, ancestry sources identified in the IA—Steppe, BMAC, and East Asian—are also observed in present-day Xinjiang Uyghurs (*38*), which suggests a genetic continuity from the IA to present-day Xinjiang populations. In the PCA analysis with present-day Xinjiang populations, HE individuals overlapped with people from Central Asian Tajikistan, Kazakhstan, and Uzbekistan; Turkmens; and Uyghur people from Xinjiang (fig. S3) (*21*), and individuals having more Steppe-related affinity shared greater drift with present-day Uyghurs.

Thus, despite multiple influences by several nomadic groups in post-IA Xinjiang, we find a surprising degree of genetic continuity, in that the mixed ancestry present in the IA is also present in historical Xinjiang populations, and the increased influences of Central Asian and East Asian ancestries dating to the IA are still shared between historical and present-day Xinjiang populations. We caution however, that the sample size of our historical sites is not as robust as those of the BA and IA sites. A broader analysis covering a larger geographical area could better verify the trends observed with the current datasets.

Discussion

Beginning with ancestral sources traced to indigenous ANE-derived and BA western and eastern Steppe populations, Xinjiang population structure can be characterized by waves of incoming gene flow and admixture from surrounding populations adding to the extant ancestry. The BA Xinjiang region contained four major ancestries, which included Tarim_EMBA1 (Xinj_BA1_TMBA1) (*16*), Afanasievo (Xinj_BA5_oAfan), Northeast Asian (Xinj_BA7_oEA), and BMAC (Xinj_BA6_aBMAC), with Tarim_EMBA1 possibly being indigenous to the region given its presence among diverse BA individuals (Fig. 3B). The Mongolian Chemurchek inhabitants near the Altai region were further linked to Xinjiang through the Chemurchek culture of BA northern Xinjiang, demonstrating the BA movement of people across the Altai region. Thus, we not only find support for both the Steppe and Bactrian oasis hypotheses (*5*, *13*–*15*), but the identification of additional ancestries suggests further complexity of the EBA populations in Xinjiang. Additional sampling of pre-BA and EBA populations will be necessary to further characterize the succession of ancestries established in Xinjiang during this time period.

Later, in the LBA, ancestry present in the Central Asian BMAC populations becomes more pronounced, which was likely to have entered Xinjiang though the IAMC route along with the Steppe_MLBA populations, such as the Andronovo, Sintashta, and Dali (Botai related ancestry) populations (Fig. 3B). The entrance of Steppe MLBA (~3900 B.P.) into Xinjiang correlates with the arrival of the eastern Fedorovo subculture of the Andronovo (~3750 to 3500 B.P.) from the Tianshan Mountains (*39*). The IA is marked by an increase in

movement and admixture of Steppe, Central Asian, and East Asian people into the Xinjiang region. The IA oversaw a continuation of Steppe_MLBA ancestry with greater genetic affinity to Central Asian populations containing BMAC ancestry. We also observed ancestries derived from South Asian Hunter Gatherer (Onge) in the LBA and IA, which suggests the movement of populations either from Central Asia already carrying this ancestry or from the Indus periphery region through the Pamir mountain regions into southern Xinjiang (*11*). Concurrently, an IA influx of East Asian ancestry from the eastern Steppe of present-day Mongolia is also observed, which may be tied to the westward expansion of the Pazyrk Xiongnu into Xinjiang. These admixed ancestries related to Steppe, East Asian, and Central Asian people established in the IA have been maintained since that time and are still prevalent in both the HE and present-day Xinjiang, linking the past with present-day populations. Whereas aspects of this reconstructed population history find support in the archaeological record, several insights can be gained by comparing the newly generated genomic data with previous archaeological and historical evidence.

First, although the diffusion of culture is not always accompanied by population movements (*32*), we observed an overall concordance between the two in Xinjiang populations. For example, the major genetic influences present at the earliest settlements of north and west Xinjiang populations can be related to the coexistence of people with different cultural backgrounds—e.g., Afanasievo, Chemurchek, and Okunevo (Fig. 3A and supplementary text). Also, the shift in population ancestries can be associated with proposed population movements. Specifically, the Afanasievo-related ancestry in BA individuals is consistent with the concurrent appearance of the Yamnaya culture in the Altai-Sai region, and the Steppe_MLBA ancestry in LBA individuals can be attributed to the expansion of the Steppe_MLBA culture into Xinjiang (see the supplementary text, which provides archaeological backgrounds). In the IA, genetic affinities with nomadic groups such as the Saka reveal the widespread presence of these groups across the entire Steppe region. Overall, we detect Tarim_EMBA ancestry surviving into the IA and HE, which implies the cohabitation of populations descended from Steppe EMBA, BMAC, and local Tarim_EMBA people in Xinjiang (Fig. 3B). These findings also give genomic evidence for the broad demographic processes underlying the spread of numerous languages in Xinjiang that have survived into historical times, such as the introduction of Tocharian languages by populations associated with the Afanasievo culture. An increasing mobility and movement of the Sakas in the IA and the establishments of the Saka states lasting into the HE aided in the expansion of Indo-Iranian languages, such as Khotanese, in Xinjiang as well.

Despite the widespread population movements documented here, the degree of genetic continuity that has been maintained in Xinjiang over the past 5000 years is noteworthy. Although genetic continuity has been observed in isolated environments or regions with relatively high cultural homogeneity, such as Northeast Asia (*40*), dynamic interactions between populations with diverse ancestries and cultures are more likely to result in major population shifts and turnover, such as in the Oceania archipelago and Europe (*28*, *41*). However, this has not been the case for Xinjiang populations, where at least two different instances of genetic continuity are observed. The first is the genetic continuity (Steppe ancestry) from BA individuals to LBA and IA individuals, which represents a case in which a core Steppe ancestry has been maintained despite the addition of an extensive influx of diverse ancestries. The second case is the stability of Xinjiang population diversity from the HE to the present day, despite the turmoil of successive external ruling powers over the past 2000 years. A major reason may be that this mixed ancestry was prevalent not just in Xinjiang but throughout Central Asia, so dynamic population movements would not result in major genetic shifts. These findings indicate that genetic and archaeological evidence can provide distinct yet complementary insights into population history. This, in turn, further emphasizes the importance of a multidisciplinary approach to uncover the complex histories of regions like Xinjiang, where persistent interactions between various populations and cultures occurred.

REFERENCES AND NOTES

1. P. B. Damgaard *et al*., *Nature* **557**, 369–374 (2018).
2. A. Betts, P. Jia, I. Abuduresule, *Archaeol. Res. Asia* **17**, 204–213 (2019).
3. A. V. G. Betts, M. Vicziany, P. Jia, A. A. Di Castro, *The Cultures of Ancient Xinjiang, Western China: Crossroads of the Silk Roads* (Archaeopress, 2020).
4. Y. Li, *Veget. Hist. Archaeobot.* **30**, 287–303 (2021).
5. J. A. Millward, *Eurasian Crossroads: A History of Xinjiang* (Columbia Univ. Press, 2007).
6. K. Li *et al*., *Sci. China Earth Sci.* **61**, 1659–1668 (2018).
7. M. Peyrot, *Indo-Euro. Linguist.* **7**, 72–121 (2019).
8. Y. Zhang, *Archaeology* **6**, 3–13 (2002).
9. J. Yu, J. Ma, *Cult. Relics* **3**, 53–57 (2013).
10. P. de Barros Damgaard *et al*., *Science* **360**, eaar7711 (2018).
11. V. M. Narasimhan *et al*., *Science* **365**, eaat7487 (2019).
12. X. Zhou *et al*., *Nat. Plants* **6**, 78–87 (2020).
13. C. Li *et al*., *BMC Genet.* **16**, 78 (2015).
14. W. Wang *et al*., *Sci. Adv.* **7**, eabd6690 (2021).
15. B. E. Hemphill, J. P. Mallory, *Am. J. Phys. Anthropol.* **124**, 199–222 (2004).
16. F. Zhang *et al*., *Nature* **599**, 256–261 (2021).
17. C. Jeong *et al*., *Cell* **183**, 890–904.e29 (2020).
18. C. Jeong *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7485–7490 (2016).
19. H. W. Bailey, *Iran* **8**, 65–72 (1970).
20. C. Ning *et al*., *Curr. Biol.* **29**, 2526–2532.e4 (2019).
21. See the supplementary materials.
22. A. W. Briggs, P. Heyn, *Methods Mol. Biol.* **840**, 143–154 (2012).
23. M.-T. Gansauge, M. Meyer, *Nat. Protoc.* **8**, 737–748 (2013).
24. Q. Fu *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2223–2227 (2013).
25. Q. Fu *et al*., *Nature* **524**, 216–219 (2015).
26. D. H. Alexander, J. Novembre, K. Lange, *Genome Res.* **19**, 1655–1664 (2009).
27. N. Patterson *et al*., *Genetics* **192**, 1065–1093 (2012).
28. W. Haak *et al*., *Nature* **522**, 207–211 (2015).
29. A. S. Kassian *et al*., *Linguistics* **59**, 949–979 (2021).
30. D. Cong, P. Jia, B. Ailisen, X. Jia, D. Baola, *West. Reg. Stud.* **4**, 15–27 (2017).
31. J. Han, *Cultures in Xinjiang from the Bronze Age to the early Iron Age* [in Chinese] (Wenwu Press, 2007).
32. C. Posth *et al*., *Nat. Ecol. Evol.* **2**, 731–740 (2018).
33. J. Mei, *A Metallurgical Study of Early Copper and Bronze Artefacts from Xinjiang, China* (Archaeopress, 2000).
34. L. Wang *et al*., *Archaeol. Anthropol. Sci.* **11**, 2407–2417 (2019).
35. J. Li *et al*., *BMC Evol. Biol.* **17**, 239 (2017).
36. L. Chaitanya *et al*., *Forensic Sci. Int. Genet.* **35**, 123–135 (2018).
37. C. Keyser *et al*., *Hum. Genet.* **126**, 395–410 (2009).
38. Q. Feng *et al*., *Mol. Biol. Evol.* **34**, 2572–2582 (2017).
39. P. Jia, A. Betts, D. Cong, X. Jia, P. Dupuy, *Antiquity* **91**, 621–639 (2017).
40. X. Mao *et al*., *Cell* **184**, 3256–3266.e13 (2021).
41. M. Lipson *et al*., *Curr. Biol.* **28**, 1157–1165.e7 (2018).
42. Y. Wang *et al*., *Genom. Proteom. Bioinform.* **15**, 14–18 (2017).
43. BIG Data Center Members, *Nucleic Acids Res.* **46**, D14–D20 (2018).

ACKNOWLEDGMENTS

We would like to thank archaeological teams from the Xinjiang and Chinese Academy of Social Science. We would also like to thank the reviewers, including V. M. Narasimhan, for their in-depth reviews and suggestions during the manuscript's review process. **Funding:** This work was supported by the National Key R&D Program of China (2016YFE0203700), the Chinese Academy of Sciences (CAS) and the Ministry of Finance of the People's Republic of China (YSBR-019 and XDB26000000), the National Natural Science Foundation of China (41925009), "Research on the roots of Chinese civilization" of Zhengzhou University (XKZDJC202006), the Tencent Foundation through the XPLORER PRIZE, and the Howard Hughes Medical Institute (grant no. 55008731). Xia.W. was supported by the Key National Social Science Foundation of China (no. 16ZDA144). **Author contributions:** Q.F. designed and supervised the research project. J.Z., Y.W., X.H., Q.R., J.Y., Xia.W., W.G., B.W., A.N., E.L., Z.T., Xin.W., M.Z., L.Z., Y.L., W.W., and W.H. assembled archaeological materials and performed dating. Q.F., P.C., R.Y., F.L., X.F., Q.D., and W.P., performed or supervised wet laboratory work. Q.F. and X.F. did preliminary data analysis. V.K. did the primary data analysis for this manuscript. E.A.B. did additional data analysis. V.K., E.A.B., Y.L., and Q.F. wrote the manuscript. All authors discussed, critically revised, and approved the final version of the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The aligned sequences and genotype calls for the newly sequenced individuals are available through the Genome Sequence Archive (*42*) in the BIG Data Center (*43*) (https://bigd.big.ac.cn/gsa-human; accession no. PRJCA005576). Mitochondrial DNA (fasta format) have been deposited in the Genome Warehouse in the National Genomics Data Center (https://bigd.big.ac.cn/gwh/; accession no. PRJCA005576) or published in (*14*) (https://bigd.big.ac.cn/gwh/; accession no. PRJCA002948). All software used are freely available online.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abk1534
Materials and Methods
Supplementary Text
Figs. S1 to S31
Tables S1 to S16
References (*44*–*113*)
MDAR Reproducibility Checklist

25 June 2021; resubmitted 3 September 2021
Accepted 3 March 2022
10.1126/science.abk1534

REPORTS

ECOLOGY

A keystone gene underlies the persistence of an experimental food web

Matthew A. Barbour[1]*, Daniel J. Kliebenstein[2], Jordi Bascompte[1]

Genes encode information that determines an organism's fitness. Yet we know little about whether genes of one species influence the persistence of interacting species in an ecological community. Here, we experimentally tested the effect of three plant defense genes on the persistence of an insect food web and found that a single allele at a single gene promoted coexistence by increasing plant growth rate, which in turn increased the intrinsic growth rates of species across multiple trophic levels. Our discovery of a "keystone gene" illustrates the need to bridge between biological scales, from genes to ecosystems, to understand community persistence.

Genes and their alleles underlie population persistence (*1*), yet their impact on ecological communities remains unclear. Variation among individual genotypes can scale up to influence the diversity and composition of associated communities (*2*–*6*); however, the causal genes and specific alleles underlying these indirect effects remain elusive. Tracing the ecological impact of specific genetic variants has the potential to transform how we act to conserve genetic and species diversity in a changing world (*7*).

To study the genetic basis of community-level effects, we used an experimental food web consisting of a plant (*Arabidopsis thaliana*), two species of aphids (*Brevicoryne brassicae* and *Lipaphis erysimi*), and a parasitoid wasp (*Diaeretiella rapae*) (*8*) (Fig. 1A). These species form a naturally occurring food web (*9*, *10*), and their interactions are partly mediated by a group of specialized metabolites called aliphatic glucosinolates (*9*). Extensive knowledge of genotype-to-phenotype causality in *Arabidopsis* aliphatic glucosinolates facilitates testing of how genetic change influences ecological interactions in a food web (Fig. 1, B and C). Specifically, loss-of-function mutations at three genes (*MAM1*, *AOP2*, and *GSOH*) control most of the natural variation in the aliphatic glucosinolate phenotype (*11*–*15*). These three genes function in a sequential pathway, such that the presence of a functional allele at a given gene allows biosynthesis to proceed to the subsequent chemical compound, whereas a loss-of-function mutation (hereafter, null allele) diverts biosynthesis, resulting in the accumulation of the preceding compound. Geographic variation at *MAM1* is correlated with the relative abundance of aphids in our food web, with evidence that *L. erysimi* selects for an increased frequency of the null allele, whereas *B. brassicae* selects for the functional allele (*9*). Aphids do not appear to impose selection on *AOP2* (*9*); however, natural variation at *AOP2* has pleiotropic effects on plant growth, phenology, jasmonic acid signaling, and circadian rhythms (*16*, *17*), which could influence food-web interactions. Although its role in aphid herbivory is unclear, *GSOH* contributes to adaptive variation in aliphatic glucosinolates of *Arabidopsis* in its native range (*18*).

To mimic natural variation at these three genes, we used three transgenic lines (gsm1, AOP2, and AOP2/gsoh) that were derived from a common genetic background (Col-0 genotype) in *Arabidopsis* (*8*). These transgenic lines are analogous to the evolutionary processes that determine natural variation in aliphatic glucosinolates and also mimic the chemical phenotype of diverse natural accessions that co-occur throughout Europe (*19*–*21*). We created experimental plant populations from different combinations of *Arabidopsis* genotypes (n = 11 combinations across 60 experimental food webs), including monocultures (n = 4), two-genotype mixtures (n = 6), and the four-genotype mixture (n = 1). This experimental

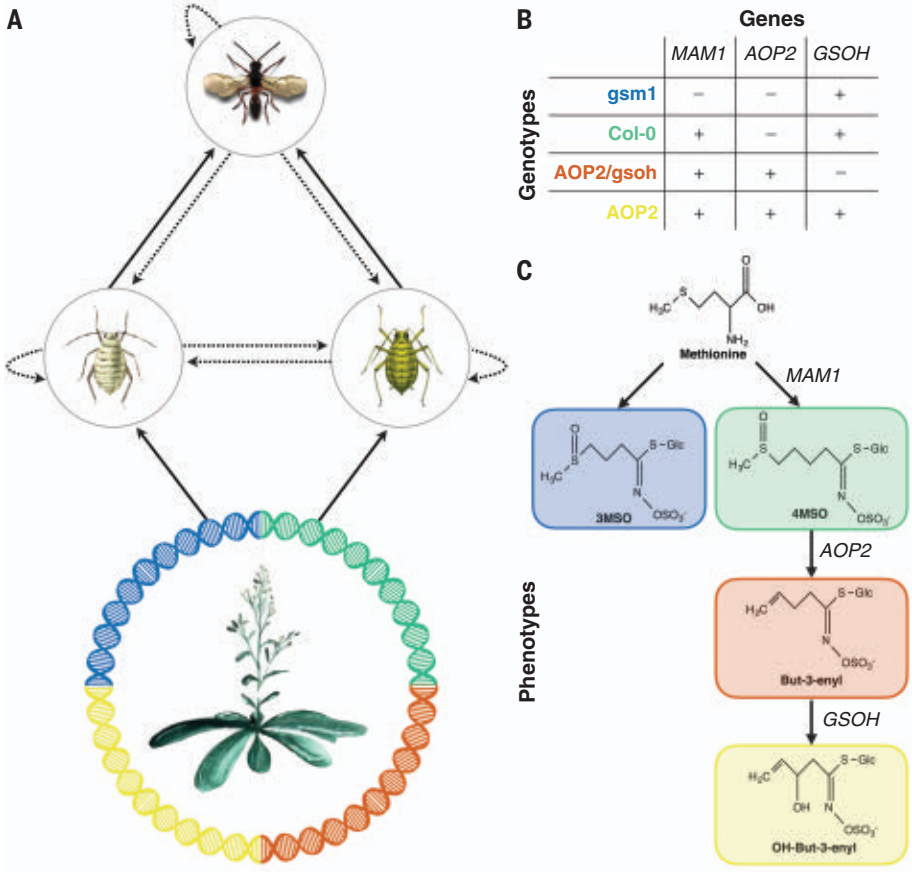


Fig. 1. Study system. (**A**) The wasp *Diaeretiella rapae* (top) parasitizes the aphids *Brevicoryne brassicae* (left) and *Lipaphis erysimi* (right), and these aphids compete for their shared resource *Arabidopsis thaliana* (bottom). Solid and dashed arrows represent positive and negative effects, respectively. (**B**) We used four *Arabidopsis* genotypes (gsm1, Col-0, AOP2/gsoh, and AOP2) that recreate natural variation in null (–) and functional (+) alleles at three genes (*MAM1*, *AOP2*, and *GSOH*) that control the biosynthesis of aliphatic glucosinolates. (**C**) The chemical phenotype (3MSO, 4MSO, But-3-enyl, or OH-But-3-enyl) of each *Arabidopsis* genotype (color-coded boxes) depends on which genes have null and functional alleles.

[1]Department of Evolutionary Biology and Environmental Studies, University of Zurich, 8057 Zurich, Switzerland. [2]Department of Plant Science, University of California Davis, Davis, CA 95616, USA.
*Corresponding author. Email: matthew.barbour@ieu.uzh.ch

design allowed us to both test a general effect of genetic diversity and isolate the effect of different alleles at specific genes (*8*). We conducted the experiment under two different temperature regimes (20°C and 23°C), which reflect the warming this food web is expected to experience over the next 25 to 50 years (*8*, *22*). After adding the experimental food web to each plant population, plants were visually inspected every 7 days, and we counted the number of each aphid species, parasitized aphids (mummies), and adult parasitoids for 17 weeks, allowing for multiple generations of aphids (~16 generations) and the parasitoid (~8 generations). Mummified aphids contain a larval-stage parasitoid, so we combined counts of mummified aphids and adult parasitoids to estimate parasitoid abundance (*8*).

We first tested for genetic effects on food-web persistence—the overall rate of species extinctions—using a Cox proportional-hazards survival analysis (*8*) (figs. S1 to S3 and table S1). We found that, on average, adding another genotype to the plant population reduced the extinction rate by 19% {extinction rate ratio [95% confidence interval (CI)] = 0.81 [0.66–0.99], Wald $Z = -2.11$, $P = 0.035$}. We also observed that specific genes varied in their contribution to food-web persistence (Fig. 2). We found that allelic differences at *AOP2* had a disproportionate effect on extinction rate, whereas different alleles at *MAM1* and *GSOH* had no clear effects (Fig. 2 and table S1). In particular, replacing the average genotype with one that had a null *AOP2*– allele reduced the extinction rate by 16% (Fig. 2, Wald $Z = -2.32$, $P = 0.021$). This reduction is larger (29%) if we make the comparison relative to the functional *AOP2*+ allele, which better characterizes the ecological impact of this loss-of-function mutation. Analogous to a keystone species that determines species diversity in a food web (*23*), our results indicate that *AOP2* functions as a keystone gene in this food web through its disproportionate contribution to maintaining species diversity (*24*). Indeed, we no longer observed an effect of genetic diversity if we first account for variation explained by genotypes with a null *AOP2*– allele (extinction rate ratio [95% CI] = 0.96 [0.74–1.26], Wald $Z = -0.27$, $P = 0.789$). *AOP2*'s keystone role was also robust to experimental warming (extinction rate ratio [95% CI] = 0.98 [0.73–1.30]; Wald $Z = -0.17$, $P = 0.865$).

To understand how variation at *AOP2* determined food-web persistence, we analyzed its role in determining transitions in food-web structure using a multistate Markov model (*8*) (fig. S4 and table S2). Figure 3 provides a map of the food-web transitions that we observed over the course of this experiment. Overall, we found that the initial food web (yellow in Fig. 3) underwent a transition where either

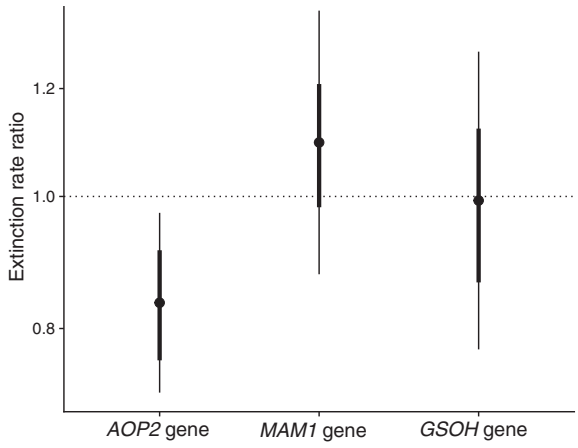


Fig. 2. Allelic differences at *AOP2* disproportionately affect food-web persistence. The dashed line represents the effect of the average allele on the overall extinction rate after controlling for genetic diversity. We arbitrarily chose the functional (+) allele of each gene as the baseline; therefore, points represent the estimated effect of the null (−) allele at each gene, with thick (SEM) and thin lines (95% CI) indicating uncertainty. Extinction rate ratio is plotted on a logarithmic scale to make positive and negative values symmetric.

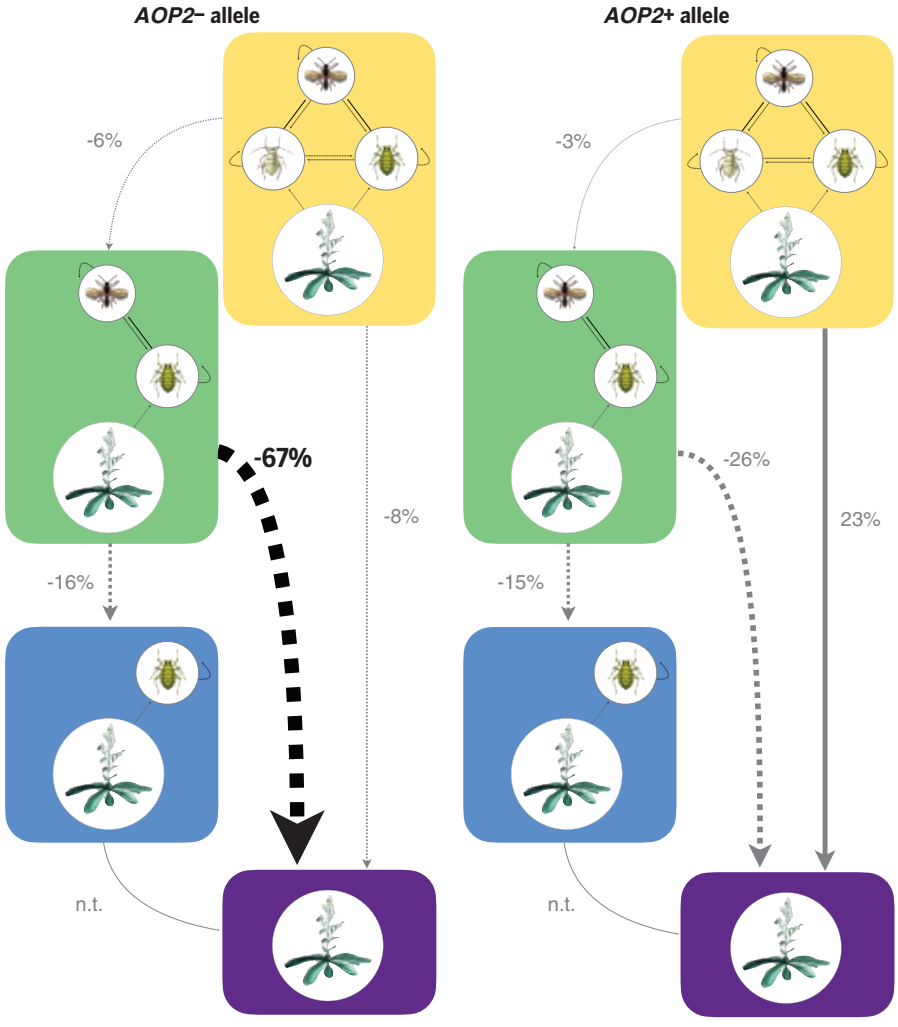


Fig. 3. *AOP2*– allele maintains species diversity by preventing the transition of the three-species food chain to the *Arabidopsis*-only state. Each color corresponds to a different food-web structure, and arrows between two colors indicate a food-web transition. Arrow thickness is proportional to the percentage change in weekly risk of a transition (numbers) when either a genotype with a null *AOP2*– allele (left) or functional *AOP2*+ allele (right) is added to the plant population. Solid and dashed arrows indicate positive and negative changes, respectively; black and gray arrows denote clear ($P < 0.05$) and unclear effects, respectively. Rare transitions were not tested (n.t.).

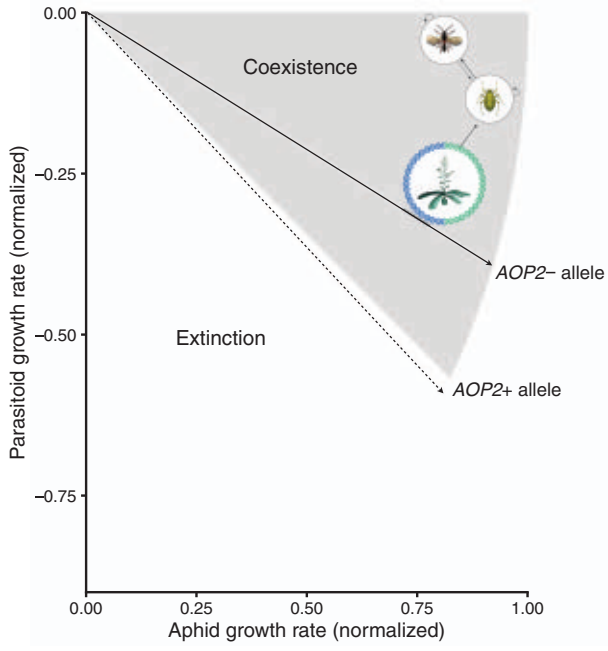


Fig. 4. *AOP2*− allele fosters coexistence by increasing the aphid and parasitoid's intrinsic growth rates. The gray area represents the range of intrinsic growth rates (normalized to length 1) where the aphid and parasitoid coexist but the parasitoid goes extinct beyond the lower boundary. Arrows represent the composite vector of aphid and parasitoid growth rates in two-genotype mixtures with either *AOP2*− alleles (Col-0 and gsm1, solid arrow) or *AOP2*+ alleles (AOP2 and AOP2/gsoh, dashed arrow) (*8*).

all insects went extinct (10 out of 60 cages; yellow to purple in Fig. 3) or, more commonly, it transitioned into a three-species food chain with the aphid *L. erysimi* (50 out of 60, yellow to green in Fig. 3). This remaining food chain either persisted for the duration of the experiment (8 out of 50), or transitioned into either an aphid-only food web (25 out of 50, green to blue in Fig. 3) or *Arabidopsis*-only state (17 out of 50, green to purple in Fig. 3). The keystone role of *AOP2* acted primarily to prevent the transition from the three-species food chain to the *Arabidopsis*-only state (green to purple), with a null *AOP2*− allele reducing the weekly risk of this food-chain transition by 67% (Fig. 3, left; hazard ratio [95% CI] = 0.33 [0.15–0.73]).

To decipher the specific mechanism by which allelic differences at *AOP2* altered the risk of the food chain collapsing, we applied theory on the structural stability of food webs (*8*). Following recent work, we say that a food web is structurally stable for as long as it remains both dynamically stable and fully feasible (*25*, *26*). The geometry of intra- and interspecific interactions defines the boundary of the food web's feasibility domain, whereas species' population growth rates at low density (hereafter, intrinsic growth rates) determine the proximity of the food web to these boundaries (*26*). At the boundary of this feasibility domain, at least one species will go extinct, thus transitioning to a simpler community. Therefore, the distance—or, to be precise, the normalized angle—from an extinction boundary is a measure of the structural stability of the food web. We applied a Bayesian multivariate autoregressive [MAR(1)] model to our time series of species' abundances to quantify *AOP2*'s effect on interactions and species' intrinsic growth rates, and thus the risk of the food chain collapsing (*8*) (figs. S5 to S9 and tables S3 and S4).

We found that the null *AOP2*− allele reduced the risk of extinctions in the food chain in a specific way (Fig. 4). This reduced risk was not because this allele altered the location of the extinction boundary (lower border of gray area in Fig. 4). Rather, it was because the *AOP2*− allele moved the vector of species' intrinsic growth rates (solid arrow) upward and into the region where the parasitoid and aphid coexist (Fig. 4 and fig. S8, Δ normalized angle [95% CI] = 13.0° [3.9–23.5]), a result that holds under nonequilibrium conditions (fig. S9). This buffering effect was primarily due to a concordant increase in the aphid's ($\Delta r_{L.\ erysimi}$ [95% CI] = 0.54 [−0.04–1.13]) and parasitoid's ($\Delta r_{D.\ rapae}$ [95% CI] = 0.66 [0.09–1.25]) intrinsic growth rates. These increases in intrinsic growth rates were likely due to a bottom-up effect of the faster relative growth rate of *Arabidopsis* genotypes with a null *AOP2*− allele (*8*) (linear mixed-effect model analysis of variance: $F_{1,5.79}$ = 15.75, P = 0.008; fig. S10 and table S5), as has been observed in other plant-herbivore-parasitoid food chains (*27*–*29*).

Our results show that variation at a single gene can control the persistence of directly and indirectly interacting species in a food web. Given that food chains of plants, insect herbivores, and their parasitoids comprise ~40% of all described Eukaryotes (*30*), this mechanism may have far-reaching consequences for the persistence and functioning of terrestrial ecosystems. These consequences may be context dependent though, as theory predicts that increased intrinsic growth rates may actually destabilize food chains in more productive environments (*31*). Field-based experiments are ultimately needed to test the relevance of our results. Still, our work highlights the potential promise of integrating tools from genetics and ecological networks to predict the consequences of genetic change for the persistence of biodiversity across scales. In particular, our study suggests that the current loss of genetic diversity (*32*) may have cascading effects that cause abrupt and catastrophic shifts in food-web structure and function.

REFERENCES AND NOTES

1. R. D. H. Barrett *et al*., *Science* **363**, 499–504 (2019).
2. R. S. Fritz, P. W. Price, *Ecology* **69**, 845–856 (1988).
3. G. M. Crutsinger *et al*., *Science* **313**, 966–968 (2006).
4. T. E. Farkas, T. Mononen, A. A. Comeault, I. Hanski, P. Nosil, *Curr. Biol.* **23**, 1835–1843 (2013).
5. T. G. Whitham *et al*., *Nat. Rev. Genet.* **7**, 510–523 (2006).
6. H. L. Barker *et al*., *Mol. Ecol.* **28**, 4404–4421 (2019).
7. M. Stange, R. D. H. Barrett, A. P. Hendry, *Nat. Rev. Genet.* **22**, 89–105 (2021).
8. Materials and methods provided as supplementary materials.
9. T. Züst *et al*., *Science* **338**, 116–119 (2012).
10. Y. Sato, R. Shimizu-Inatsugi, M. Yamazaki, K. K. Shimizu, A. J. Nagano, *BMC Plant Biol.* **19**, 163 (2019).
11. J. Kroymann, S. Donnerhacke, D. Schnabelrauch, T. Mitchell-Olds, *Proc. Natl. Acad. Sci. U.S.A.* **100** (suppl. 2), 14587–14592 (2003).
12. D. J. Kliebenstein, V. M. Lambrix, M. Reichelt, J. Gershenzon, T. Mitchell-Olds, *Plant Cell* **13**, 681–693 (2001).
13. B. G. Hansen *et al*., *Plant Physiol.* **148**, 2096–2108 (2008).
14. D. J. Kliebenstein, N. I. Cacho, *Advances in Botanical Research* (Academic Press, 2016), vol. 80, chap. 3.
15. Complete mechanistic details are described in (*11*–*14*).
16. R. E. Kerwin *et al*., *Plant Cell* **23**, 471–485 (2011).
17. M. Burow *et al*., *Mol. Plant* **8**, 1201–1212 (2015).
18. B. Brachi *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4032–4037 (2015).
19. E. K. F. Chan, H. C. Rowe, J. A. Corwin, B. Joseph, D. J. Kliebenstein, *PLOS Biol.* **9**, e1001125 (2011).
20. E. Katz *et al*., *eLife* **10**, e67784 (2021).
21. For example, gsm1 ~ Oy-0 and Wa-1, AOP2 ~ NFA-10 and Sq-1, and AOP2/gsoh ~ T670 and IP-Mot-0 (*19*). Relative frequencies of functional and null alleles in European populations: *MAM1* (+35%, −65%), *AOP2* (+75%, −10%), and *GSOH* (+27%, −73%) (*20*).
22. E. M. Zubler *et al*., *Clim. Change* **125**, 237–252 (2014).
23. R. T. Paine, *Am. Nat.* **103**, 91–93 (1969).
24. L. H. Skovmand *et al*., *Trends Ecol. Evol.* **33**, 689–700 (2018).
25. U. Bastolla *et al*., *Nature* **458**, 1018–1020 (2009).
26. R. P. Rohr, S. Saavedra, J. Bascompte, *Science* **345**, 1253497 (2014).
27. T. Züst, B. Joseph, K. K. Shimizu, D. J. Kliebenstein, L. A. Turnbull, *Proc. Biol. Sci.* **278**, 2598–2603 (2011).
28. A. Karami, Y. Fathipour, A. A. Talebi, G. V. P. Reddy, *Arthropod-Plant Interact.* **12**, 291–301 (2018).
29. C. S. Awmack, S. R. Leather, *Annu. Rev. Entomol.* **47**, 817–844 (2002).
30. D. A. Grimaldi, M. S. Engel, *Evolution of the Insects* (Cambridge Univ. Press, 2005).
31. K. S. McCann, *Food Webs* (Princeton Univ. Press, 2011).
32. D. M. Leigh, A. P. Hendry, E. Vázquez-Domínguez, V. L. Friesen, *Evol. Appl.* **12**, 1505–1512 (2019).
33. M. A. Barbour, mabarbour/genes-to-foodweb-stability: Keystone Gene (v3.0), *Zenodo* (2022); https://doi.org/10.5281/zenodo.5901259.

ACKNOWLEDGMENTS

We thank D. Trujillo-Villegas and X. Muenger for help conducting the experiment and M. Furler for greenhouse support. C. Vorburger and P. Rodriguez provided guidance on maintaining insect colonies. B. Schmid gave valuable feedback on the experimental design. A. Ferrera gave insight into quantifying the proximity of extinction boundaries. We also thank K. Shimizu and Y. Sato for valuable discussions and University of Zurich's URPP Global Change and Biodiversity program for supporting this work. This manuscript was improved by three anonymous reviewers. **Funding:** Swiss National Science Foundation grants 31003A_160671 and 310030_197201 to J.B. **Author contributions:** Conceptualization: M.A.B., J.B.;

Methodology, investigation, formal analysis, visualization, and data curation: M.A.B. Resources: M.A.B. (insects), D.J.K. (plant genotypes); Funding acquisition: J.B.; Writing – original draft: M.A.B.; Writing – review & editing: M.A.B., D.J.K., J.B. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data and code to reproduce the reported results are publicly available on GitHub (https://mabarbour.github.io/genes-to-foodweb-stability/) and have been archived on Zenodo (*33*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abf2232
Materials and Methods
Figs. S1 to S10
Tables S1 to S5
References (*34*–*67*)
MDAR Reproducibility Checklist

12 October 2020; accepted 9 February 2022
10.1126/science.abf2232

SOLAR CELLS

Damp heat–stable perovskite solar cells with tailored-dimensionality 2D/3D heterojunctions

Randi Azmi[1], Esma Ugur[1], Akmaral Seitkhan[1], Faisal Aljamaan[1], Anand S. Subbiah[1], Jiang Liu[1], George T. Harrison[1], Mohamad I. Nugraha[1], Mathan K. Eswaran[1], Maxime Babics[1], Yuan Chen[2], Fuzong Xu[1], Thomas G. Allen[1], Atteq ur Rehman[1], Chien-Lung Wang[2], Thomas D. Anthopoulos[1], Udo Schwingenschlögl[1], Michele De Bastiani[1]†, Erkan Aydin[1], Stefaan De Wolf[1]*

If perovskite solar cells (PSCs) with high power conversion efficiencies (PCEs) are to be commercialized, they must achieve long-term stability, which is usually assessed with accelerated degradation tests. One of the persistent obstacles for PSCs has been successfully passing the damp-heat test (85°C and 85% relative humidity), which is the standard for verifying the stability of commercial photovoltaic (PV) modules. We fabricated damp heat–stable PSCs by tailoring the dimensional fragments of two-dimensional perovskite layers formed at room temperature with oleylammonium iodide molecules; these layers passivate the perovskite surface at the electron-selective contact. The resulting inverted PSCs deliver a 24.3% PCE and retain >95% of their initial value after >1000 hours at damp-heat test conditions, thereby meeting one of the critical industrial stability standards for PV modules.

Commercialization of any photovoltaic (PV) technology requires a guaranteed product lifetime of at least 25 to 30 years, as is common for conventional crystalline silicon (c-Si) PV modules. Lifetime predictions of PV technologies are usually accomplished through standardized accelerated degradation tests. After the demonstration of excellent power conversion efficiencies (PCEs) of perovskite solar cells (PSCs), the main challenge toward market entry of PSCs is successfully passing standard industrial lifetime assessment tests of the International Electrotechnical Commission (IEC 61215:2016)—in particular, damp-heat testing at 85°C and 85% relative humidity (*1*, *2*). A stabilized PCE performing like a commercial c-Si solar cell (PCE ~20%) would need to pass a damp-heat test for >1000 hours with <5% relative loss in PCE (*3*, *4*).

Degradation of encapsulated PSCs is usually caused by leakage in the packaging (allowing atmospheric agents to interact with the perovskite) and device-related material instabilities. To address this, we developed leakage-free device packaging that seals the PSC within two glass sheets, using a vacuum-laminated encapsulant and edge sealing via butyl rubber. Despite this, damp-heat testing of our encapsulated control devices resulted in fast degradation (see below), implying an intrinsic thermal instability of the perovskite absorber layer itself.

The instability of three-dimensional (3D) perovskite films, when used as the photoactive absorber layer in PSCs, is mainly attributed to high defect densities as well as ion migration at grain boundaries and interfaces, which is exacerbated at higher operational temperatures (*5*–*9*). Several approaches have been reported to passivate these defects (*7*, *10*–*13*). Specifically, growing 2D perovskite layers on the top surface of 3D perovskites creates a 2D/3D perovskite heterojunction that can effectively passivate surface defects and suppress ion migration (*3*, *6*, *14*–*18*).

At the device level, integrating such 2D perovskite passivation layers in PSCs can enhance their PCE and lifetime (*3*, *4*, *14*–*17*). So far, this strategy has been most successful for "regular structured" PSCs in which phase-pure 2D perovskites (n = 1, where n represents the dimensionality of the 2D perovskite by counting the number of its octahedral inorganic sheets) were inserted between the 3D perovskite surface and the (opaque) hole-selective top-contact stack (*3*, *14*–*17*). For "inverted" devices, this top-contact passivation approach (now at the electron-selective side) has consistently failed in PCE and lifetime; this represents a persistent challenge in the perovskite community (*19*, *20*), as inverted PSCs are arguably easier to fabricate and scale up (*11*).

We found that tailoring the dimensionality (n) of the 2D perovskite fragments at the electron-selective interface of inverted PSCs is essential to enable efficient top-contact passivation through 2D perovskite passivation layers. This interface has frequently been ignored because it is assumed that the conventional electron-selective layer, C_{60} (or its derivatives), provides sufficient passivation of 3D perovskites (*21*); instead, attention has predominantly focused on the hole-selective interface of inverted PSCs, situated at the (transparent) bottom contact of the device (*22*–*25*). However, recent reports have revealed that C_{60} is only weakly bonded to perovskite layers, which induces a high energetic disorder between perovskite and C_{60} layers that limits device performance at elevated operating temperatures (*5*, *6*, *26*). Moreover, a thin layer of C_{60} is insufficient to effectively protect the 3D perovskite film underneath from moisture or oxygen ingress. Implementing 2D perovskite passivation layers is a promising approach to solve all of the issues mentioned above.

We post-treated the surface defects of the 3D perovskites by applying oleylammonium iodide (OLAI) molecules to form Ruddlesden-Popper–phase 2D perovskite layers, which resulted in higher PCEs and prolonged stabilities of inverted PSCs. We tailored the dimensionality, n, of the 2D-perovskite fragments (which also dictates their optical and electronic properties) by tuning the annealing conditions from lower to higher temperatures, in that higher-n layers have a lower formation energy (*27*). Indeed, all 2D perovskite passivation layers prepared through thermal annealing (2D-TA) showed a dominant emission peak at ~510 nm (as evidenced in fig. S1), which belongs to n = 1, in accordance with previous studies (*4*, *6*, *19*, *28*). However, the formation of higher-dimensionality 2D perovskite layers ($n \geq 2$) became more pronounced when the post-treatment was performed at room temperature (2D-RT) with the OLAI molecule (Fig. 1A).

We investigated the formation of 2D perovskite passivation films on 3D perovskites with grazing-incidence wide-angle x-ray scattering (GIWAXS; Fig. 1, B and C). The 2D perovskite passivation films exhibited diffraction q_z peaks at ~0.2 to ~0.5 A^{-1}, corresponding to the (001)

[1]Physical Science and Engineering Division, KAUST Solar Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. [2]Department of Applied Chemistry, National Yang Ming Chiao Tung University, Hsin-Chu, Taiwan.
***Corresponding author. Email: stefaan.dewolf@kaust.edu.sa**
†Present address: Department of Chemistry & INSTM, Universitá di Pavia, 27100 Pavia, Italy.

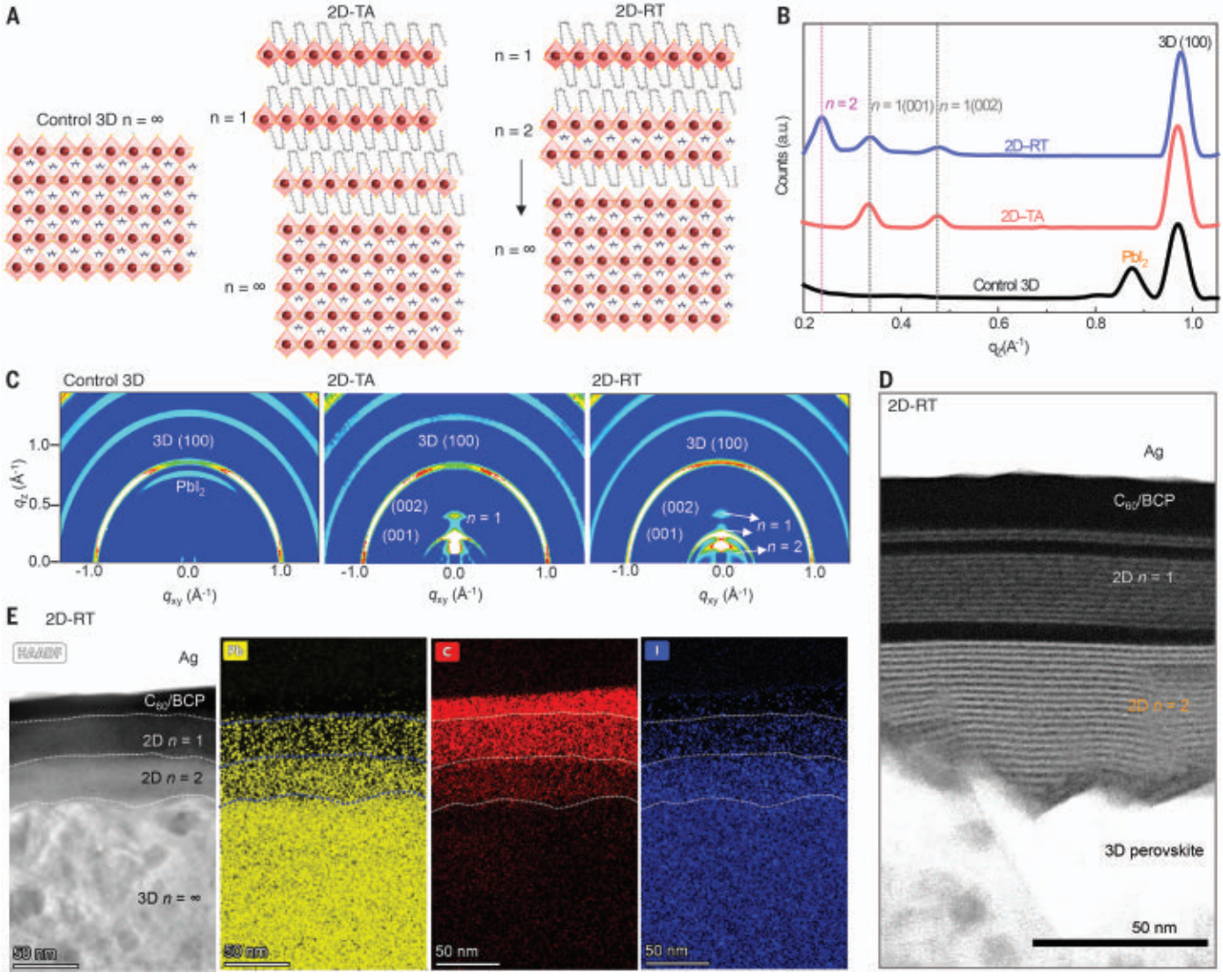


Fig. 1. Structure of 2D perovskite dimensionality control by tuning the annealing conditions. (**A**) Schematic illustration of 2D perovskite passivation with different *n* layers under thermal annealing at 100°C (TA) and room-temperature process (RT). (**B**) Integrated intensity of GIWAXS data along q_z. (**C**) GIWAXS maps of each film. (**D** and **E**) Cross-sectional HR-STEM image (D) and high-angle annular dark-field (HAADF)/energy-dispersive x-ray spectroscopy elemental map (E) of the cross-sectional STEM image of the 2D-RT samples.

and (002) planes of 2D perovskite crystals (*28*). As expected, the 2D-TA films were dominated by $n = 1$ layers (with a prominent peak at $q_z \approx 0.35$ A^{-1}). In contrast, 2D-RT films exhibited the diffraction peaks of $n = 1$ and $n = 2$, with a more substantial $n = 2$ peak at lower q_z (*16*). The strong intensity in the *z*-direction for 2D perovskite films was indicative of a highly oriented lateral direction of the top 3D perovskite layers.

Cross-sectional high-resolution scanning transmission electron microscopy (HR-STEM) images also showed $n = 1$ and $n = 2$ layers in 2D-RT samples (Fig. 1D and fig. S2A) but only $n = 1$ in 2D-TA (fig. S2C), which is consistent with the GIWAXS results. To differentiate between $n = 1$ and $n = 2$ layers, we performed elemental mapping images and profiling positions of 2D perovskite layers; Fig. 1E and fig. S2B show two-dimensionalities of 2D layers in 2D-RT samples as the reduction of the density of C, Pb, and I elements, which correspond to $n = 1$ and $n = 2$ layers. In comparison, 2D-TA samples showed only $n = 1$ (fig. S2D). Further, profiling the position of 2D perovskites from TEM images confirmed the dimensionality of $n = 2$ and $n = 1$ by analyzing average distances between the two closest octahedral inorganic sheets. As a result, higher dimensionality ($n = 2$) had a wider distance (~1.5 nm) relative to $n = 1$ (~1.2 nm) (fig. S2, E to H).

Scanning electron microscopy (SEM) top-view images revealed that the surface morphology of the perovskite films after 2D perovskite passivation did not change substantially (fig. S3, A and B). Further, the 2D perovskite passivation films exhibited stronger photoluminescence (PL) emission with a longer PL decay lifetime than control 3D perovskite films because of the suppression of nonradiative recombination associated with trap states at the surface (figs. S4 and S5). Interestingly, the 2D perovskite ($n = 2$) capping layer formed uniformly on top of 3D perovskite surfaces for 2D-RT, as shown in the PL images in Fig. 2A. In addition, PL spectra in 2D-TA samples showed a dominant emission peak corresponding to $n = 1$, whereas a PL emission associated with a higher dimensionality of $n = 2$ is more pronounced in 2D-RT samples (see Fig. 2B), in accordance with GIWAXS and TEM results.

The energy-level diagrams of [2-(9*H*-carbazol-9-yl)ethyl]phosphonic acid (2PACz) anchored on indium tin oxide (ITO) [which we used as the hole-selective contact (*22*)], perovskites, and C_{60} are shown in Fig. 2C. With the OLAI

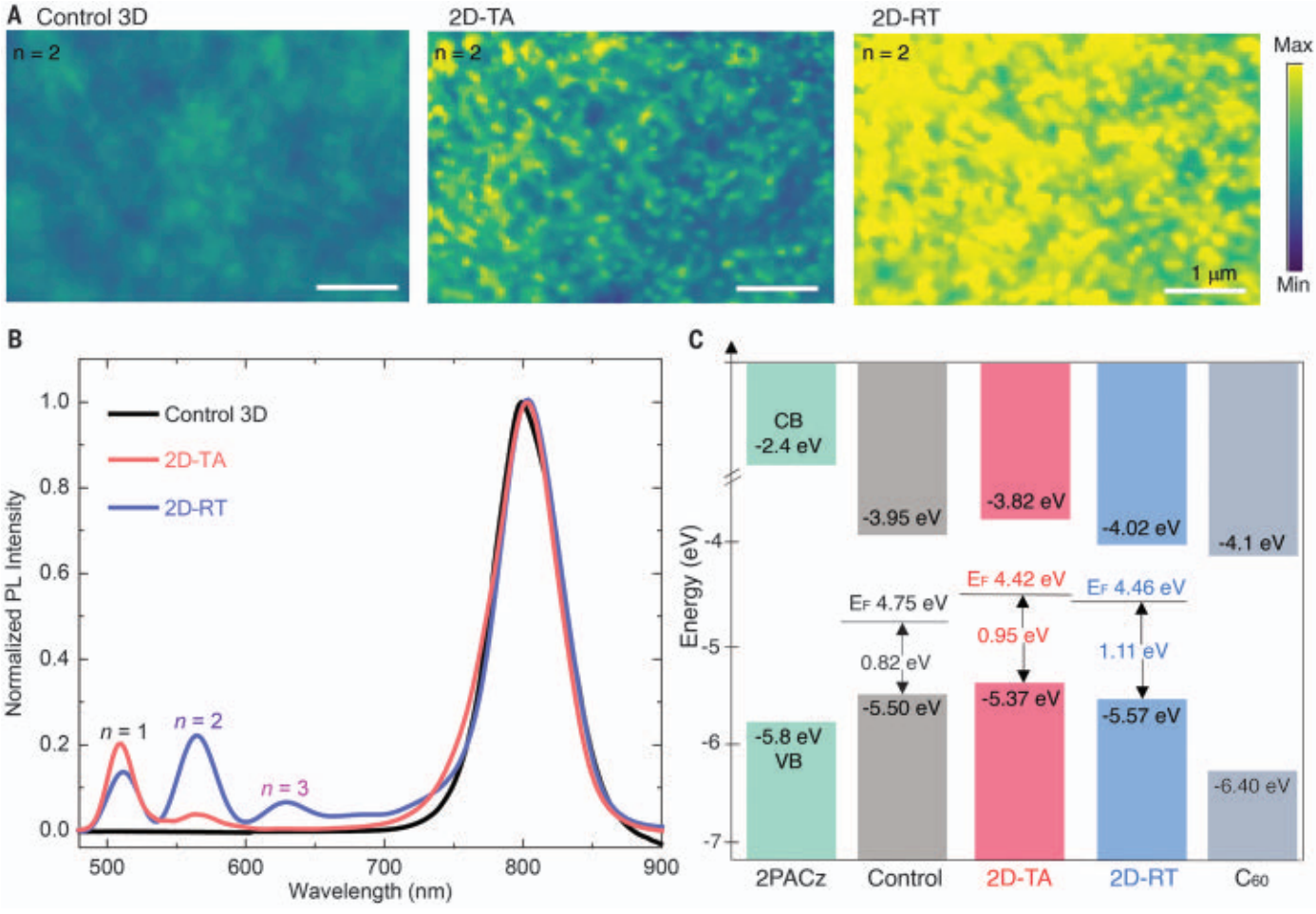


Fig. 2. Optical characterization and energetic alignments of perovskite films with and without 2D perovskite passivation. (**A**) PL images of control 3D, 2D-TA, and 2D-RT films at wavelength ~570 nm, which corresponds to n = 2 layers (images extracted using PHySpecV2). (**B**) Normalized PL spectra of each film from low to high wavelengths. (**C**) Energy level scheme for the control and OLAI-treated films extracted from UPS data. The VBM was obtained as $h\nu - (E_{cutoff} - E_{VB,min})$. The position of the CBM with respect to the VBM was defined by optical-E_g (1.55 eV).

post-treatment, the secondary electron cutoff (E_{cutoff}) shifted to a higher binding energy, indicating that the ion exchange–induced 3D-to-2D perovskite phase transition could lower the Fermi level (E_F) of post-treated perovskite films (fig. S6). Notably, the energetic gap between E_F and the valence band maximum (VBM) of the 2D-RT sample was wider, indicating the enhanced n-type character of post-treated 3D perovskite films, which we attribute to a successful 2D perovskite passivation strategy (*19*). The conduction band minimum (CBM) of 2D-RT films was also closer to the CBM of C_{60} at the n-type contact, which resulted in more efficient charge transfer at the 2D/3D perovskite interface and the C_{60} electron-selective layer. In contrast, the CBM of 2D-TA films was much higher than the CBM of C_{60} with less n-type character and resulted in less efficient charge transfer of this 2D/3D perovskite interface at the electron-selective contact. In addition, the effects of 2D perovskite capping layers also enhanced the resilience against moisture of 3D perovskite films, as shown by contact angle measurement in fig. S7.

Next, we fabricated inverted PSCs with a structure of glass/ITO/2PACz/3D perovskite/2D perovskite/C_{60}/bathocuproine (BCP)/Ag (Fig. 3A); Fig. 3B shows the false-colored cross-sectional SEM view of these devices. As shown in the current density–voltage (J-V) characteristics of the devices in Fig. 3C, the 2D-RT devices demonstrated substantially improved PCEs, with a maximum PCE of 24.3% and stabilized PCE of ~24% (open-circuit voltage, V_{OC}, of ~1.20 V and fill factor, FF, of ~82%; fig. S8, A and C). These results represent an absolute ~2% PCE gain upon 2D-RT passivation and can be compared with PCEs for other inverted PSCs (see fig. S9A).

Also, 2D-RT passivation enables minimization of the device energy loss ($E_{loss} = E_g - qV_{OC}$, where q is electron charge and E_g is the optical bandgap) up to 0.34 eV, which represents ~96% of the thermodynamic limit of the V_{OC} (1.262 V) for E_g of 1.55 eV (fig. S8, D and E). The reduced nonradiative loss in 2D-RT–based devices is comparable with state-of-the-art GaAs solar cells (V_{OC} of 1.127 V, yielding ~98% of the thermodynamic limit of the V_{OC}) (*29*). The 2D-TA passivated devices suffered from lower FF values (<79%; figs. S8B and S10), indicative of an energy level mismatch at the electron-selective contact, as derived from ultraviolet photoelectron spectroscopy (UPS) results (*20*). The narrow statistical distribution of the PCE, V_{OC}, FF, and short-circuit current density (J_{SC}) values of the devices (figs. S8B and S11) confirmed the high reproducibility of our approach. We also confirmed the effectiveness of our approaches by showing less than 0.5% deviation for person-to-person variations of seven different researchers (fig. S12A).

Further, our proposed passivation approach was universal for various perovskite compositions (various bandgaps) and deposition techniques (such as one-step, two-step, and blade-coating), as demonstrated by the

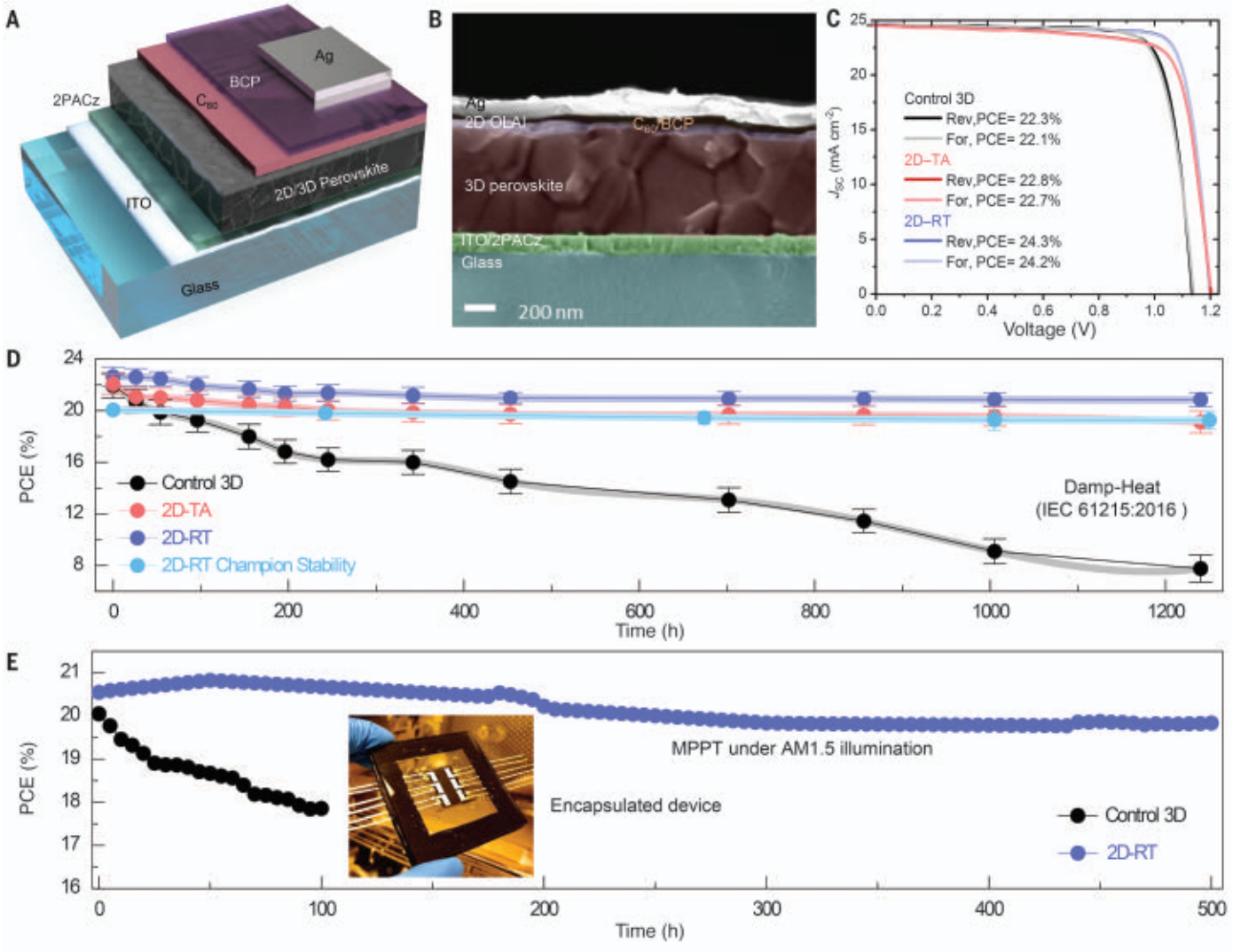


Fig. 3. Device performance and stability of 2D/3D perovskite heterojunction. (**A**) Device architecture of inverted PSCs. (**B**) Cross-sectional SEM image of inverted cells. (**C**) *J-V* scan of champion PSCs. (**D**) Variation of PCEs during damp-heat test of encapsulated devices. The continuous lines are guides to the eye; error bars denote SD. (**E**) Continuous MPPT for the encapsulated control and 2D-RT cells under AM 1.5 illumination in ambient air. Inset: Photograph of the encapsulated device.

systematic absolute PCE enhancement of 1.5 to 2.0% of device performance in fig. S12B. The reduced trap-assisted recombination of 2D/3D perovskite heterojunction devices was also investigated with transient photovoltage decay and light intensity dependent under open-circuit conditions (fig. S13). The 2D-passivated devices exhibited a longer charge recombination lifetime and lower ideality factor than control devices, confirming the reduced trap-assisted recombination at 3D/C_{60} interfaces by the 2D perovskite passivation.

Finally, we subjected our 2D perovskite passivation–treated PSCs to a set of rigorous stability tests. First, we evaluated the stability of our encapsulated devices when subjected to industry-relevant damp-heat tests (fig. S14). Here, our 2D-perovskite passivation simultaneously served as ion migration–blocking moisture/oxygen ingress barriers and as defect passivation layers, particularly at elevated operating temperatures (see fig. S15) (*3*, *14*–*17*). Indeed, the 2D-RT–based device retained >95% of the initial PCE (T_{95}) after >1200 hours for champion stability cells (Fig. 3D). Remarkably, after the damp-heat test, three devices showed an average PCE of 19.3 ± 0.69%.

Our results represent the successful encapsulation of PSCs passing the industry-relevant damp-heat test according to the IEC 61215:2016 protocols (*1*, *2*). Also, our final PCE of >19% after >1000 hours of damp-heat test represents a very high retained PCE (fig. S9B and table S1). There was no substantial change in the structural and optical properties of the 2D perovskite passivation films (both 3D and 2D perovskites) after >500 hours of thermal annealing at 85°C under dark condition, confirming the robustness of our 2D perovskite passivation approach (fig. S16).

Note that our encapsulated devices used for stability tests exhibited slightly lower initial PCEs than the unencapsulated devices because of J_{SC} losses originating from the encapsulant and glass sheets (fig. S17). We also tested unencapsulated devices in our damp-heat chamber, applying thermal tests in ambient air with relative humidity of >50% (figs. S18 and S19), representative of extreme outdoor conditions. Our 2D capping layer introduced a substantially enhanced resistance of the devices against high moisture and thermal stress. Finally, we performed maximum power point tracking (MPPT) measurements for encapsulated cells under simulated 1-sun illumination in ambient air for >500 hours (Fig. 3E). Here, the 2D-RT–based devices retained up to

~95% of their initial PCE after an MPPT test of >500 hours, whereas the control devices retained their PCE of <90% for only ~100 hours.

REFERENCES AND NOTES

1. A. Mei *et al.*, *Joule* **4**, 2646–2660 (2020).
2. L. Shi *et al.*, *Science* **368**, eaba2412 (2020).
3. Y.-W. Jang *et al.*, *Nat. Energy* **6**, 63–71 (2021).
4. K. T. Cho *et al.*, *Energy Environ. Sci.* **11**, 952–959 (2018).
5. S. Yang *et al.*, *Science* **365**, 473–478 (2019).
6. Y. Lin *et al.*, *J. Phys. Chem. Lett.* **9**, 654–658 (2018).
7. Y.-H. Lin *et al.*, *Science* **369**, 96–102 (2020).
8. J. Xu *et al.*, *Science* **367**, 1097–1104 (2020).
9. Y. Zhao *et al.*, *Nat. Commun.* **11**, 6328 (2020).
10. J. Jeong *et al.*, *Nature* **592**, 381–385 (2021).
11. S. Chen *et al.*, *Science* **373**, 902–907 (2021).
12. R. Wang *et al.*, *Science* **366**, 1509–1513 (2019).
13. H. Zhou *et al.*, *Science* **345**, 542–546 (2014).
14. J. J. Yoo *et al.*, *Energy Environ. Sci.* **12**, 2192–2199 (2019).
15. E. H. Jung *et al.*, *Nature* **567**, 511–515 (2019).
16. A. H. Proppe *et al.*, *Nat. Commun.* **12**, 3472 (2021).
17. G. Yang *et al.*, *Nat. Photonics* **15**, 681–689 (2021).
18. Q. Yao *et al.*, *Adv. Mater.* **32**, e2000571 (2020).
19. D. H. Cao, C. C. Stoumpos, O. K. Farha, J. T. Hupp, M. G. Kanatzidis, *J. Am. Chem. Soc.* **137**, 7843–7850 (2015).
20. J. Yang *et al.*, *Adv. Energy Mater.* **10**, 2000687 (2020).
21. Y. Shao, Z. Xiao, C. Bi, Y. Yuan, J. Huang, *Nat. Commun.* **5**, 5784 (2014).
22. A. Al-Ashouri *et al.*, *Energy Environ. Sci.* **12**, 3356–3369 (2019).
23. A. Al-Ashouri *et al.*, *Science* **370**, 1300–1309 (2020).
24. E. Aktas *et al.*, *Energy Environ. Sci.* **14**, 3976–3985 (2021).
25. I. Levine *et al.*, *Joule* **5**, 2915–2933 (2021).
26. S. Gharibzadeh *et al.*, *Energy Environ. Sci.* **14**, 5875–5893 (2021).
27. J. Xing *et al.*, *Nat. Commun.* **9**, 3541 (2018).
28. A. H. Proppe *et al.*, *J. Am. Chem. Soc.* **141**, 14180–14189 (2019).
29. M. A. Green *et al.*, *Prog. Photovolt. Res. Appl.* **27**, 3–12 (2018).

ACKNOWLEDGMENTS

We acknowledge the use of KAUST Solar Center and Core Lab facilities and the support from its staff. We thank A. Prasetio for help in preparing the figures with Adobe Illustrator, and W. Yan for assistance with contact angle measurement. **Funding:** Supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under awards OSR-CARF/CCF-3079, OSR-CRG2019-4093, OSR-CRG2020-4350, IED OSR-2019-4208, IED OSR-2019-4580, and REI/1/4833-01-01. **Author contributions:** R.A. conceived and conducted the research. R.A., E.A., M.D.B., T.G.A., and S.D.W. wrote the original draft. E.A., E.U., F.A., J.L., and A.S.S. fabricated the devices. E.U. and F.A. performed optical spectroscopy measurements and did data analysis. A.S. conducted TEM measurement and analyses. F.A. helped to measure XRD patterns. G.T.H. performed XPS and U.P.S. measurements and analyses. M.K.E. performed DFT calculations. A.S.S., F.A., and F.Z. performed SEM measurements. Y.C. performed the 2D-GIWAXS measurements. M.I.N. performed the AFM and impedance measurements. A.S.S. performed the characterization of the device, including impedance and transient photo-voltage and current measurements. R.A., M.D.B., and M.B. performed the stability tests. J.L. and A.U.R. helped to fabricate the aperture mask. C.-L.W., T.D.A., U.S., and S.D.W. supervised the project and contributed to the manuscript; funding acquisition by S.D.W. **Competing interests:** R.A., S.D.W., E.A., and M.D.B. are inventors on a patent application related to this work filed by King Abdullah University of Science and Technology (KAUST reference 2022-067). The other authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm5784
Materials and Methods
Figs. S1 to S19
Table S1
References (*30*–*38*)

27 September 2021; resubmitted 13 December 2021
Accepted 2 February 2022
Published online 17 February 2022
10.1126/science.abm5784

GAMMA-RAY ASTRONOMY

Time-resolved hadronic particle acceleration in the recurrent nova RS Ophiuchi

H.E.S.S. Collaboration*†

Recurrent novae are repeating thermonuclear explosions in the outer layers of white dwarfs, due to the accretion of fresh material from a binary companion. The shock generated when ejected material slams into the companion star's wind can accelerate particles. We report very-high-energy [VHE; ≥100 giga–electron volts] gamma rays from the recurrent nova RS Ophiuchi, up to 1 month after its 2021 outburst, observed using the High Energy Stereoscopic System (H.E.S.S.). The temporal profile of VHE emission is similar to that of lower-energy giga–electron volt emission, indicating a common origin, with a 2-day delay in peak flux. These observations constrain models of time-dependent particle energization, favoring a hadronic emission scenario over the leptonic alternative. Shocks in dense winds provide favorable environments for efficient acceleration of cosmic rays to very high energies.

RS Ophiuchi (RS Oph) is a recurrent nova system composed of a white dwarf and a companion red giant star. Novae are a source of high-energy particles and photons (*1*, *2*), with nonthermal gamma-ray emission in the range of ~100 MeV to ~10 GeV (*3*). We adopt a distance of ~1.4 kpc from Earth to the RS Oph system (*4*); analysis of astrometric data suggests larger distances of 2.3 kpc (*5*) or 2.7 kpc (*6*), but we regard these estimates as less reliable because of the binary orbital motion in RS Oph. The binary components have a separation of 1.48 astronomical units, close enough for the white dwarf to continually accrete material from its companion star (*7*). At irregular intervals, enough material accumulates on the surface of the white dwarf to trigger a thermonuclear explosion, driving a quasi-spherical shock (*8*) into the red giant's wind (fig. S2). Eight outbursts were observed between 1898 and 2006, recurring at intervals of 9 to 26 years (*9*).

On 8 August 2021, an outburst of RS Oph was identified in optical observations (*10*), reaching a peak naked-eye visual magnitude of 4.5, more than 1000 times as bright as the quiescent visual magnitude of 12.5. We observed RS Oph with the High Energy Stereoscopic System (H.E.S.S.), an array of five atmospheric Cherenkov telescopes (*8*). Observations commenced on 9 August 2021 and continued for five nights, until 13 August 2021. A higher optical background due to moonlight prevented good quality observations for the next 10 nights. During each of the first five nights, H.E.S.S. detected point-like gamma-ray emission from the direction of RS Oph, with a significance of >6σ on each night (table S1). The combined data for those five nights are shown in Fig. 1. Observations recommenced on 25 August 2021, ~17 days after the initial outburst. We find evidence for a much weaker signal, ~3σ above the background, in ~15 hours (after quality cuts) of data accumulated over the subsequent 14 days.

We performed a spectral analysis of the H.E.S.S. data for the first five observation nights separately. We also separated the data from the array of four 106-m^2 mirror area telescopes (designated CT1 to CT4) and a fifth low-threshold telescope (CT5) with a mirror area of 612 m^2 (*8*). We find that the very-high-energy (VHE) flux is variable, with a spectral index >3 throughout (table S2).

Figure 2 shows the time evolution of the gamma-ray flux from RS Oph for photon energies between 250 GeV and 2.5 TeV. The VHE gamma-ray flux rises smoothly from T_0, the time of peak optical emission in the V band (*10*), until a VHE peak on the third night of observations. The VHE gamma-ray flux then decays by an order of magnitude over a 2-week period. We obtained 60-MeV to 500-GeV data taken by the *Fermi*-LAT (Large Area Telescope) instrument for the same time period as the H.E.S.S. observations, which are also shown in Fig. 2. The *Fermi*-LAT flux varies in the range of ~1×10^{-8} to 2×10^{-10} erg cm^{-2} s^{-1}, with a peak on T_0 + 1 day. The VHE gamma-ray emission peak is delayed by a further 2 days.

We used a power-law model $t^{-\alpha}$, with exponent α and time t, to fit the decay of the flux after the peak. For the choice of T_0 = 1 day, the best-fitting values are $\alpha_{HESS} = 1.43 \pm 0.18$ for H.E.S.S. and $\alpha_{LAT} = 1.31 \pm 0.07$ for *Fermi*-LAT. We find consistent results for the flux and temporal decay if the *Fermi*-LAT data are analyzed in bins of 24-hour duration (*8*); the higher signal-to-noise ratio of these larger bins enables a more detailed spectral analysis.

The combined H.E.S.S. and *Fermi*-LAT data allow us to measure wide-band gamma-ray

***Corresponding authors: contact.hess@hess-experiment.eu; Alison Mitchell (alison.mw.mitchell@fau.de); Stefan Ohm (stefan.ohm@desy.de); Brian Reville (brian.reville@mpi-hd.mpg.de)**
†H.E.S.S. Collaboration authors and affiliations are listed in the supplementary materials.

spectra over more than four orders of magnitude in energy and to follow their temporal evolution (Fig. 3). The RS Oph spectra are consistent with a log-parabola model. Comparison of spectra taken on different nights shows a general trend for the flux normalization to decrease and the parabola to widen over time (*8*) (table S2).

The similarity between the spectra of the *Fermi*-LAT and H.E.S.S. data and their similar decay profiles after their respective peaks indicate a common origin for the gamma rays from 1 day to 1 month after the explosion (supplementary text). We assume that the gamma rays are emitted by particles accelerated at the external shock, as it propagates into the wind of the red giant (fig. S2). Optical spectroscopic measurements of the 2021 nova indicate shock velocities (u_{sh}) ranging from 4000 to 5000 km s^{-1} (*11*), compatible with measurements from the previous 2006 outburst of RS Oph (*12*, *13*). High-resolution images of the 2006 event (*14*) indicated that the polar regions of the shock expanded at ≈5000 km s^{-1} over the first 5 months. We therefore assume that, during the first week after the 2021 outburst, the shock velocity did not fall below several thousand kilometers per second.

The images of the 2006 nova showed a quasi-spherical outflow, pinched at an equatorial ring (*14*, *15*). This is consistent with a shock expanding into the wind of the red giant, orthogonal to the orbital plane of the binary but inhibited close to the plane by the denser gas (*7*, *16*). We expect particles to undergo diffusive shock acceleration at the external fast-moving shocks, above and below the orbital plane of the binary. We consider two scenarios to explain the observed spectral and temporal properties of the gamma-ray emission from RS Oph: (i) gamma-ray production from accelerated protons colliding with dense gas in the downstream volume (hadronic π^0 decay model) and (ii) gamma-ray production from energetic electrons scattering low-energy photons from the nova (leptonic inverse Compton model). For both models, the observations place strong constraints on the physical conditions, particularly the acceleration efficiencies required to match the measured fluxes and maximum photon energies (supplementary text).

VHE gamma-ray emission requires acceleration of particles to energies above the tera–electron volt level. The maximum energy a particle attains at a shock is determined either by the time taken before radiative cooling dominates over acceleration or by when particles become too energetic and thus escape upstream of the shock (*17*). This confinement limit applies when the accelerating particles are unable to excite magnetic field fluctuations to a sufficient level ahead of the shock. Because particles spend more time diffusing upstream than downstream, the details of the downstream magnetic fields can be neglected (*18*). For upstream magnetic field amplification to be effective, a sufficient flux of particles, typically protons, must escape upstream of the shock. This requires both efficient transfer of the shock kinetic energy to relativistic protons (i.e., those traveling at speeds close to the speed of light) and that a fraction of these protons penetrate far upstream. Escaping particles have energies concentrated close to the maximum particle energy; less-energetic particles are confined to the shock. The escaping flux per unit area at a given shock radius can be parameterized as

$$j_{esc} = e\xi_{esc}F_{\varepsilon}/E_{max} \quad (1)$$

where e is the elementary charge, $F_{\varepsilon} = \frac{1}{2}\rho_{up}u_{sh}^3$ is the energy flux density for a high–Mach-number nonrelativistic shock, ρ_{up} is the immediate upstream gas density at that radius, E_{max} is the maximum particle energy that dominates the escaping flux, and the efficiency parameter ξ_{esc} depends on the assumed particle spectrum (supplementary text). For a wind-like density profile, and neglecting radiative losses, the confinement limit on the maximum energy for a particle with atomic number Z, and thus charge $|Z|e$, is

$$E_{max} = 1.5|Z|\left(\frac{\xi_{esc}}{0.01}\right)\left(\frac{\dot{M}/v_{wind}}{10^{11}\ \mathrm{kg\ m^{-1}}}\right)^{1/2}\left(\frac{u_{sh}}{5000\ \mathrm{km\ s^{-1}}}\right)^{2}\ \mathrm{TeV} \quad (2)$$

where $\dot{M}$ and v_{wind} are the mass-loss rate and the wind velocity, respectively, of the red giant. ξ_{esc} is predicted to be ~1% for high–Mach-number shocks (*17*). For RS Oph, $\dot{M}/v_{wind} = 6 \times 10^{11}$ kg m^{-1} (*15*), which, together with the inferred shock velocities, indicates a maximum energy $E_{max} \approx 10$ TeV. This is compatible with the measured maximum photon energies $E_{\gamma,max} \approx 1$ TeV (Fig. 3).

In the hadronic scenario, the gamma-ray light curves are consistent with an expanding shock in a decreasing density profile. With the adopted distance of 1.4 kpc (*4*), the measured gamma-ray fluxes require that >10% of the post-shocked medium's internal energy goes into accelerating protons or other nuclei. The delay between the peaks in the *Fermi*-LAT and H.E.S.S. light curves would then reflect the finite acceleration time of the >1 TeV protons or, more specifically, the time taken to populate the high-energy tail of the distribution (*19*). A simple calculation of the acceleration time (supplementary text), on the basis of a comparison of the confinement and Hillas limits, implies that this should happen on the order of days. This is consistent with the spectral evolution seen in Fig. 3: a reduction in the *Fermi*-LAT flux, accompanied by a hardening in H.E.S.S. flux and increased $E_{\gamma,max}$ over the first few days after the outburst. Attenuation of gamma-rays, due to the nova's optical and infrared photon fields, is minor below 1 TeV a few hours after the explosion; therefore, attenuation alone cannot account for the observed hardening (supplementary text and fig. S10).

For the alternative leptonic scenario, in which tera–electron volt gamma-rays are produced by VHE electrons, the acceleration needs to overcome the strong radiative losses due to inverse Compton cooling in the strong photon fields of the nova, as well as synchrotron cooling in the magnetic field in the shock region. To achieve this, electrons must accelerate at close to the Bohm rate—i.e., the scattering rate equal to the rate of gyration in the magnetic field (*18*). Such efficient scattering requires strong self-generated magnetic fluctuations upstream of the shock, which implies the presence

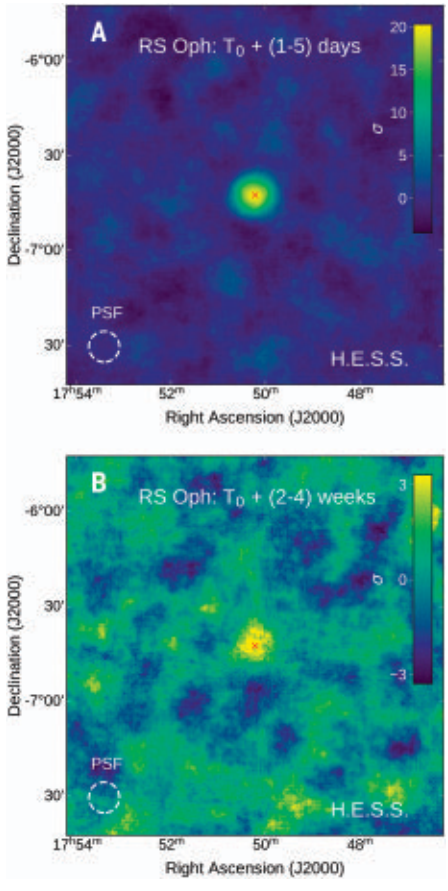


Fig. 1. RS Oph significance maps. Significance maps derived from the H.E.S.S. >100-GeV gamma-ray observations for the early (**A**) and late (**B**) phases of the RS Oph 2021 outburst. T_0 is the time of peak optical emission, modified Julian day 59435.25. The dashed white circles indicate the H.E.S.S. point spread function (PSF).

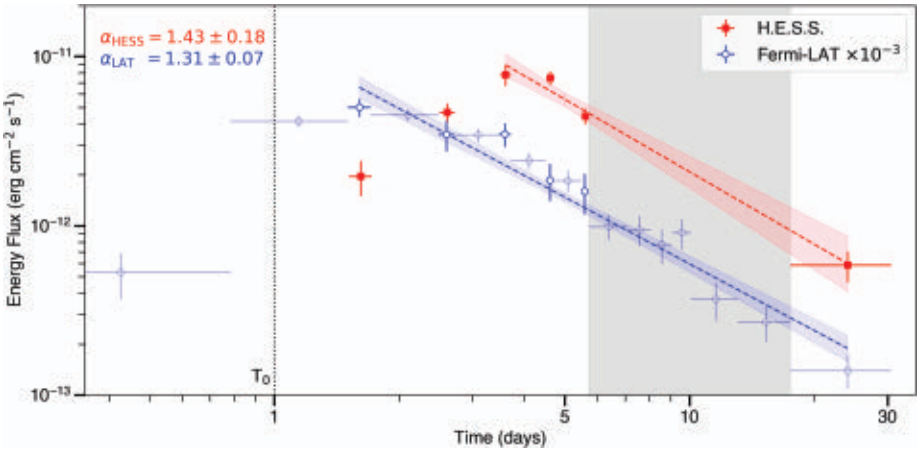


Fig. 2. Gamma-ray light curves of RS Oph. Light curves of gamma-ray emission from RS Oph, including data from *Fermi*-LAT and H.E.S.S. observations. The H.E.S.S. data (red squares) cover a period of five nights, after which observations paused for 10 days (shaded gray band) owing to bright moonlight and then recommenced for a period of 14 days. The H.E.S.S. flux is integrated from 250 GeV to 2.5 TeV, whereas the *Fermi*-LAT flux is integrated from 60 MeV to 500 GeV. *Fermi*-LAT data are shown in 6-hour bins (bright blue circles) corresponding to the time windows of the H.E.S.S. observations; data collected outside of these times are shown with semitransparent markers. Error bars denote 1σ statistical uncertainties. A power-law model was fitted to the temporal decay after the time of peak flux for both instruments (red and blue dashed lines, with uncertainties indicated by the shaded regions). The vertical dotted black line indicates T_0, the peak of the outburst in the optical waveband.

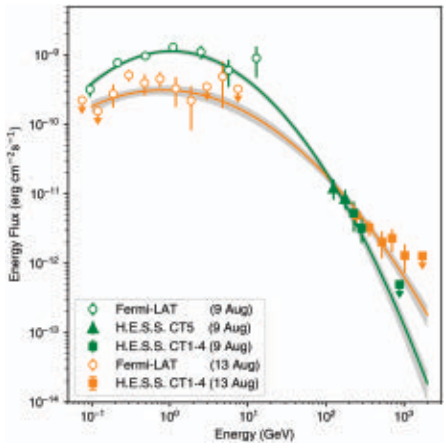


Fig. 3. RS Oph gamma-ray spectra. H.E.S.S. and *Fermi*-LAT spectra for 9 August (green) and 13 August (orange), fitted with a log-parabola model. The analysis is applied separately for the H.E.S.S. CT1 to CT4 (squares) and CT5 (triangles) detectors (*8*). The *Fermi*-LAT data (open circles) are integrated over 24 hours centered at the H.E.S.S. observation times. There is spectral evolution from 9 to 13 August, with a reduction in the *Fermi*-LAT flux and an increase in the maximum energy of the tera–electron volt spectrum. Error bars denote 1σ statistical uncertainty, and upper limits are 95% confidence levels.

of an energetic relativistic hadronic component. In this scenario, the differences between the spectral slopes in the *Fermi*-LAT and H.E.S.S. energy ranges are a consequence of the energy-dependent cooling rates in time-dependent photon fields. Electrons that radiate in the VHE band cool on a time scale less than the age of the nova remnant at the times the observations were taken, whereas lower-energy uncooled electrons accumulate downstream over time. The *Fermi*-LAT light curve in this scenario then reflects the evolution of the energy density of soft-photon targets, whereas the H.E.S.S. light curve traces the full radiative output of high-energy electrons up to the VHE peak (supplementary text). After the peak, owing to the rapidly decreasing photon energy density, the cooling time increases faster than the remnant's age, and the VHE emitting electrons are also slow cooling.

In this time-dependent numerical single-zone model, parameters can be found to approximately describe the light curves and spectra in both leptonic and hadronic scenarios, providing quantitative estimates for the acceleration efficiencies at early times t ($t \leq T_0 + 5$ days). We are unable to account for the temporal decay at later times, probably because we do not consider the complex internal structure of the nova remnant, its nonspherical geometry, or the escape of particles from the emission zone. Both the leptonic and hadronic models at early times are consistent with continuous injection of particles following a power law spectrum in energy $E^{-2.2}$ with a high-energy cutoff. To approximately match the observed flux, the hadronic model requires >10% of the shocked gas's internal energy to be transferred to nonthermal protons, whereas the leptonic model requires >1% efficiency for nonthermal electrons.

Such a high fraction of the total energy in nonthermal electrons is inconsistent with theories of injection at high–Mach-number shocks, for which the ion injection efficiency is expected to be much higher than that of electrons (*20*). Numerical simulations of high–Mach-number shocks find a ratio of $\ll 10^{-2}$ for electron to ion energy densities (*21*), consistent with multiwavelength models of supernova remnants (*22*). A >1% efficiency of conversion to nonthermal electrons cannot be realized in a purely leptonic model. For this reason, we prefer the hadronic scenario discussed above, for which both the implied high–proton-acceleration efficiencies and inferred maximum energy are in line with theoretical predictions (*17*). Our findings support previous hadronic models of gamma-ray novae (*23*–*25*).

The VHE detection of RS Oph demonstrates that particle acceleration to energies around the tera–electron volt level can occur within the dense wind environments of recurrent novae. The total kinetic energy from each nova of RS Oph is estimated to be ~10^{43} erg [10^{-7} solar masses ($M_\odot$) of ejecta at ~4000 km s^{-1} (*13*)], with a large fraction of this being converted to relativistic protons and heavier nuclei, which are the main constituents of Galactic cosmic rays. Each nova event generates enough cosmic rays to fill a cubic parsec volume with an energy density of ~0.1 eV cm^{-3}, similar to the local Galactic cosmic-ray energy density of ~(0.8 to 1.0) eV cm^{-3} (*26*) sustained by supernovae. In the case of RS Oph, the cosmic-ray energy input recurs approximately every Δt = 15 to 20 years, leading to an almost-continuous injection of nonthermal particles. For a diffusion coefficient D in the neighborhood of RS Oph, the cosmic-ray output from each nova is spread over a diffusion length $\ell_{\rm diff} = \sqrt{4D\Delta t}$. Using a Galactic average $D \approx (3-5) \times 10^{28}$ cm^2 s^{-1} (*27*), we find $\ell_{\rm diff} > 1$ pc, and the contribution from each nova is less than the average Galactic cosmic-ray population. If the diffusion coefficient in the neighborhood of the nova is much lower than the Galactic average, perhaps due to enhanced turbulence after previous novae, such a sustained source of cosmic rays will raise the local abundance. If efficient acceleration of particles to tera–electron volt energies in recurrent novae is commonplace, with spectral energy distribution harder than that of the Galactic cosmic-ray background $\propto E^{-2.7}$, the local contribution from novae would dominate at tera–electron volt energies over volumes above the cubic parsec level. The size of the affected region will depend on the value of the diffusion coefficient, which can be constrained with measurements of

the diffuse gamma-ray emission at energies ~10 to 100 GeV (*28*).

Our time-resolved gamma-ray emission measurements have implications for the origin of cosmic rays. Acceleration of cosmic rays to peta–electron volt energies in supernova remnants requires substantial amplification of magnetic fields. Fast shocks (~10,000 km s^{-1}) propagating through the dense winds ($\dot{M}/v_{wind} \sim 10^{13}$ kg m^{-1}) associated with the progenitors of supernova remnants from massive ($\gtrsim 8\,M_{\odot}$) stars provide the only known environments where the required conditions can (in theory) be met (*17*, *29*). However, observational confirmation of this prediction has not been found. The detection of VHE gamma-rays from RS Oph provides an example of a Galactic accelerator reaching the theoretical limit for the maximum achievable particle energy via diffusive shock acceleration (*17*). If our results can be extrapolated to the most optimistic supernova conditions, they support the prevailing model of Galactic peta–electron volt cosmic rays originating in supernova remnants from massive stars (*17*, *29*).

REFERENCES AND NOTES

1. A. A. Abdo *et al*., *Science* **329**, 817–821 (2010).
2. Fermi-LAT Collaboration *et al*., *Science* **345**, 554–558 (2014).
3. L. Chomiuk, B. D. Metzger, K. J. Shen, *Annu. Rev. Astron. Astrophys.* **59**, 391–444 (2021).
4. R. K. Barry *et al*., in *RS Ophiuchi (2006) and the Recurrent Nova Phenomenon*, A. Evans, M. F. Bode, T. J. O'Brien, M. J. Darnley, Eds., vol. 401 of *Astronomical Society of the Pacific Conference Series* (Astronomical Society of the Pacific, 2008), pp. 52–60.
5. Gaia Collaboration *et al*., *Astron. Astrophys.* **616**, A1 (2018).
6. L. Lindegren *et al*., *Astron. Astrophys.* **649**, A4 (2021).
7. R. A. Booth, S. Mohamed, P. Podsiadlowski, *Mon. Not. R. Astron. Soc.* **457**, 822–835 (2016).
8. Materials and methods are available as supplementary materials.
9. E. Brandi, C. Quiroga, J. Mikołajewska, O. E. Ferrer, L. G. Garcia, *Astron. Astrophys.* **497**, 815–825 (2009).
10. S. Kafka, Observations from the AAVSO International Database (2021); www.aavso.org.
11. J. Mikolajewska, E. Aydi, D. Buckley, C. Galan, M. Orio, The Astronomer's Telegram 14852 (2021).
12. M. F. Bode *et al*., *Astrophys. J.* **652**, 629–635 (2006).
13. J. L. Sokoloski, G. J. M. Luna, K. Mukai, S. J. Kenyon, *Nature* **442**, 276–278 (2006).
14. M. F. Bode *et al*., *Astrophys. J.* **665**, L63–L66 (2007).
15. T. J. O'Brien *et al*., *Nature* **442**, 279–281 (2006).
16. R. Walder, D. Folini, S. N. Shore, *Astron. Astrophys.* **484**, L9–L12 (2008).
17. A. R. Bell, K. M. Schure, B. Reville, G. Giacinti, *Mon. Not. R. Astron. Soc.* **431**, 415–429 (2013).
18. P. O. Lagage, C. J. Cesarsky, *Astron. Astrophys.* **125**, 249–257 (1983).
19. L. O. Drury, *Mon. Not. R. Astron. Soc.* **251**, 340–350 (1991).
20. M. A. Malkov, L. O. Drury, *Rep. Prog. Phys.* **64**, 429–481 (2001).
21. J. Park, D. Caprioli, A. Spitkovsky, *Phys. Rev. Lett.* **114**, 085003 (2015).
22. H. J. Völk, E. G. Berezhko, L. T. Ksenofontov, *Astron. Astrophys.* **483**, 529–535 (2008).
23. B. D. Metzger *et al*., *Mon. Not. R. Astron. Soc.* **457**, 1786–1795 (2016).
24. K.-L. Li *et al*., *Nat. Astron.* **1**, 697–702 (2017).
25. E. Aydi *et al*., *Nat. Astron.* **4**, 776–780 (2020).
26. A. C. Cummings *et al*., *Astrophys. J.* **831**, 18 (2016).
27. A. W. Strong, I. V. Moskalenko, V. S. Ptuskin, *Annu. Rev. Nucl. Part. Sci.* **57**, 285–327 (2007).
28. M. Ackermann *et al*., *Astrophys. J.* **750**, 3 (2012).
29. A. Marcowith, V. V. Dwarkadas, M. Renaud, V. Tatischeff, G. Giacinti, *Mon. Not. R. Astron. Soc.* **479**, 4470–4485 (2018).

ACKNOWLEDGMENTS

We appreciate the excellent work of the technical support staff in Berlin, Zeuthen, Heidelberg, Palaiseau, Paris, Saclay, Tübingen, and Namibia in the construction and operation of the H.E.S.S. equipment. This work benefited from services provided by the H.E.S.S. Virtual Organisation, supported by the national resource providers of the EGI Federation. **Funding:** The support of the Namibian authorities and of the University of Namibia in facilitating the construction and operation of H.E.S.S. is gratefully acknowledged, as is support from the German Ministry for Education and Research (BMBF), the Max Planck Society, the German Research Foundation (DFG), the Helmholtz Association, the Alexander von Humboldt Foundation, the French Ministry of Higher Education, Research and Innovation, the Centre National de la Recherche Scientifique (CNRS/IN2P3 and CNRS/INSU), the Commissariat à l'énergie atomique et aux énergies alternatives (CEA), the UK Science and Technology Facilities Council (STFC), the Irish Research Council (IRC), the Science Foundation Ireland (SFI), the Knut and Alice Wallenberg Foundation, the Polish Ministry of Education and Science (agreement no. 2021/WK/06), the South African Department of Science and Technology and National Research Foundation, the University of Namibia, the National Commission on Research, Science & Technology of Namibia (NCRST), the Austrian Federal Ministry of Education, Science and Research and the Austrian Science Fund (FWF), the Australian Research Council (ARC), the Japan Society for the Promotion of Science, the University of Amsterdam, and the Science Committee of Armenia (grant 21AG-1C085).
Author contributions: A. Mitchell and S. Ohm led the H.E.S.S. observations of RS Oph. R. Konno carried out initial on-site analysis, the main H.E.S.S. CT1 to CT4 stereo data analysis and the atmospheric correction. S. Steinmassl performed the CT5 mono analysis, and J. P. Ernenwein performed the cross-check analysis. S. Ohm coordinated the multiple H.E.S.S. analyses and evaluated the systematic errors. E. de Oña Wilhelmi and T. Unbehaun carried out the *Fermi*-LAT data analysis. B. Reville, D. Khangulyan, J. Mackey, and E. de Oña Wilhelmi developed the interpretation and modeling. The manuscript was prepared by A. Mitchell, B. Reville, S. Ohm, D. Khangulyan, E. de Oña Wilhelmi, R. Konno, and S. Steinmassl. S. Wagner is the collaboration spokesperson. Other H.E.S.S. Collaboration authors contributed to the design, construction, and operation of H.E.S.S. and the development and maintenance of data handling, data reduction, or data analysis software. All authors meet the journal's authorship criteria and have reviewed, discussed, and commented on the results and the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The H.E.S.S. data are available at www.mpi-hd.mpg.de/hfm/HESS/pages/publications/auxiliary/2022_RS_Oph/. This includes the sky maps in Fig. 1, the light-curve data plotted in Fig. 2, the spectral energy distributions used to produce Fig. 3, and the data values plotted in figs. S8 and S9. The Fermi-LAT data were downloaded from https://fermi.gsfc.nasa.gov/cgi-bin/ssc/LAT/LATDataQuery.cgi; we used 60-MeV to 500-GeV photons within 15° of RS Oph from 7 August 2021 to 13 September 2021. Our emission modeling software code is available at https://github.com/dmikha/RS_Oph.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn0567
H.E.S.S. Collaboration Authors and Affiliations
Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 to S4
References (*30*–*71*)

1 November 2021; accepted 1 March 2022
Published online 10 March 2022
10.1126/science.abn0567

PALEONTOLOGY

Brawn before brains in placental mammals after the end-Cretaceous extinction

Ornella C. Bertrand[1]*, Sarah L. Shelley[1,2], Thomas E. Williamson[3], John R. Wible[2], Stephen G. B. Chester[4,5,6], John J. Flynn[7,8,9,10], Luke T. Holbrook[11], Tyler R. Lyson[12], Jin Meng[7], Ian M. Miller[13,12], Hans P. Püschel[1], Thierry Smith[14], Michelle Spaulding[15], Z. Jack Tseng[16], Stephen L. Brusatte[1,3]*

Mammals are the most encephalized vertebrates, with the largest brains relative to body size. Placental mammals have particularly enlarged brains, with expanded neocortices for sensory integration, the origins of which are unclear. We used computed tomography scans of newly discovered Paleocene fossils to show that contrary to the convention that mammal brains have steadily enlarged over time, early placentals initially decreased their relative brain sizes because body mass increased at a faster rate. Later in the Eocene, multiple crown lineages independently acquired highly encephalized brains through marked growth in sensory regions. We argue that the placental radiation initially emphasized increases in body size as extinction survivors filled vacant niches. Brains eventually became larger as ecosystems saturated and competition intensified.

Mammals have the largest brains, absolutely and relative to body size, of all vertebrates, reaching an extreme in the hyperinflated brain of humans (*1*, *2*). The mammalian brain was assembled over 200 million–plus years of evolution, beginning with encephalization pulses—increases in brain size relative to body size—on the mammal stem lineage in the Mesozoic. These increases were associated with heightened olfaction and the origin of the neocortex, a new part of the cerebrum involved in higher cognition and sensory integration (*3*). Among extant mammals, the more than 6000 species of placentals (*4*) exhibit remarkable diversity of encephalization (*5*) and sensory abilities (*6*), the result of complex interplay between brain and body size changes over time and

across taxa (*7*). It has long been recognized that the earliest fossils of extant placental orders (herein "placental crown orders") from the Eocene (56 million to 34 million years ago) had brains similar in structure to (*8*, *9*), but smaller than, those of their modern-day counterparts (*10*, *11*) and that relative brain size generally increased over the Cenozoic (*12*, *13*). However, much about the origin of the placental brain, and when and how it encephalized to modern levels, remains unclear.

In particular, little is known about the transition from the ancestral brains of Mesozoic mammals to the more modern brains of Eocene crown placentals. The principal gap in understanding is the Paleocene (66 million to 56 million years ago), the interval after the end-Cretaceous mass extinction, when placentals and close kin radiated into niches vacated by dinosaurs (*14*), ballooned in body size (*15*, *16*), and inaugurated the Age of Mammals (*17*, *18*). Jerison (*1*) posited that the "archaic" placentals replacing dinosaurs (herein "placental stem taxa," those that do not clearly belong to extant orders) were overgrown versions of Mesozoic mammals, which passively increased their absolute brain sizes to keep pace with body size expansion and then, later in the Eocene, actively encephalized to increase relative brain sizes. This hypothesis was based on a small sample of skulls, of uncertain Paleocene or Eocene age, whose brain cavities were measured with basic volumetric techniques unable to distinguish sensory regions. Recently, a broad study of mammals instead identified increases in relative brain size after the end-Cretaceous extinction (*7*), but this alternative hypothesis was inferred from phylogenetic comparative data that did not directly include Paleocene fossils. Testing these competing hypotheses has proven difficult because well-preserved Paleocene mammal skulls have been notoriously rare.

We assessed the early evolution of the placental brain with an expansive dataset of Mesozoic and Cenozoic mammals, including newly discovered Paleocene skulls from the San Juan Basin of New Mexico (*19*) and Denver Basin of Colorado (*20*). To do so, we used high-resolution computed tomography (CT) to measure the size of the brain and its individual sensory components (neocortex, olfactory bulbs, and cerebellar petrosal lobules involved in control of eye movements) (Fig. 1) (*21*). Our dataset includes 34 new CT scans of Paleogene taxa (17 Paleocene and 17 Eocene) (table S1), alongside data from previous work. We examined patterns over time and across taxa to quantify body size, brain size, and sensory regions change during the placental radiation and after the end-Cretaceous extinction. This enabled us to test the competing hypotheses for placental encephalization and more broadly to investigate when and how the modern placental brain and sensory repertoire emerged, and what role the end-Cretaceous extinction played.

We found that Mesozoic and Paleocene mammals had smaller brains relative to their body mass than those of Eocene stem taxa and crown orders (Fig. 1D). When expressed as a phylogenetic encephalization quotient (PEQ), a measure of relative brain size compared with its value predicted by allometry and phylogeny (*22*), Paleocene forms were less encephalized than Mesozoic and Eocene taxa (Fig. 2, A and B, and table S2). After the Paleocene low in PEQ, both average PEQ and variance significantly expanded in the Eocene, in both stem taxa and crown orders (Fig. 2, A and B). These results are robust to phylogenetic uncertainty (figs. S1 and S2 and tables S3 to S5), differences in body mass estimates (figs. S3 to S9 and tables S6 to S8), and whether placentals originated in the Cretaceous or immediately after the end-Cretaceous extinction (fig. S10). Our results are broadly in line with Jerison's "body-before-brain" hypothesis rather than the alternative but reveal an unexpected wrinkle. The earliest placentals were neither scaled up versions of Mesozoic mammals nor did they encephalize rapidly. Instead, their relative brain sizes actually decreased as they increased in body sizes more than in brain size while radiating after the end-Cretaceous extinction.

Proportions of sensory regions also changed markedly over time. The olfactory bulbs and petrosal lobules of Paleocene species and Eocene placental stem taxa essentially maintained their Mesozoic sizes, relative to both body mass and brain volume (endocranial volume) (Fig. 3, fig. S11, and table S2). The olfactory bulbs became a significantly smaller component of the brain, and the petrosal lobules became a larger component, in Eocene crown orders (Fig. 3, fig. S11, and table S2). Additionally, the neocortex enlarged over time and covered significantly more of the brain surface in Eocene crown orders versus Paleocene and Eocene stem forms (Fig. 3 and table S2). Thus, the Eocene increase in relative brain size was underpinned by expansions of the petrosal lobules and especially the neocortex, but not the olfactory bulbs. Most of these changes occurred in the Eocene crown orders, helping explain why they had significantly higher PEQs than those of their Eocene stem contemporaries, despite insignificant differences in body mass among them (Fig. 2 and table S2). These results are also robust to phylogenetic uncertainty and placental origin timing (figs. S12 to S17 and tables S3 to S5).

Phylogenetic context helps untangle the trends and tempo of these changes. Both body mass and brain volume significantly increased from the Mesozoic to the Paleocene (Fig. 2, D and F, and table S9), as did the rate of change in both measures on the phylogeny (figs. S18 and S19), regardless of placental origin time (fig. S20). However, during the Paleocene, the majority of branches exhibited faster rates of body mass increase than brain volume increase (fig. S21), and although this largely held in the Eocene, crown orders accelerated to faster relative increases in brain volume rate compared with those of Eocene stem and Paleocene taxa. This discrepancy explains why there was a drop in PEQ in the Paleocene, followed by relative brain expansion in the Eocene.

Rates of change in body mass, brain volume, and PEQ were stable across most of the Mesozoic but increased dramatically at or near the origin of Placentalia, whether it occurred in the Cretaceous (Fig. 4B and figs. S18 and S19) or immediately after the end-Cretaceous extinction (figs. S20 and S22). This makes it difficult to differentiate the effects of the extinction itself, but because both dating approaches place the key pulses at or near the origin of Placentalia, this implies that fundamental changes to the mammalian brain and its allometric relationships happened around this event, perhaps associated with changes in metabolic rate and reproductive style (*23*). The body mass, brain volume, and PEQ rate increases for early Placentalia were not followed by stasis or decline; rather, high rates continued through the Paleocene and Eocene. Brain evolutionary rate increased in placentals, although it would take until the Eocene for those persistently high rates to assemble greatly encephalized modern placental brains with relatively large neocortices and petrosal lobules and small olfactory bulbs (figs. S23 to S26).

Phylogenetic character mapping reveals lineage-specific changes in body, brain, and sensory region sizes. The Paleocene decline

[1]School of GeoSciences, University of Edinburgh, Grant Institute, Edinburgh, Scotland EH9 3FE, UK. [2]Section of Mammals, Carnegie Museum of Natural History, Pittsburgh, PA, USA. [3]New Mexico Museum of Natural History and Science, Albuquerque, NM, USA. [4]Department of Anthropology, Brooklyn College, City University of New York, Brooklyn, NY, USA. [5]Department of Anthropology, The Graduate Center, City University of New York, New York, NY, USA. [6]New York Consortium in Evolutionary Primatology, New York, NY, USA. [7]Division of Paleontology, American Museum of Natural History, New York, NY, USA. [8]Department of Earth and Environmental Sciences, Columbia University, New York, NY, USA. [9]Ecology, Evolutionary Biology, and Behavior subprogram, PhD Program in Biology, The Graduate Center, City University of New York, New York, NY, USA. [10]PhD Program in Earth and Environmental Sciences, The Graduate Center, City University of New York, New York, NY, USA. [11]Department of Biological Sciences, Rowan University, Glassboro, NJ, USA. [12]Denver Museum of Nature & Science, Denver, CO, USA. [13]National Geographic Society, Washington, DC, USA. [14]Directorate Earth and History of Life, Royal Belgian Institute of Natural Sciences, Brussels, Belgium. [15]Department of Biological Sciences, Purdue University Northwest, Westville, IN, USA. [16]Department of Integrative Biology and Museum of Paleontology, University of California, Berkeley, CA, USA.
***Corresponding author. Email: ornella.bertrand@ed.ac.uk (O.C.B.); stephen.brusatte@ed.ac.uk (S.L.B.)**

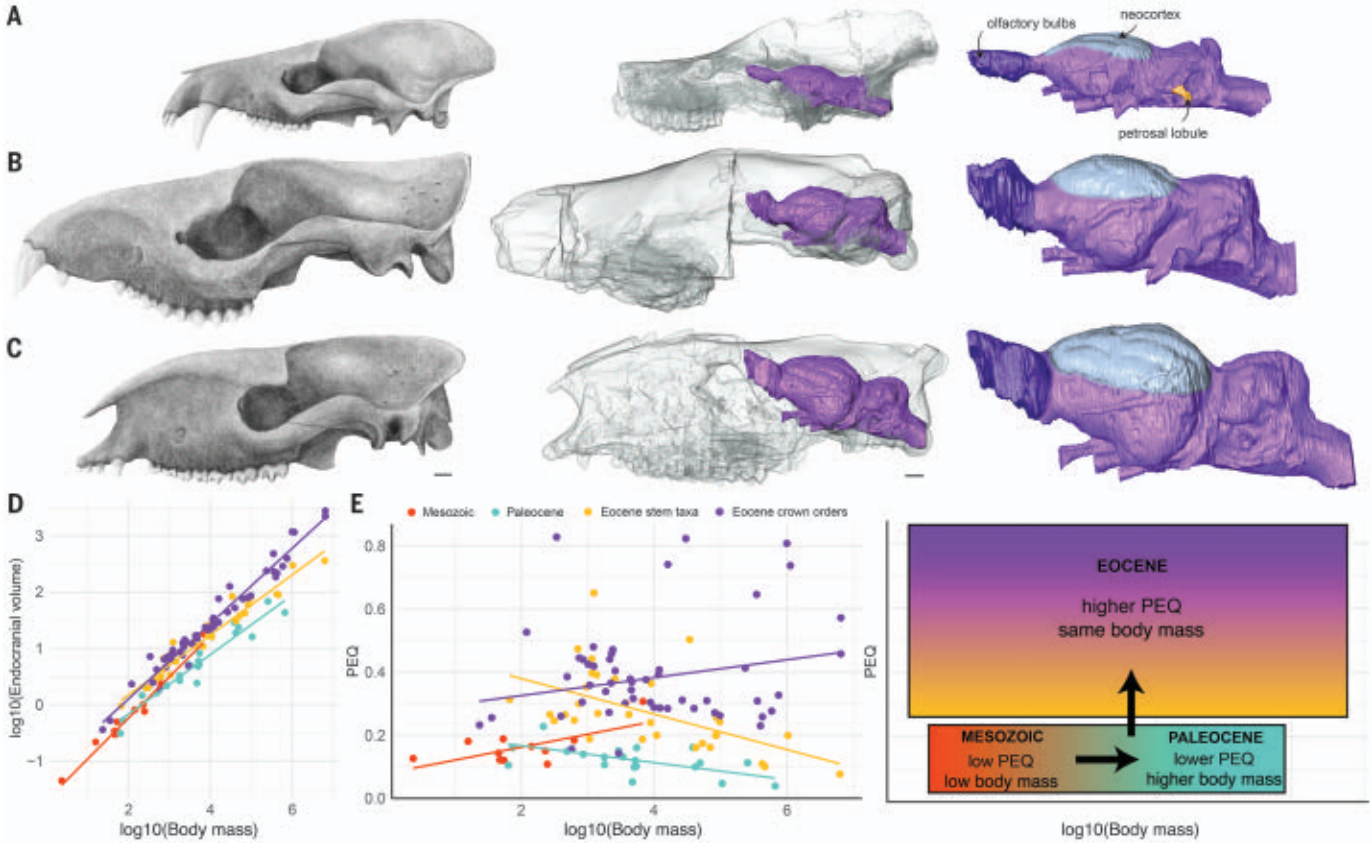


Fig. 1. Virtual endocasts of Cenozoic mammal exemplars used in this study with sensory regions highlighted and phylogenetic generalized least squares (PGLS) regressions for Mesozoic and early Cenozoic mammals. (**A**) Late Paleocene stem placental condylarth *Arctocyon primaevus* (IRSNB M2332). (**B**) Middle Eocene stem placental tillodont *Trogosus hillsii* (USNM 17157). (**C**) Middle Eocene crown placental perissodactyl *Hyrachyus modestus* (AMNH FM 12664). (**D**) Phylogenetically corrected PGLS regression of $\log_{10}$(Endocranial volume) versus $\log_{10}$(Body mass). (**E**) Phylogenetically corrected PGLS regression of PEQ versus $\log_{10}$(Body mass) and graphical summary of the results. The petrosal lobules are absent in (B) and (C). Scale bar, 10 mm. Body mass was measured in grams, and endocranial volume was measured in cubic centimeters.

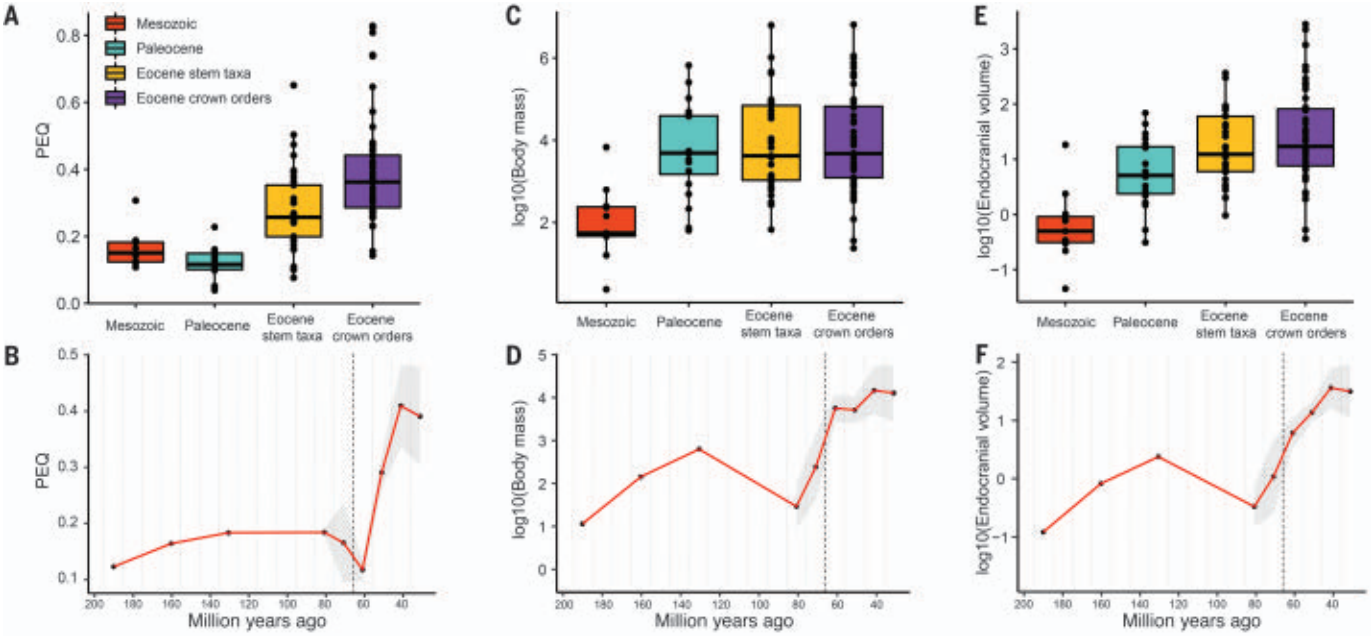


Fig. 2. PEQ, endocranial volume, and body mass of Mesozoic and early Cenozoic mammals. (**A**) Boxplot of PEQ for Mesozoic, Paleocene, Eocene stem taxa, and Eocene crown orders. (**B**) PEQ average through time per 10-million-year bins. (**C**) Boxplot of $\log_{10}$(Body mass). (**E**) Boxplot of $\log_{10}$(Endocranial volume) for groups in (A). (**D**) $\log_{10}$(Body mass) average through time per 10-million-year bins for groups in (A). (**F**) $\log_{10}$(Endocranial volume) average through time per 10-million-year bins. Body mass was measured in grams, and endocranial volume was measured in cubic centimeters.

in PEQ was the net result of several independent decreases in relative brain volume, largely because of independent increases in body mass in archaic groups such as "condylarths," pantodonts, and taeniodonts (Fig. 4 and figs. S18 to S20, S22, and S27 to S29). Then in the Eocene, there were several independent increases in PEQ, particularly in crown orders such as artiodactyls (including cetaceans), perissodactyls, carnivorans, euprimates, and rodents (Fig. 4 and figs. S18 to S20, S22, and S27 to S29). Omnivores and carnivores were statistically indistinguishable in PEQ compared with contemporary herbivores in the Paleocene, but both guilds significantly increased PEQ in the Eocene, with omnivores and carnivores eclipsing herbivores (fig. S30 and table S10). The two most speciose placental subclades experienced different fates: Raw body mass increased markedly in Paleocene-Eocene laurasiatherians (carnivorans, perissodactyls, and artiodactyls) but not in euarchontoglirans (rodents and primates). Both subclades underwent substantial encephalization (Fig. 4), but rates of change in both body mass and brain volume were higher in laurasiatherians (Fig. 4 and figs. S18 to S20 and S22).

Small relative brain size after the end-Cretaceous extinction—even more so a decline in encephalization—is surprising and counter to the convention that mammal brains have steadily gotten larger over time [(*10*–*13*), but

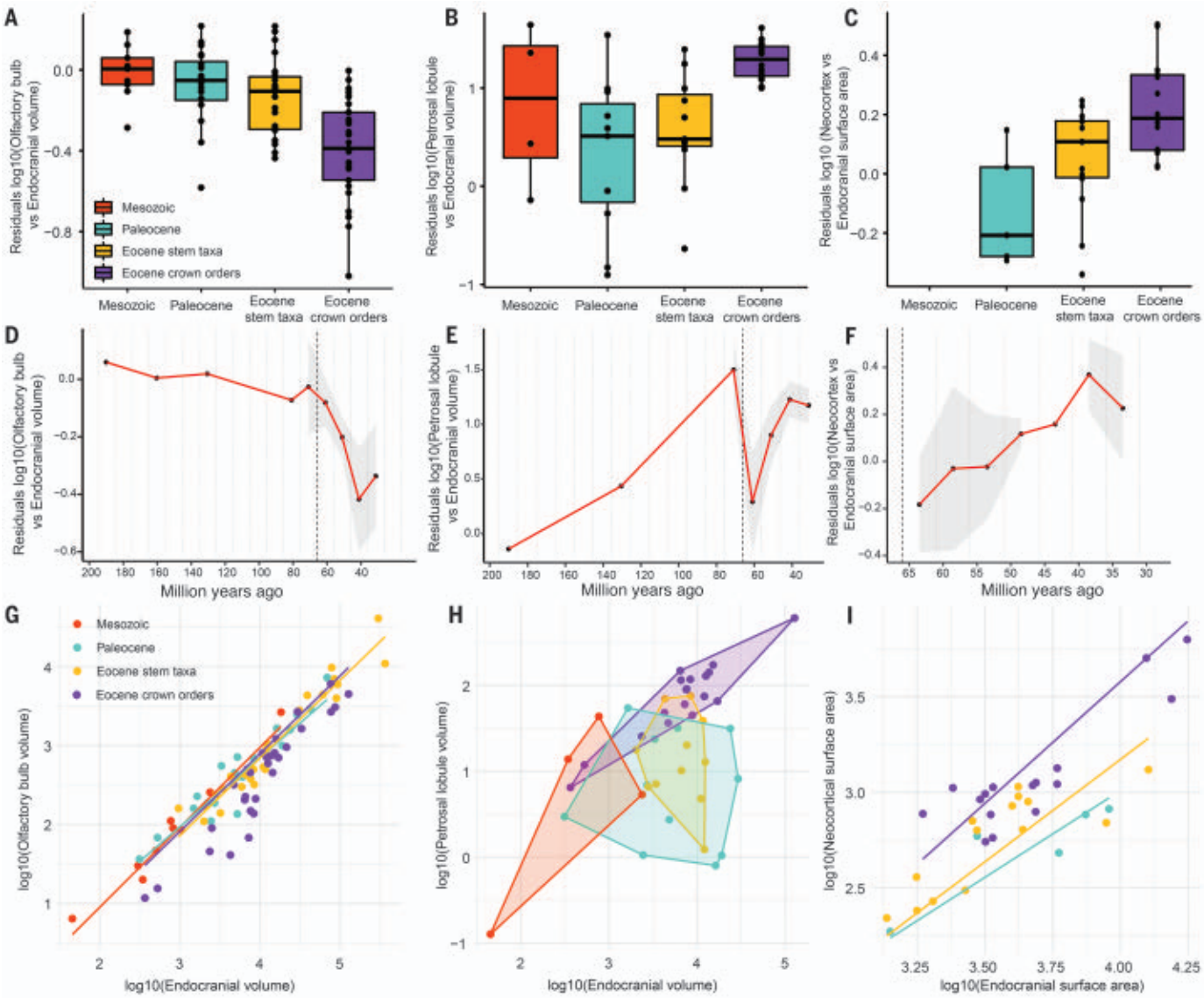


Fig. 3. Relative size of the olfactory bulbs, petrosal lobules, and neocortex of Mesozoic and early Cenozoic mammals. (**A**) Boxplot of the residuals from a PGLS regression of $\log_{10}$(Olfactory bulb volume versus Endocranial volume) for Mesozoic, Paleocene, Eocene stem taxa, and Eocene crown orders. (**B**) Boxplot of the residuals from a PGLS regression of $\log_{10}$(Petrosal lobule volume versus Endocranial volume) for the groups in (A). (**C**) Boxplot of the residuals from a PGLS regression of $\log_{10}$(Neocortical surface area versus Endocranial surface area) for the groups in (A). (**D**) Residuals of $\log_{10}$(Olfactory bulb volume versus Endocranial volume) average through time per 10-million-year bins. (**E**) Residuals of $\log_{10}$(Petrosal lobule volume versus Endocranial volume) average through time per 10-million-year bins. (**F**) Residuals of $\log_{10}$(Neocortical surface area versus Endocranial surface area) average through time per 5-million-year bins. (**G**) PGLS regression of $\log_{10}$(Olfactory bulb volume versus Endocranial volume) for the groups in (A). (**H**) PGLS regression of $\log_{10}$(Petrosal lobule volume versus Endocranial volume) for the groups in (A). (**I**) PGLS regression of $\log_{10}$(Neocortical surface area versus Endocranial surface area) for the groups in (A). Mesozoic neocortical data were unavailable. Volumetric measurements are in cubic millimeters, and surface area measurements are in square millimeters.

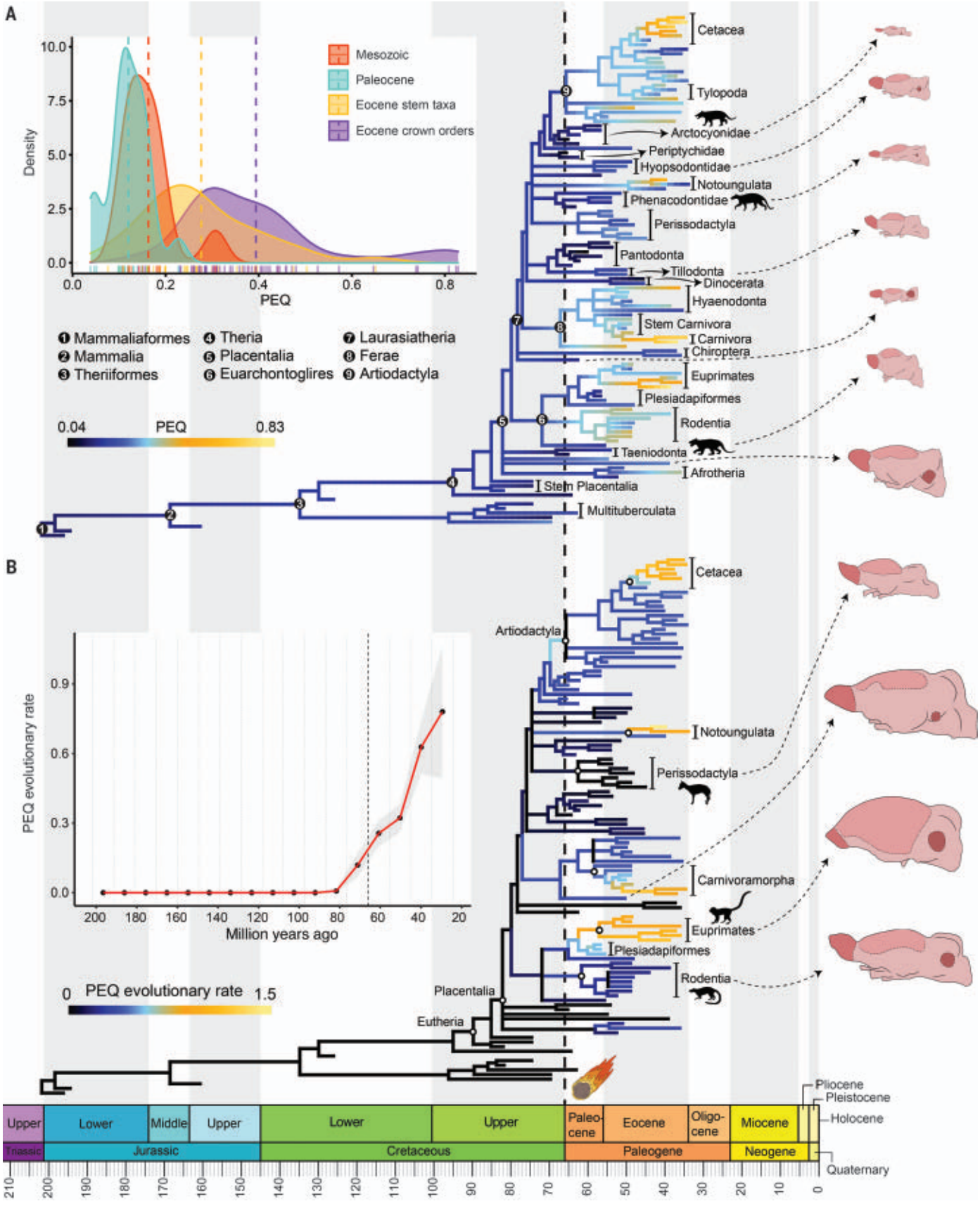


Fig. 4. PEQ evolutionary change across time and phylogeny in Mesozoic and early Cenozoic mammals. (**A**) PEQ density plot and ancestral state reconstruction mapped onto a phylogeny for Mesozoic, Paleocene, and Eocene mammals. (**B**) PEQ evolutionary rate mapped onto a phylogenetic

tree and averaged through time per 10-million-year bins. Virtual endocasts illustrated to the right with highlighted brain regions are scaled in terms of their PEQ and are, from top to bottom, *Arctocyon primaevus* (IRSNB M2332), *Hyopsodus paulus* (USNM 17980), *Meniscotherium chamense* (USNM 22673), *Trogosus hillsii* (USNM 17157), *Acmeodon secans* (KU 7912), *Ectoganus copei* (USNM 12714), *Leptictis* sp. (AMNH 62369), *Hyrachyus modestus* (AMNH FM 12664), *Metacheiromys marshi* (USNM 452349), *Smilodectes gracilis* (UM 32773), and *Paramys delicatus* (AMNH 12506). [Silhouettes are from http://phylopic.org (Euprimates, Perissodactyla, and Rodentia).]

see (*7*)]. Although relative brain size is but one aspect of cognition, and we lack information on neuron density and connectivity in virtually all fossil mammals, experimental evidence indicates that extant mammals with relatively larger brains are better problem solvers (*24*). There also is ample evidence that larger relative brain size is linked to greater behavioral flexibility and capacity to cope with new or altered environments (*25*–*28*). All of these skills, presumably, would have been beneficial after a mass extinction, yet the placentals that diversified as devastated ecosystems recovered and new food webs emerged did not develop relatively larger brains, which suggests that other factors were behind their radiation.

It appears that the postextinction placental radiation emphasized changes to body size rather than brain size, as survivors proliferated to fill vacant niches, no longer constrained to shrew to badger sizes by incumbent dinosaurs (*15*). Brain size increased as a result of body size expansion, and because there were no marked relative increases in key sensory regions (olfactory bulbs and petrosal lobules), the growth was primarily in regions that permit control of larger bodies (brain stem, diencephalon, and striatum) (*29*, *30*). This suggests that selection acted differently on brain and body size and adds to growing evidence that body size shifts, rather than pronounced alterations in brain size, drove much of the variation in mammalian encephalization (*7*, *31*). At the very least, relatively enlarged brains were not necessary for Paleocene placentals.

As the Paleocene transitioned to the Eocene, the mode of placental brain and body evolution shifted. Relative brain size greatly increased, in both average and variance, signaling a new regime in which brain size changes were more paramount than those in body size. Brains were not only becoming larger; growth was focused in regions involved in advanced senses as these mammals added better balance, vision, eye movement, head control, and sensory integration to their preexisting keen olfaction, thus greatly expanding the placental sensory toolkit. This corroborates arguments that mosaic evolution of brain regions, not merely changes in overall size, underlie the adaptive potential of the mammalian brain—not only across phylogeny (*32*) but also in deep time. Primarily, encephalization and sensory enhancements occurred in the crown orders, with predators developing significantly larger relative brain size than those of prey groups as ecosystems saturated. It was these crown placentals—among them the first horses, whales, dogs, bats, and euprimates—that achieved high PEQs, whereas the smaller-brained and more olfactory-driven stem taxa waned, with groups once so instrumental in the end-Cretaceous recovery and initial placental radiation—such as condylarths and pantodonts—ultimately succumbing to extinction.

REFERENCES AND NOTES

1. H. J. Jerison, *Evolution of the Brain and Intelligence* (Academic Press, 1973).
2. M. Tsuboi *et al*., *Nat. Ecol. Evol.* **2**, 1492–1500 (2018).
3. T. B. Rowe, T. E. Macrini, Z. X. Luo, *Science* **332**, 955–957 (2011).
4. C. J. Burgin, J. P. Colella, P. L. Kahn, N. S. Upham, *J. Mammal.* **99**, 1–14 (2018).
5. J. R. Burger, M. A. George Jr., C. Leadbetter, F. Shaikh, *J. Mammal.* **100**, 276–283 (2019).
6. B. Finlay, in *Encyclopedia of Neuroscience* (2009), vol. 2, pp. 337–345.
7. J. B. Smaers *et al*., *Sci. Adv.* **7**, eabe2101 (2021).
8. L. Radinsky, *Science* **194**, 626–627 (1976).
9. H. J. Jerison, *Biolinguistics* **6**, 383–392 (2012).
10. E. Lartet, *Comptes Rendus des Séances de l'Académie des Sciences* **66**, 1119–1122 (1868).
11. O. C. Marsh, *Ann. Mag. Nat. Hist.* **14**, 167–167 (1874).
12. H. J. Jerison, *Science* **133**, 1012–1014 (1961).
13. H. J. Jerison, *Science* **170**, 1224–1225 (1970).
14. D. M. Grossnickle, S. M. Smith, G. P. Wilson, *Trends Ecol. Evol.* **34**, 936–949 (2019).
15. J. Alroy, *Syst. Biol.* **48**, 107–118 (1999).
16. F. A. Smith *et al*., *Science* **330**, 1216–1219 (2010).
17. G. G. Simpson, *Biol. Rev. Camb. Philos. Soc.* **12**, 1–46 (1937).
18. K. D. Rose, *The Beginning of the Age of Mammals* (Johns Hopkins Univ. Press, 2006).
19. T. E. Williamson, *The Beginning of the Age of Mammals in the San Juan Basin, New Mexico: Biostratigraphy and Evolution of Paleocene Mammals of the Nacimiento Formation: Bulletin 8* (New Mexico Museum of Natural History and Science, 1996), vol. 8.
20. T. R. Lyson *et al*., *Science* **366**, 977–983 (2019).
21. H. Rambold, A. Churchland, Y. Selig, L. Jasmin, S. G. Lisberger, *J. Neurophysiol.* **87**, 912–924 (2002).
22. X. Ni, J. J. Flynn, A. R. Wyss, C. Zhang, *Sci. Adv.* **5**, eaav7913 (2019).
23. B. K. McNab, *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **151**, 5–28 (2008).
24. S. Benson-Amram, B. Dantzer, G. Stricker, E. M. Swanson, K. E. Holekamp, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2532–2537 (2016).
25. D. Sol, S. Bacher, S. M. Reader, L. Lefebvre, *Am. Nat.* **172** (Suppl. 1), S63–S71 (2008).
26. D. Sol, *Biol. Lett.* **5**, 130–133 (2009).
27. J. M. Ratcliffe, M. B. Fenton, S. J. Shettleworth, *Brain Behav. Evol.* **67**, 165–176 (2006).
28. J. T. van Woerden, E. P. Willems, C. P. van Schaik, K. Isler, *Evolution* **66**, 191–199 (2012).
29. S. Herculano-Houzel, K. Catania, P. R. Manger, J. H. Kaas, *Brain Behav. Evol.* **86**, 145–163 (2015).
30. S. Herculano-Houzel, *Curr. Opin. Behav. Sci.* **16**, 1–7 (2017).
31. J. B. Smaers, D. K. N. Dechmann, A. Goswami, C. Soligo, K. Safi, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18006–18011 (2012).
32. R. A. Barton, P. H. Harvey, *Nature* **405**, 1055–1058 (2000).
33. O. C. Bertrand, Bertrand-Ornella/Brain_Evolution_Placental: Brain evolution in early Placental mammals. *Zenodo* (*2022*).

ACKNOWLEDGMENTS

We thank F. Mao, M. Hill Chase, R. O'Leary, M. Novacek, and S. Goldberg for scanning specimens and logistic support at the AMNH; A. Harrington and D. Boyer for providing access to several Euprimates virtual endocasts accessioned in Morphosource; S. Hoffman for providing access to the virtual endocast of *Vintana*; U. Lefèvre and the IRSNB for scanning *Arctocyon* and *Orthaspidotherium*; S. Anderson, P. dePolo, N. Egi, S. Holpin, and Z. Kynigopoulou for supplying postcranial measurements for several Paleocene and Eocene specimens; and T. Stubbs, B. Moon, and A. Meade for fruitful discussions about the evolutionary rates methods. We thank two anonymous reviewers for helpful comments that improved this manuscript. **Funding:** This work was supported by Marie Skłodowska-Curie Actions: Individual Fellowship, H2020-MSCA-IF-2018-2020, no. 792611 (O.C.B.); European Research Council (ERC) Starting Grant 756226 under the European Union's Horizon 2020 Research and Innovation Programme (S.L.B.); Philip Leverhulme Prize (S.L.B.); NSF DEB 1654952 (T.E.W. and S.L.B.); NSF DEB 1654949 (J.R.W. and S.L.S.); NSF DEB1456826 (L.T.H. and S.G.B.C.); Belgian Science Policy Office project BRAIN BR/121/A3/PalEurAfrica (T.S.); Lisa Levin Appel Family Foundation, M. Cleworth, Lyda Hill Philanthropies, David B. Jones Foundation, M. L. and S. R. Kneller, T. and K. Ryan, and J. R. Tucker as part of the Denver Museum of Nature & Science No Walls Community Initiative (T.R.L. and I.M.M.); and National Agency for Research and Development (ANID)/PFCHA/Doctorado en el extranjero Becas Chile/2018-72190003 (H.P.P.). Funding to J.J.F. and for scanning American Museum specimens was provided by the Microscopy and Imaging Facility, American Museum of Natural History Frick Fund, Division of Paleontology, American Museum of Natural History, and US National Science Foundation (DEB-1257572).
Author contributions: Conceptualization: O.C.B., S.L.B., S.L.S., T.E.W., and J.R.W. Methodology: O.C.B., S.L.B., S.L.S., J.J.F., Z.J.T., and H.P.P. Formal analyses: O.C.B., S.L.S., Z.J.T., and H.P.P. Investigation: O.C.B., S.L.B., S.L.S., T.E.W., and J.R.W. Results acquisition: O.C.B., S.L.B., S.L.S., T.E.W., and J.R.W. Raw data acquisition (specimens and CT data): O.C.B., S.L.B., S.L.S., T.E.W., J.R.W., S.G.B.C., J.J.F., L.T.H., T.R.L., J.M., I.M.M., T.S., M.S., and Z.J.T. Supervision: S.L.B. Writing, original draft: O.C.B. and S.L.B. Writing, review and editing: all authors. Visualization: O.C.B. and S.L.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Code and data to run analyses using topology 1 (unconstrained) is provided with this article and is also available at (*33*). Raw data are available in the supplementary materials. All fossils are accessioned in public collections and, upon curator approval, should be available for study by qualified researchers. All (other) data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials. Raw CT scans of American Museum of Natural History specimens are available from the American Museum of Natural History under a materials transfer agreement with that institution. The 3D models of the brains will be stored in Morphosource at www.morphosource.org/projects/0000C1223?locale=en. Access to the 3D models will be approved by O.C.B. The endocasts cannot be described or figured without permission from O.C.B. or S.L.B.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl5584
Materials and Methods
Supplementary Text
Figs. S1 to S35
References (*34*–*295*)
MDAR Reproducibility Checklist
Tables S1 to S24
Data File S1

21 July 2021; accepted 1 February 2022
10.1126/science.abl5584

CHRONIC PAIN

A spinal microglia population involved in remitting and relapsing neuropathic pain

Keita Kohno[1,2]†, Ryoji Shirasaka[2]†, Kohei Yoshihara[2], Satsuki Mikuriya[1], Kaori Tanaka[3], Keiko Takanami[4,5], Kazuhide Inoue[6], Hirotaka Sakamoto[4], Yasuyuki Ohkawa[3], Takahiro Masuda[1,2], Makoto Tsuda[1,2]*

Neuropathic pain is often caused by injury and diseases that affect the somatosensory system. Although pain development has been well studied, pain recovery mechanisms remain largely unknown. Here, we found that CD11c-expressing spinal microglia appear after the development of behavioral pain hypersensitivity following nerve injury. Nerve-injured mice with spinal CD11c$^+$ microglial depletion failed to recover spontaneously from this hypersensitivity. CD11c$^+$ microglia expressed insulin-like growth factor-1 (IGF1), and interference with IGF1 signaling recapitulated the impairment in pain recovery. In pain-recovered mice, the depletion of CD11c$^+$ microglia or the interruption of IGF1 signaling resulted in a relapse in pain hypersensitivity. Our findings reveal a mechanism for the remission and recurrence of neuropathic pain, providing potential targets for therapeutic strategies.

Damage to the nervous system alters the somatosensory system, which can result in the development of neuropathic pain. These alterations occur in neurons and in non-neuronal cells, especially microglia, a type of resident immune cells. In response to peripheral nerve injury (PNI, a model of neuropathic pain), microglia in the spinal dorsal horn (SDH) respond rapidly, with alterations in morphology, cell number, and gene expression levels, processes that are necessary for pain development (*1*, *2*). However, microglia are highly plastic cells that exist in heterogeneous states characterized by distinct gene expression, and microglial populations can change during the course of neuronal development and disease (*3*–*5*).

Here, we focused on microglia characterized by the expression of integrin αX (ITGAX, also known as CD11c) (*6*) because CD11c$^+$ microglia in the brain increase in several neurological disease models (*7*). By using *Itgax-Venus* mice, in which CD11c$^+$ cells express the fluorescent protein Venus, we found that Venus$^+$ cells in the SDH were rarely observed in naive and sham-operated mice (fig. S1A) but increased after PNI (Fig. 1A and fig. S1B). Venus expression was highly restricted to cells with ionized calcium-binding adapter molecule 1 (IBA1)

[1]Department of Life Innovation, Graduate School of Pharmaceutical Sciences, Kyushu University, Fukuoka, Japan. [2]Department of Molecular and System Pharmacology, Graduate School of Pharmaceutical Sciences, Kyushu University, Fukuoka, Japan. [3]Division of Transcriptomics, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan. [4]Ushimado Marine Institute, Graduate School of Natural Science and Technology, Okayama University, Setouchi, Japan. [5]Mouse Genomics Resources Laboratory, National Institute of Genetics, Shizuoka, Japan. [6]Kyushu University Institute for Advanced Study, Fukuoka, Japan.
*Corresponding author. Email: tsuda@phar.kyushu-u.ac.jp
†These authors equally contributed to this work.

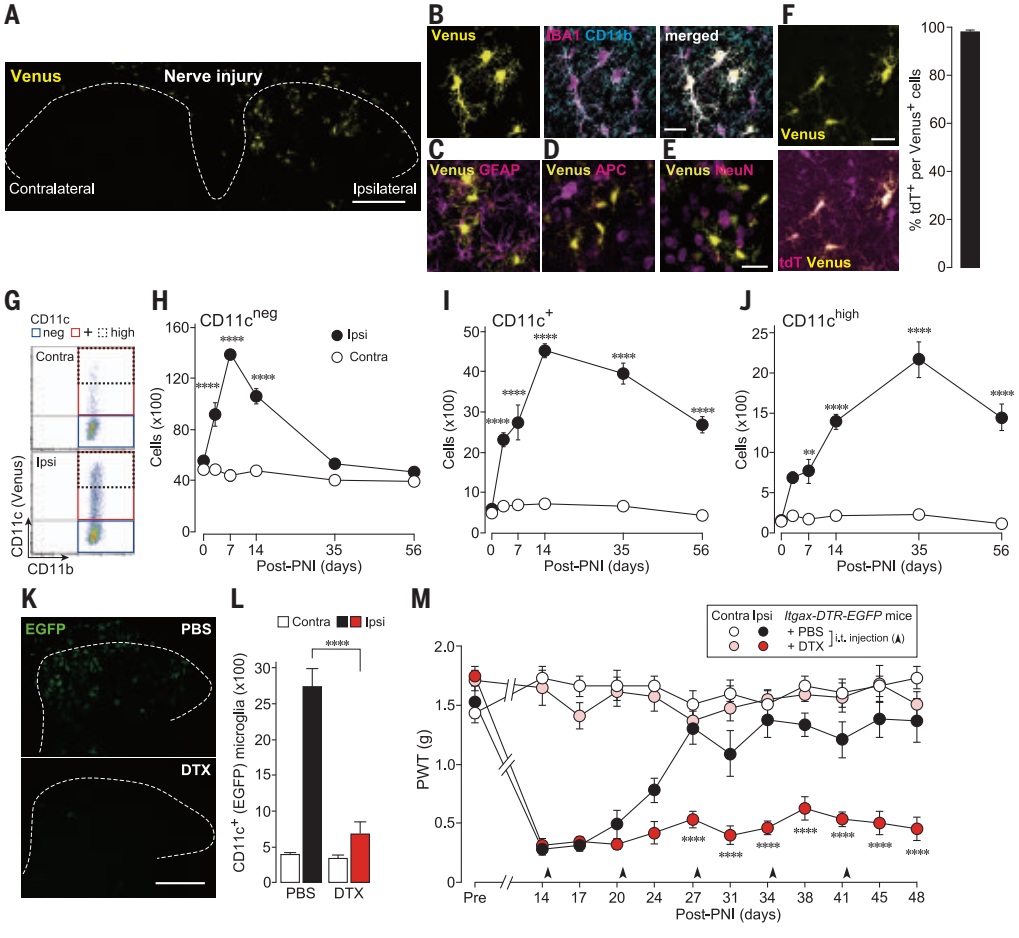


Fig. 1. CD11c$^+$ spinal microglia are necessary for the remission of pain hypersensitivity after PNI. (**A**) Venus fluorescence in the SDH of *Itgax-Venus* mice 14 days after PNI. (**B** to **E**) Venus and cell-type marker immunostaining: (B) IBA1/CD11b, myeloid markers including microglia; (C) glial fibrillary acidic protein (GFAP), an astrocyte marker; (D) adenomatous polyposis coli (APC), an oligodendrocyte marker; and (E) neuronal nuclei (NeuN), a neuron marker. (**F**) Venus and tdT fluorescence in the SDH of *Hexb*$^{tdT/tdT}$;*Itgax-Venus* mice on day 14. Shown is the percentage of tdT$^+$ cells per total Venus$^+$ cells (n = 4 mice). (**G** to **J**) Temporal changes in the number of CD11c^{neg} (H), CD11c$^+$ (I), and CD11c^{high} (J) spinal microglia after PNI (n = 3 to 5 mice) (flow cytometry). (**K** and **L**) EGFP in the SDH of *Itgax-DTR-EGFP* mice 2 days after intrathecal injection of PBS or DTX (0.5 ng) on day 14 (K). Also shown is flow cytometric quantification of CD11c$^+$ (EGFP) microglia (n = 5 mice) (L). (**M**) PWT of *Itgax-DTR-EGFP* mice injected with PBS or DTX before (Pre) and after PNI (n = 7 to 8 mice). Scale bars, 200 μm [(A) and (K)] and 20 μm (B). Data are shown as means ± SEM. $^{**}P < 0.01$ and $^{****}P < 0.0001$. For more details, see the extended caption in the supplementary materials.

and CD11b (Fig. 1, B to E), the purinergic receptor P2Y12R (*3*, *7*) (fig. S1C), and tandem Tomato (tdT) (Fig. 1F) in *Hexb*$^{tdT/tdT}$ mice, which enables the visualization of microglia (*8*). Venus^{+} cells were observed in a portion of tdT^{+} (Fig. 1F) and P2Y12R^{+} microglia (fig. S1C). In addition, data from bone marrow–chimera mice (fig. S1, D and E) and splenectomized mice (fig. S1F) excluded an involvement of peripherally derived cells in the appearance of Venus^{+} cells in the SDH after PNI. Thus, we hereafter referred to Venus^{+} SDH cells as CD11c^{+} microglia. Flow cytometry analyses (Fig. 1, G to J) revealed that CD11c^{+} microglia peaked at day 14 after PNI and were observed until the last time point tested (day 56) (Fig. 1I), a temporal pattern that was in contrast with CD11c^{neg} microglia (Fig. 1H). Moreover, CD11c^{high} microglia (Fig. 1G and fig. S1G) peaked on day 35 after PNI (Fig. 1J), a time point that the paw withdrawal threshold (PWT) to mechanical stimuli returned to the basal level (fig. S1H). Thus, CD11c^{+} microglia appeared in the SDH after the development of PNI-induced pain hypersensitivity and remained even after the hypersensitivity had remitted.

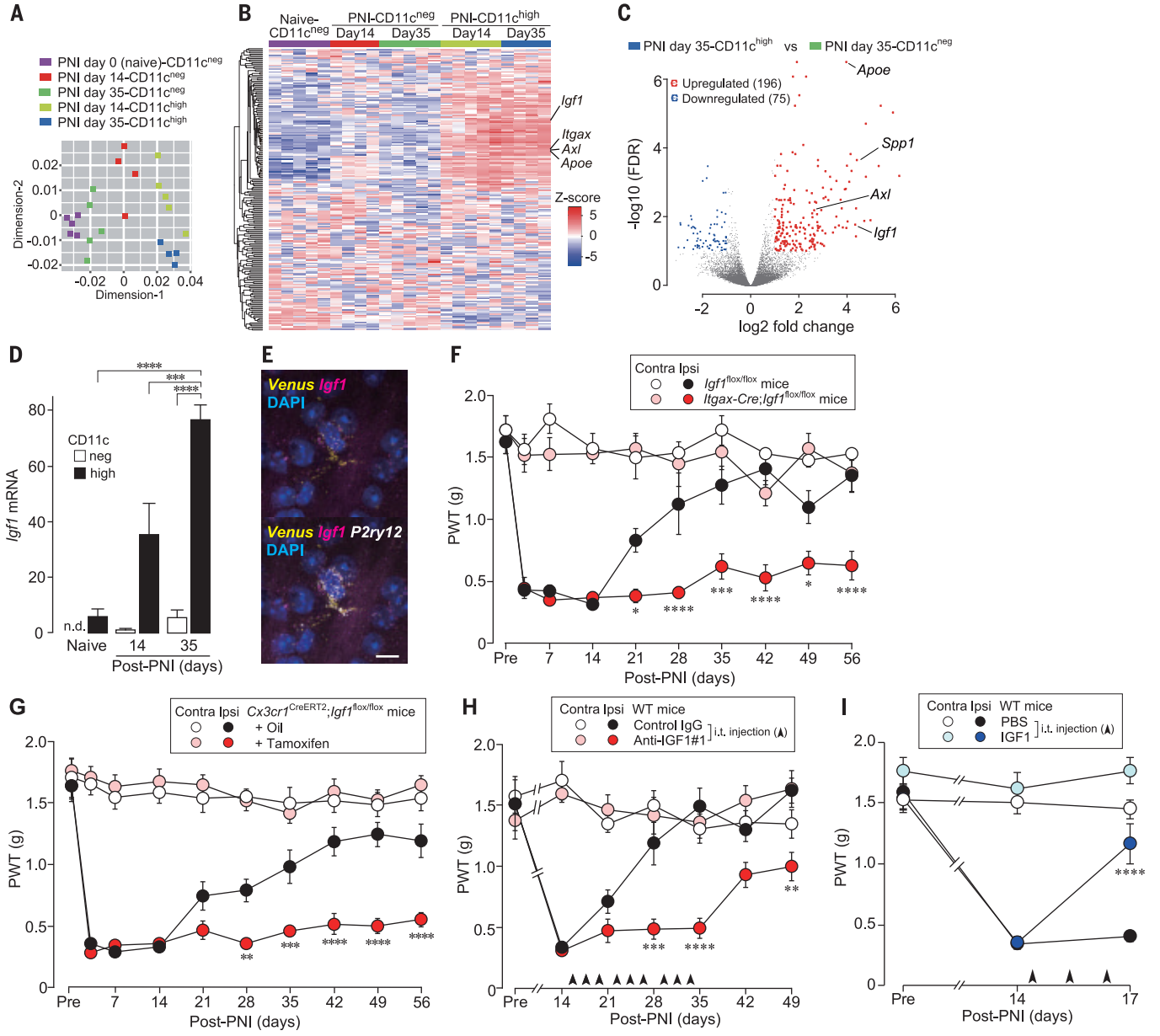


Fig. 2. IGF1 is a CD11c^{+} microglial factor required for recovery from pain hypersensitivity. (**A**) Multidimensional scaling plot representing relationships between CD11c^{neg} and CD11c^{high} microglia (n = 4 to 5 mice). Each point represents a single sample. (**B**) Heatmap of the top 200 DEGs in microglia for the different conditions. Each row represents a single sample. DEGs are listed in table S1. (**C**) Volcano plot showing the average log2-fold change and −log10 false discovery rate (FDR) for all DEGs between CD11c^{neg} and CD11c^{high} microglia on day 35. DEGs are listed in table S2. (**D**) qPCR analysis of *Igf1* mRNA in FACS-sorted CD11c^{neg} and CD11c^{high} SDH microglia before (Naive) and after PNI (n = 4 mice). (**E**) In situ hybridization of mRNAs of *Venus*, *Igf1*, and *P2ry12* in the SDH on day 35. Scale bar, 20 μm. (**F** to **H**) PWTs of *Itgax-Cre*;*Igf1*$^{flox/flox}$ mice (n = 5 mice) (F), *Cx3cr1*^{CreERT2};*Igf1*$^{flox/flox}$ mice (n = 8 or 9 mice) (G), and IGF1-neutralizing antibody (anti-IGF1#1)–treated WT mice (n = 6 to 7 mice) (H). (**I**) Reversal of PNI-induced pain hypersensitivity by intrathecal injection of IGF1 (n = 6 to 8 mice). Data are shown as mean ± SEM. *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001. For more details, see the extended caption in the supplementary materials.

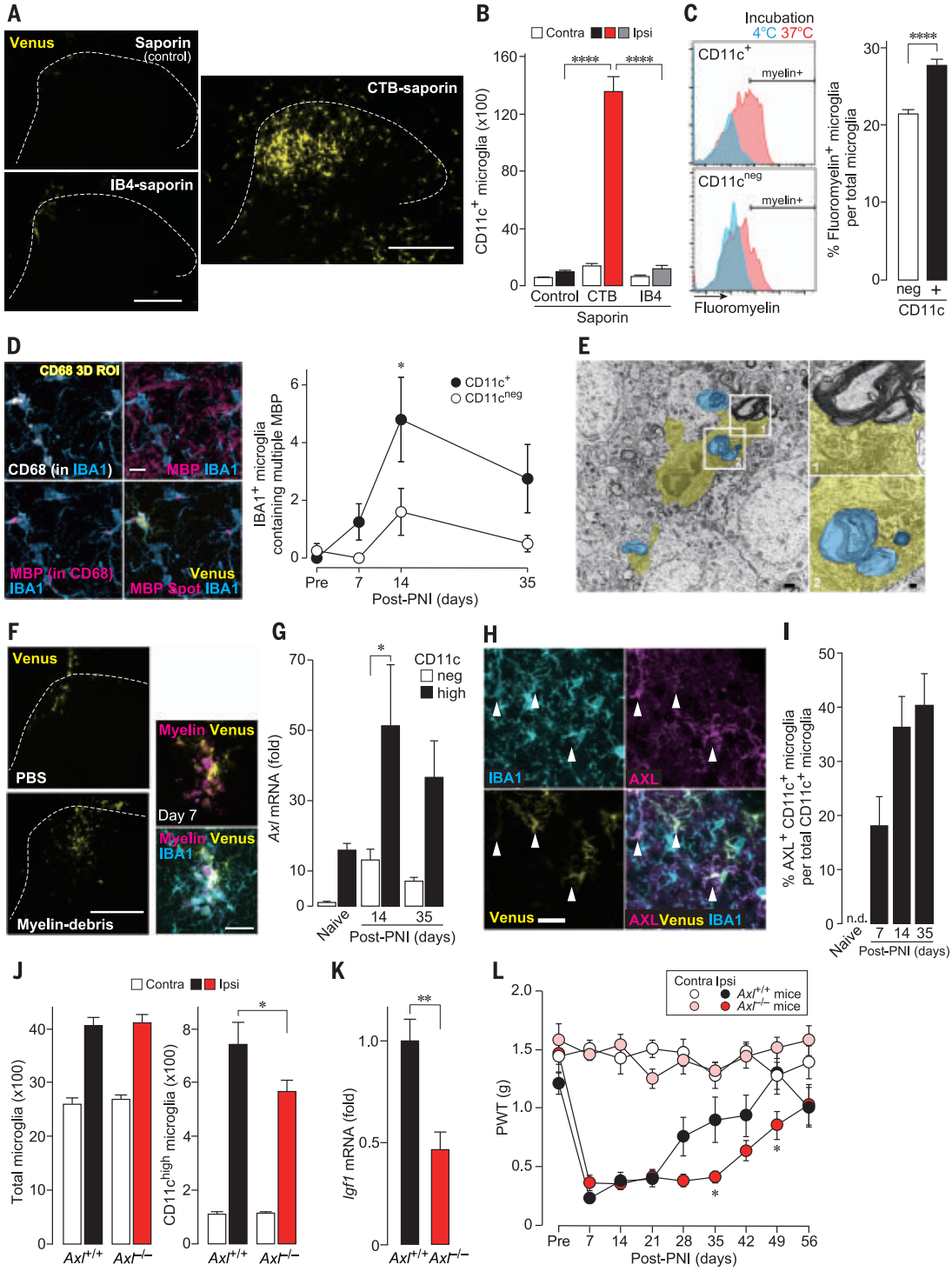


Fig. 3. Appearance of CD11c$^+$ SDH microglia involves AXL. (**A**) Venus in the SDH of *Itgax-Venus* mice on day 14 after injection of CTB-saporin or IB4-saporin. (**B**) CD11c$^+$ microglia after saporin injection (n = 3 to 4 mice) (flow cytometry). (**C**) Phagocytosis of myelin debris by cultured spinal microglia from *Itgax-Venus* mice with PNI (day 14). Percentage of fluoromyelin$^+$ microglia (n = 7 wells). (**D**) IBA1, CD68, and MBP and Venus in the SDH on day 14. IBA1$^+$ cells including multiple MBP signals within CD68$^+$ lysosomes were counted (n = 4 to 5 mice). (**E**) Electron microscopy image of Venus (yellow) and myelin particles (blue) in the SDH (day 14). (**F**) Venus in the SDH on days 14 (left) or 7 (right) after intraspinal injection of myelin debris. (**G**) *Axl* mRNA in sorted CD11c^{neg} and CD11c^{high} SDH microglia after PNI (n = 4 mice) (qPCR). (**H** and **I**) Representative images for AXL and Venus in the SDH on day 14 (arrowheads: AXL$^+$Venus$^+$ microglia). Percentage of AXL$^+$CD11c$^+$ microglia (n = 3 to 5 mice). (**J**) Total (left) and CD11c^{high} microglia (right) in the SDH of *Axl*$^{-/-}$ mice on day 35 (n = 10 or 11 mice). (**K**) *Igf1* mRNA in sorted microglia in the ipsilateral SDH of *Axl*$^{-/-}$ mice on day 35 (n = 7 mice) (qPCR). (**L**) PWT of *Axl*$^{-/-}$ mice (n = 10 mice). Data are shown as mean ± SEM. *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001. Scale bars, 200 μm [(A) and (F), left], 20 μm [(F), right, and (H)], 10 μm (D), 1 μm [(E), left)], and 200 nm [(E), right]. For more details, see the extended caption in the supplementary materials.

To test the role of CD11c$^+$ microglia, we used a cell-depletion strategy in *Itgax-DTR-EGFP* mice, in which CD11c$^+$ cells expressed diphtheria toxin receptors (DTRs) and enhanced green fluorescent protein (EGFP) (fig. S2). Intrathecal injection of diphtheria toxin (DTX) depleted EGFP$^+$ cells (CD11c$^+$ microglia) (Fig. 1, K and L, and fig. S2). In a behavioral test 14 days after PNI, phosphate-buffered saline (PBS)–treated *Itgax-DTR-EGFP* mice (Fig. 1M) and DTX-treated wild-type (WT) mice (fig. S3) displayed markedly decreased PWTs, which gradually increased. Mice with CD11c$^+$ spinal microglia depletion from day 14 failed to display spontaneous recovery from pain hypersensitivity (Fig. 1M and fig. S4). By contrast, CD11c$^+$ microglial depletion on day 7 had no effect (fig. S5). In addition, splenectomy did not change the PNI-induced pain (fig. S6). Thus, CD11c$^+$ microglia played a role in the remission of neuropathic pain.

RNA sequencing of SDH microglia of *Itgax-Venus* mice with or without PNI revealed a distinct transcriptional signature of CD11c^{neg} and CD11c^{high} microglia (Fig. 2, A and B, and table S1). The differentially expressed genes (DEGs) in CD11c^{high} SDH microglia on day 35

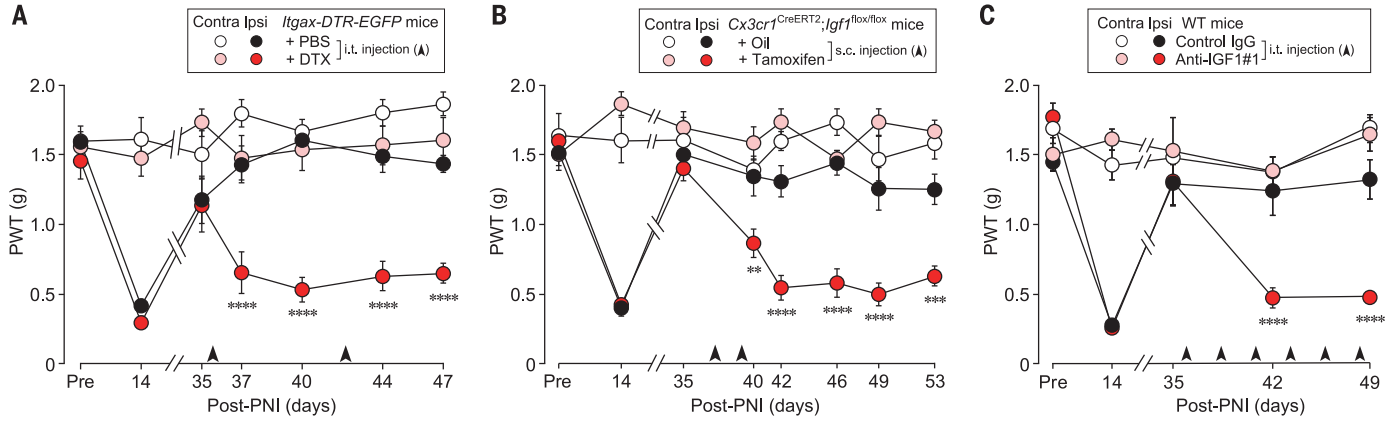


Fig. 4. Depleting CD11c$^+$ microglia relapses PNI-induced pain hypersensitivity. (**A** to **C**) Recurred pain hypersensitivity in mice with depletion of CD11c$^+$ microglia (A), with microglia-selective IGF1-knockout (B), and with spinal IGF1 neutralization (C). (A) DTX or PBS was intrathecally injected into *Itgax-DTR-EGFP* mice at days 35 and 42 after PNI (n = 7 mice). (B) Tamoxifen (4 mg) or corn oil (control) were subcutaneously injected into *Cx3cr1*^{CreERT2};*Igf1*$^{flox/flox}$ mice at days 37 and 39 after PNI (n = 7 mice). (C) Anti-IGF1#1 was intrathecally injected into WT mice from day 35 after PNI (n = 6 mice). Data are shown as mean ± SEM. $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$. For more details, see the extended caption in the supplementary materials.

after PNI included insulin-like growth factor 1 (*Igf1*) (Fig. 2C and table S2). Quantitative polymerase chain reaction (qPCR) analysis confirmed that *Igf1* mRNA was increased in CD11c$^+$ microglia after PNI and peaked on day 35 (Fig. 2D and fig. S7). *Igf1* mRNA was detected in CD11c$^+$ SDH microglia by in situ hybridization (Fig. 2E). To determine the role of IGF1, we used *Itgax-Cre*;*Igf1*$^{flox/flox}$ mice to knock out IGF1 in CD11c$^+$ cells. After PNI, these mice also showed no recovery from pain hypersensitivity (Fig. 2F). Similar results were obtained from tamoxifen-treated *Cx3cr1*^{CreERT2};*Igf1*$^{flox/flox}$ mice (Fig. 2G), enabling deletion of the *Igf1* gene in CX3CR1$^+$ cells (*9*), including all microglia and central nervous system (CNS)–associated macrophages (fig. S8) (*10*). The repeated intrathecal administration of an IGF1-neutralizing antibody, anti-IGF1#1, to PNI mice prevented pain recovery (Fig. 2H). A similar effect was obtained using another IGF1-neutralizing antibody, anti-IGF1#2 (fig. S9). The repeated intrathecal administration of recombinant IGF1 to PNI mice accelerated pain recovery (Fig. 2I). IGF1 was also observed in neurons (fig. S10, A to F), but *Igf1* knockdown in SDH neurons did not change PNI-induced hypersensitivity (fig. S10, G and H). Thus, IGF1 derived from CD11c$^+$ microglia was necessary for the spontaneous remission of neuropathic pain.

From the data showing no increase in CD11c$^+$ SDH microglia in a model of inflammatory pain (fig. S11, A and B), the mechanism underlying the PNI-induced appearance of CD11c$^+$ microglia was predicted to involve an alteration specific to PNI. CD11c$^+$ SDH microglia increased after the intraplantar injection of saporin conjugated with cholera toxin B (CTB) but not with isolectin B4 (IB4) (Fig. 3, A and B), which kill myelinated and unmyelinated primary afferents, respectively (*11*, *12*). PNI did not induce proliferation of existing CD11c$^+$ microglia (fig. S11C). These results, together with the fact that CD11c$^+$ SDH microglia expressed genes associated with phagocytosis (table S1), raised the possibility that the appearance of CD11c$^+$ SDH microglia after PNI could be related to the phagocytosis of myelin from damaged primary afferents. Primary cultured CD11c$^+$ spinal microglia from adult *Itgax-Venus* mice after PNI showed the increased uptake of myelin debris (Fig. 3C). In the SDH, CD11c$^+$ microglia had myelin-basic protein (MBP) in their lysosomes (CD68$^+$) after PNI increased (Fig. 3D). Electron microscopy revealed CD11c$^+$ SDH microglia containing phagocytosed myelin particles after PNI (Fig. 3E). The intraspinal injection of purified myelin debris into normal *Itgax-Venus* mice induced the engulfment of myelin debris and caused microglia to become CD11c$^+$ (Fig. 3F). Furthermore, after PNI, CD11c^{high} SDH microglia expressed *Axl* mRNA (Figs. 2, B and C, and 3G), a member of the TAM (TYRO3, AXL, and MERTK) family of tyrosine kinase receptors implicated in the engulfment of myelin debris (*13*–*15*). *Mertk* but not *Tyro3* mRNA was also expressed, but its level was comparable between CD11c^{neg} and CD11c^{high} microglia and was not changed after PNI (fig. S12A). AXL protein was expressed in CD11c$^+$ SDH microglia in PNI mice (Fig. 3, H and I). Moreover, *Axl*$^{-/-}$ mice showed a reduction of the PNI-induced increase in CD11c^{high} SDH microglia (Fig. 3J) and of the *Igf1* mRNA up-regulation in microglia (Fig. 3K), along with prolonged pain hypersensitivity (Fig. 3L). In sorted SDH microglia, the peak expression of *Axl* matched that of *Itgax* and preceded that of *Igf1* mRNA (fig. S12). Thus, AXL participated in the PNI-induced appearance of CD11c^{high} SDH microglia, presumably through the phagocytosis of myelin debris.

CD11c$^+$ microglia remained in the SDH for extended periods of time, even after the apparent behavioral signs of pain hypersensitivity disappeared. Depleting CD11c$^+$ microglia by intrathecal injection of DTX on day 35 (fig. S13A), when PWT had returned to the basal level, resulted in a relapse in pain hypersensitivity (Fig. 4A). The role of CD11c$^+$ SDH microglia in relapsing pain was supported by the data showing no effect of DTX on PWT in WT mice (fig. S13B) and no effect of deletion of other CD11c$^+$ cell types in the periphery (fig. S14). A similar recurrence of pain hypersensitivity was observed in *Cx3cr1*^{CreERT2};*Igf1*$^{flox/flox}$ mice when tamoxifen was administered at days 37 and 39 after PNI (Fig. 4B), a treatment that efficiently knocked down *Igf1* mRNA expression in SDH microglia (fig. S15). Moreover, intrathecally administered the neutralizing antibody anti-IGF1#1 (Fig. 4C) or the IGF1 receptor inhibitor JB-1 (fig. S16) also caused the reemergence of hypersensitivity. Thus, CD11c$^+$ microglia-derived IGF1 in the spinal cord repressed the manifestation of a neuropathic, hypersensitive state in an ongoing fashion. In addition, there were no sex differences in the appearance of CD11c$^+$ microglia in the SDH after PNI or in their role in neuropathic pain (fig. S17).

This study revealed that CD11c$^+$ microglia are essential for the remission and recurrence of neuropathic pain. This implies that, although mice appeared to exhibit normal behavioral responses several weeks after PNI, this was not due to the normalization of a pathologically altered hypersensitive state but rather to the dynamic equilibrium between the ongoing pain facilitation and repression, the latter of which primarily involves IGF1-expressing CD11c$^+$ microglia. Shifting microglia toward

this state partly involved AXL and myelin phagocytosis. Pain repression by spinal IGF1 might be associated with its reported pleiotropic effects, including the modulation of neuronal activity (*16*) and astrocytic and microglial inflammatory responses (*17*). CD11c$^+$ microglia have also been shown to increase in the brain in neurological disease models (*7*, *18*). Thus, our findings could yield an avenue for the development of therapeutic strategies to treat neuropathic pain and other pathological symptoms.

REFERENCES AND NOTES

1. K. Inoue, M. Tsuda, *Nat. Rev. Neurosci.* **19**, 138–152 (2018).
2. G. Chen, Y.-Q. Zhang, Y. J. Qadri, C. N. Serhan, R.-R. Ji, *Neuron* **100**, 1292–1311 (2018).
3. T. Masuda, R. Sankowski, O. Staszewski, M. Prinz, *Cell Rep.* **30**, 1271–1281 (2020).
4. M. Prinz, S. Jung, J. Priller, *Cell* **179**, 292–311 (2019).
5. V. Stratoulias, J. L. Venero, M.-È. Tremblay, B. Joseph, *EMBO J.* **38**, e101997 (2019).
6. K. Bulloch *et al*., *J. Comp. Neurol.* **508**, 687–710 (2008).
7. A. Benmamar-Badel, T. Owens, A. Wlodarczyk, *Front. Immunol.* **11**, 430 (2020).
8. T. Masuda *et al*., *Nat. Immunol.* **21**, 802–815 (2020).
9. C. N. Parkhurst *et al*., *Cell* **155**, 1596–1609 (2013).
10. T. Goldmann *et al*., *Nat. Immunol.* **17**, 797–805 (2016).
11. I. J. Llewellyn-Smith, C. L. Martin, L. F. Arnolda, J. B. Minson, *J. Neurosci. Methods* **103**, 83–90 (2000).
12. J. W. Tarpley, M. G. Kohler, W. J. Martin, *Brain Res.* **1029**, 65–76 (2004).
13. J. G. Weinger *et al*., *J. Neuroinflammation* **8**, 49 (2011).
14. H. J. Hoehn *et al*., *Brain Res.* **1240**, 1–11 (2008).
15. C. V. Rothlin, E. A. Carrera-Silva, L. Bosurgi, S. Ghosh, *Annu. Rev. Immunol.* **33**, 355–391 (2015).
16. A. Pristerà *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **116**, 3817–3826 (2019).
17. J. L. Labandeira-Garcia, M. A. Costa-Besada, C. M. Labandeira, B. Villar-Cheda, A. I. Rodríguez-Perez, *Front. Aging Neurosci.* **9**, 365 (2017).
18. A. Wlodarczyk *et al*., *Front. Immunol.* **6**, 463 (2015).

ACKNOWLEDGMENTS

We thank D. Tomiyama for assisting behavioral experiments; L. Giles from Edanz Group (https://en-author-services.edanzgroup.com/ac) for editing a draft of this manuscript; S. Yoneda and M. Fujita (Shionogi & Co., Ltd.) for helpful comments and advice during this study; the University of Pennsylvania vector core for providing pZac2.1, pAAV2/9, and pAd DeltaF6 plasmid; and the Research Support Center, Research Center for Human Disease Modeling, Kyushu University Graduate School of Medical Sciences, for technical assistance. **Funding:** This work was supported by the Japanese Society for the Promotion of Science (JSPS KAKENHI grants JP19H05658 and JP20H05900 to M.T.; JP20K22687, JP21H02752, and JP21H00204 to T.M.; and JP16H06280 to H.S.); the Japan Science and Technology Agency [JST Moonshot R&D grant JPMJMS2024 to M.T. and a Grant-in-Aid for Scientific Research on Innovative Areas – Platforms for Advanced Technologies and Research Resources "Advanced Bioimaging Support (ABiS)" to H.S.]; CREST (AMED grant JP21gm0910006 to M.T.; AMED grant JP20gm6310016 AMED-PRIME JP21wm0425001 to T.M.); the Platform Project for Supporting Drug Discovery and Life Science Research (BINDS AMED grant JP21am0101091 to M.T.); and by an academic research grant from Shionogi & Co., Ltd. **Author contributions:** K.K. designed the experiments, performed almost all of the experiments, analyzed the data, and wrote the manuscript. R.S. performed experiments and analyzed the data. K.Y. and S.M. assisted with some experiments. K.Tan. and Y.O. performed RNA-seq experiments and analyzed the data. K.Tak. and H.S. performed electron microscopy experiments. K.I. analyzed the data. T.M. analyzed the RNA-seq data, provided comments on experiments, and wrote the manuscript. M.T. conceived and initiated this project, designed and performed experiments, supervised the overall project, and wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data generated supporting the findings of this study are available in the manuscript or the supplementary materials. Additional information and materials are available from the corresponding author upon reasonable request.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abf6805
Materials and Methods
Extended Captions for Figs. 1 to 4
Figs. S1 to S17
Tables S1 and S2
References (*19*–*43*)
MDAR Reproducibility Checklist

11 November 2020; resubmitted 20 January 2022
Accepted 25 February 2022
10.1126/science.abf6805

MEMBRANES

An integrated materials approach to ultrapermeable and ultraselective CO_2 polymer membranes

Marius Sandru[1]*, Eugenia M. Sandru[1], Wade F. Ingram[2], Jing Deng[3], Per M. Stenstad[1], Liyuan Deng[3], Richard J. Spontak[2,4]*

Advances in membrane technologies that combine greatly improved carbon dioxide (CO_2) separation efficacy with low costs, facile fabrication, feasible upscaling, and mechanical robustness are needed to help mitigate global climate change. We introduce a hybrid-integrated membrane strategy wherein a high-permeability thin film is chemically functionalized with a patchy CO_2-philic grafted chain surface layer. A high-solubility mechanism enriches the concentration of CO_2 in the surface layer hydrated by water vapor naturally present in target gas streams, followed by fast CO_2 transport through a highly permeable (but low-selectivity) polymer substrate. Analytical methods confirm the existence of an amine surface layer. Integrated multilayer membranes prepared in this way are not diffusion limited and retain much of their high CO_2 permeability, and their CO_2 selectivity is concurrently increased in some cases by more than ~150-fold.

An alternative to chemical absorption for the capture of CO_2 derives from polymer gas-separation membranes (*1*–*3*) that can be chemically tailored to achieve different levels of CO_2 selectivity and permeability. Cross-linked elastomeric membranes produced from polyethers are reverse selective—which means that their selectivity is based on solubility rather than size (i.e., diffusion) considerations—and display a strong chemical affinity for CO_2 (*4*–*7*). Alternatively, glassy polymers, such as polyimides and polysulfones, operate on molecular-size sieving and can be modified in various ways to alter the free-volume pathways through which penetrant molecules migrate (*8*–*10*). In addition to these mature membrane technologies, recent studies have demonstrated that humidified block ionomers promote high CO_2 permeabilities at moderate selectivities (*11*, *12*), whereas membranes containing mobile CO_2-philic carriers can yield high CO_2 selectivities at relatively low permeabilities (*13*). Although these advances, as well as those regarding other classes of polymer membranes, are encouraging, they confirm the trade-off between permeability and selectivity, which is manifested as the empirical Robeson upper bound (*14*).

In efforts to improve the gas flux through a membrane, polymer membranes have evolved from self-standing films measuring ~50 to 100 μm thick to supported thin films typically one to two orders of magnitude thinner (*15*, *16*), as illustrated in Fig. 1A. The motivation for doing so is that thin films retain the polymer characteristics responsible for selectivity and reduce the diffusive path length (because macroporous supports afford a negligible barrier to diffusion). A special case of CO_2 permeation is based on chemically augmented facilitated transport, wherein CO_2 molecules react with specific chemical groups in a hydrated membrane and consequently permeate more quickly (*17*–*21*). Details of these transport mechanisms are provided in the supplementary materials (section 1). A polymer that is technologically promising is hydrated polyvinylamine (PVAm) (*19*, *22*), which has been reported (*20*) to exhibit CO_2/N_2 selectivities up to 500, depending on test conditions and specimen preparation. Unfortunately, CO_2 permeability is diffusion hindered, even through hydrated PVAm membranes

[1]Department of Biotechnology and Nanomedicine, SINTEF Industry, 7034 Trondheim, Norway. [2]Department of Materials Science and Engineering, North Carolina State University, Raleigh, NC 27695, USA. [3]Department of Chemical Engineering, Norwegian University of Science and Technology, 7491 Trondheim, Norway. [4]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27695, USA.
***Corresponding author. Email: marius.sandru@sintef.no (M.S.); rich_spontak@ncsu.edu (R.J.S.)**

that are less than a micrometer thick. An important design paradigm drawn from this observation is that the ultrahigh CO_2 selectivity afforded by facilitated transport should be preserved, whereas the extent to which diffusive limitations negatively affect CO_2 permeability should be lessened.

One route to fabricate a polymer membrane having both characteristics is included in Fig. 1A. Here, a high-solubility and fast-diffusion membrane is generated by growing an ultrathin amine-containing surface layer on a supported high-permeability polymer thin film. These hybrid-integrated (HI) membranes exploit the tunable surface functionalization introduced through surface polymerization (*23*, *24*) so that their surface is covered by a highly CO_2-philic and hydrophilic nanoscale polymer layer, which is conceptually similar to surface-modified polymer membranes produced for liquid separation (*25*). Within this nanofabricated design (achieved using the setup pictured in fig. S1; supplementary materials, section 2), a mixture of CO_2, N_2, and H_2O gases initially encounters a molecularly thin amine-rich layer (region I in Fig. 1B), and the CO_2 permeates through this hydrated layer as a result of facilitated transport without appreciable diffusive resistance. In essence, this layer serves to concentrate CO_2 before it enters the underlying substrate. An elevated population of CO_2 molecules then enters and quickly permeates through a dense polymer support (region II) and then a macroporous support (region III), yielding both high CO_2/N_2 selectivity and high CO_2 permeability. This scenario, depicted in Fig. 1B, represents a combinatorial separation mechanism that marries fast facilitated transport (high CO_2 selectivity) at the membrane surface in region I with solution-diffusion transport (high CO_2 permeability) through region II.

Two amorphous commercialized gas-separation membranes having relatively high CO_2 permeability—elastomeric polydimethylsiloxane (PDMS) and glassy polytetrafluoroethylene (PTFE AF)—were selected for surface modification. The PDMS membranes were procured as commercial membrane systems (DeltaMem AG, Allschwil, Switzerland) composed of PDMS thin films measuring 6 μm thick and supported on porous polyacrylonitrile mounts. The PTFE AF membranes were produced in house by casting Teflon AF (amorphous fluoropolymer) 2400 (Chemours, Geneva, Switzerland) from 3M Fluorinert Electronic Liquid FC-72 (Nordmann Nordic AB, Stockholm, Sweden) into Teflon petri dishes measuring 10 cm in diameter. The resultant films were first dried at 25°C for 24 hours and then vacuum-dried at 200°C for at least 24 hours to yield films measuring 25 to 50 μm thick. To avoid edge effects, they were cut into circles (8 cm diameter) and mechanically supported on porous poly(vinylidene fluoride) mounts (MFP2, with a pore size of 0.2 μm) manufactured by Alfa Laval (Soborg, Denmark).

The series of reactions used to chemically alter the polymer surfaces is portrayed in Fig. 1C and begins with the introduction of initiator sites by proton extraction under ultraviolet (UV) conditions. For this purpose, benzophenone is chemically attached to each membrane surface to introduce reactive hydroxyl moieties (*26*). The surface density of initiator sites was controlled by removing unbound initiator with methanol or acetonitrile and varying the initiator concentration (from 1 to 15% in acetonitrile), with the best CO_2 permeation results achieved at 10% benzophenone. These sites are then reacted with glycidyl methacrylate (GMA) monomer (3 to 15% in methanol or acetonitrile, with 10% yielding the most promising results) and polymerized under UV radiation to yield poly (glycidyl methacrylate) (PGMA) chains on the surface of each membrane. Epoxy ring–opening the PGMA chains with 20 to 50% ethylene diamine (EDA) in aqueous solution yielded the desired amine functionality. Additional details regarding the surface modification are provided in the supplementary materials (section 3). According to x-ray photoelectron spectroscopy (XPS) spectra collected on a Kratos Ultra spectrometer (Fig. 1D), both surface-modified polymers exhibit evidence of surface nitrogen, confirming the presence of surface amine groups after the final reaction step. Complementary spectroscopic results are included in figs. S2 to S4.

Incorporation of a patchy ultrathin amine layer on the surface of PDMS, an elastomer, and PTFE AF, a glassy polymer, results in profound changes in CO_2 permeation and selectivity, as evidenced by the pressure- and temperature-dependent measurements compiled in Fig. 2. Gas permeability tests were performed with a real-time automated data collection setup (fig. S5) wherein feed pressure, humidity, temperature, and He sweep are all precisely regulated. All the results reported here—measured from a mixed 10/90 CO_2/N_2 gas feed (at 100% relative humidity) that is

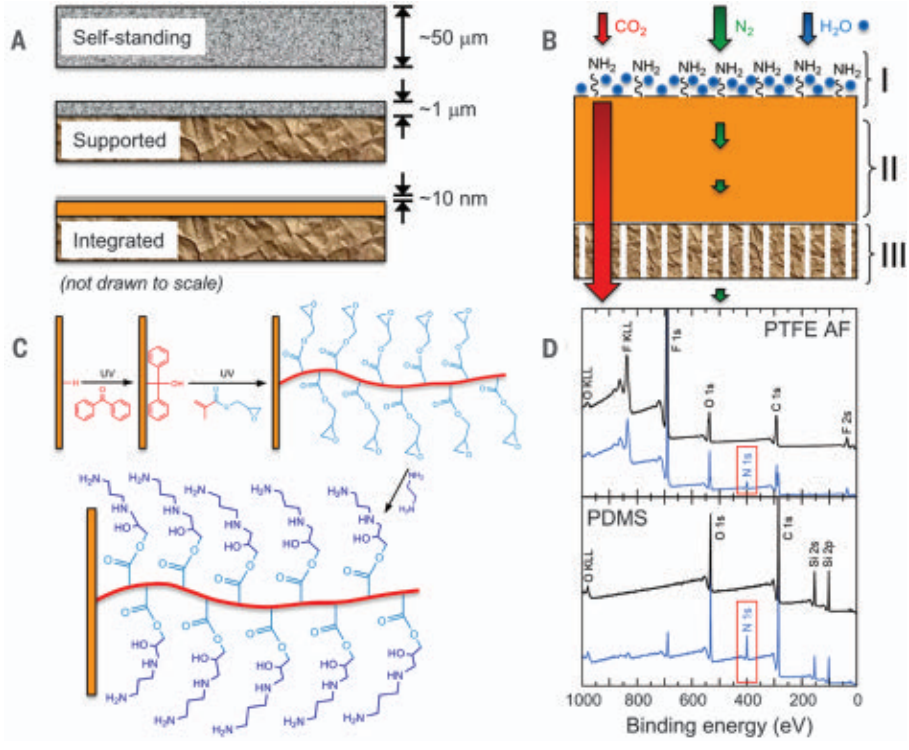


Fig. 1. Conceptual and experimental details of the studied HI membranes. (**A**) Schematic diagram displaying the evolution of gas-separation polymer membranes from self-standing to supported to integrated (labeled). (**B**) Illustration of the underlying mechanism behind HI membranes. Here, a surface-grown amine-rich layer (region I) is grown on the surface of a highly permeable polymer thin film (region II) supported by conventional macroporous polymers (region III). Selectivity is exclusively achieved in region I, which requires hydration. (**C**) The chemical route used to grow amine-modified chains on polymer substrates. See the text and the supplementary materials (section 2) for details. (**D**) Representative XPS spectra acquired from PTFE AF (top) and PDMS (bottom) membranes before (black lines) and after (blue lines) surface modification. The presence of the nitrogen (N 1s) peak associated with the amine-rich surface layer is highlighted in each surface-modified spectrum.

representative of a broad spectrum of CO_2-related separations and a standard feed pressure (1.2 bar) and temperature (25° ± 1°C), unless otherwise indicated—reflect the most common conditions encountered for gases containing CO_2. The dependence of CO_2 permeability on pressure displayed in Fig. 2A confirms that the parent PDMS and PTFE AF polymer thin films have comparable permeabilities and remain largely unaffected by increasing pressure over the range examined, as expected for solution-diffusion permeation. By contrast, the CO_2 permeabilities of the amine-functionalized membranes (am-PDMS and am–PTFE AF) are lower than those of their parent membranes, assuming negligible resistance from the membrane supports, and they decrease with increasing pressure, which is indicative of facilitated transport (*17*, *19*, *20*).

Similarly, the mixed-gas CO_2/N_2 selectivities of the unmodified and modified polymers decrease slightly with increasing pressure in Fig. 2C, whereas the am-PDMS and am–PTFE AF polymers exhibit highly elevated selectivity levels. The temperature-dependent permeabilities provided in Fig. 2B reveal that the parent polymers are either insensitive to or increase slightly with increasing temperature. Although analogous behavior is observed for am–PTFE AF, the am-PDMS membrane appears to be more highly temperature dependent. This response difference is attributed to the elastomeric (flexible) nature of PDMS versus the more rigid nature of glassy PTFE AF, because the materials are expected to expand differently upon heating, in accord with their coefficients of thermal expansion: 9.6×10^{-4} °C^{-1} (PDMS) and 3.0×10^{-4} °C^{-1} (PTFE AF). The corresponding CO_2/N_2 selectivities included in Fig. 2D indicate that those of the parent polymers are comparably low and decrease slightly further with increasing temperature. By contrast, the selectivities of the surface-functionalized polymers are much higher but, in similar fashion to their CO_2 permeabilities, exhibit different thermal behavior. Additional measurement conditions are considered in figs. S6 to S8 in section 4.

The improvements in CO_2/N_2 selectivity apparent in Fig. 2 in the presence of moisture are a consequence of the hydrated, short-chain patchy amine layers on the surfaces of PDMS and PTFE AF. Low-voltage scanning electron microscopy (SEM) images of membranes acquired on an ultrahigh-resolution FEI Verios 460L Schottkey emitter electron microscope at an accelerating voltage of 1.0 kV before and after surface functionalization are displayed in Fig. 3. In Fig. 3A, the surface of the unmodified PTFE AF thin film reveals the presence of surface pits (~80 to 450 nm in diameter), which, upon functionalization in Fig. 3B, become completely obscured. Because the surface of neat PDMS is featureless, Fig. 3, C and D, focuses on surface characteristics that develop during surface polymerization. These features resemble dewetting patterns, which implies that one of the steps in the reaction sequence did not occur uniformly on the PDMS surface. However, closer examination of the membranes before and after surface functionalization reveals a more complex picture. Height and amplitude atomic force microscopy (AFM) images collected on an Asylum MFP-3D probe microscope are displayed for membranes based on am–PTFE AF in Fig. 3, E and F, and for membranes based on am-PDMS in Fig. 3, G and H. The insets correspond to the unmodified polymer films, and the height images yield root mean square (RMS) roughness values of 6.85 and 0.99 nm for virgin PTFE AF and PDMS, respectively.

The modified surfaces consist of discrete nanodroplets, confirming the likelihood of surface dewetting during functionalization. The diameters of these features measure 42 to 264 nm (127 nm average) on PTFE AF and 41 to 179 nm (95.9 nm average) on PDMS, and their mean heights vary from 97.9 nm on PTFE AF to 12.3 nm on PDMS. These nanodroplets amplify the RMS surface roughness by ~10× (68.6 nm on am–PTFE AF and 8.93 nm on am-PDMS). Increases in the membrane surface area owing to the presence of topological features (relative to a flat surface) are 13.2 and 3.6% for am–PTFE AF and am-PDMS, respectively, thereby providing greater opportunity for CO_2 molecules to interact with and transport through the amine-rich layer. A geometric interpretation of this increase in surface area is provided in figs. S9 and S10. By changing the intensity and exposure area of the surface polymerization (supplementary materials, section 5), macroscopic dewetting can be eliminated on PDMS, yielding a second-generation HI membrane designated as am-PDMS-2. According to SEM images of am-PDMS-2 membranes, no large features are evident (fig. S11A), whereas a nanoscale topology is discernible (fig. S11B). Corresponding height and phase AFM images are presented in Fig. 3, I and J [as well as three-dimensional (3D) topographical images in fig. S12], that confirm the existence of a patchy nanoscale topology with a substantially reduced RMS surface roughness (1.68 nm) on the am-PDMS-2 membranes. The circled features in Fig. 3, I and J, identify protrusions measuring up to ~12 nm in height that are soft relative to the surrounding matrix, indicating that they represent low-density, surface-grafted chain patches. Avoiding surface heterogeneities on the functionalized PDMS membranes not only increases uniformity but also improves both CO_2 permeability (Fig. 2B) and CO_2/N_2 selectivity (Fig. 2D).

The performance of these HI membranes relative to other polymer membranes intended for CO_2 separation is routinely assessed on a Robeson plot, which evaluates membrane selectivity as a function of permeability. This representation identifies the empirical upper bound that reflects the trade-off between selectivity and permeability. As membranes have continued to improve since the original upper bound was identified (*27*), the position of the upper bound has progressively shifted upward. The upper bound identified by Robeson (*14*)

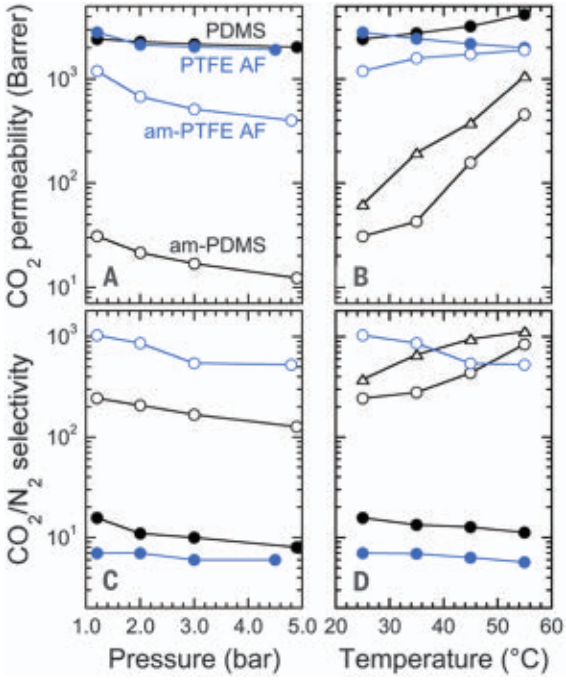


Fig. 2. Molecular transport through parent and amine-modified PDMS and PTFE AF membranes. (**A** to **D**) CO_2 permeability [(A) and (B)] and CO_2/N_2 selectivity values [(C) and (D)] measured as functions of feed pressure [(A) and (C)] and temperature [(B) and (D)] from mixed gases for membranes derived from PDMS (black symbols and lines) and PTFE AF (blue symbols and lines) before and after surface modification [labeled in (A)]. Included in (B) and (D) are measurements from am-PDMS-2 membranes (indicated by triangles). The solid lines serve to connect the data. In variable-pressure tests, the temperature is maintained at 25° ± 1°C, whereas the pressure in variable-temperature tests is kept at 1.2 bar, and the feed stream is fully humidified (100% relative humidity).

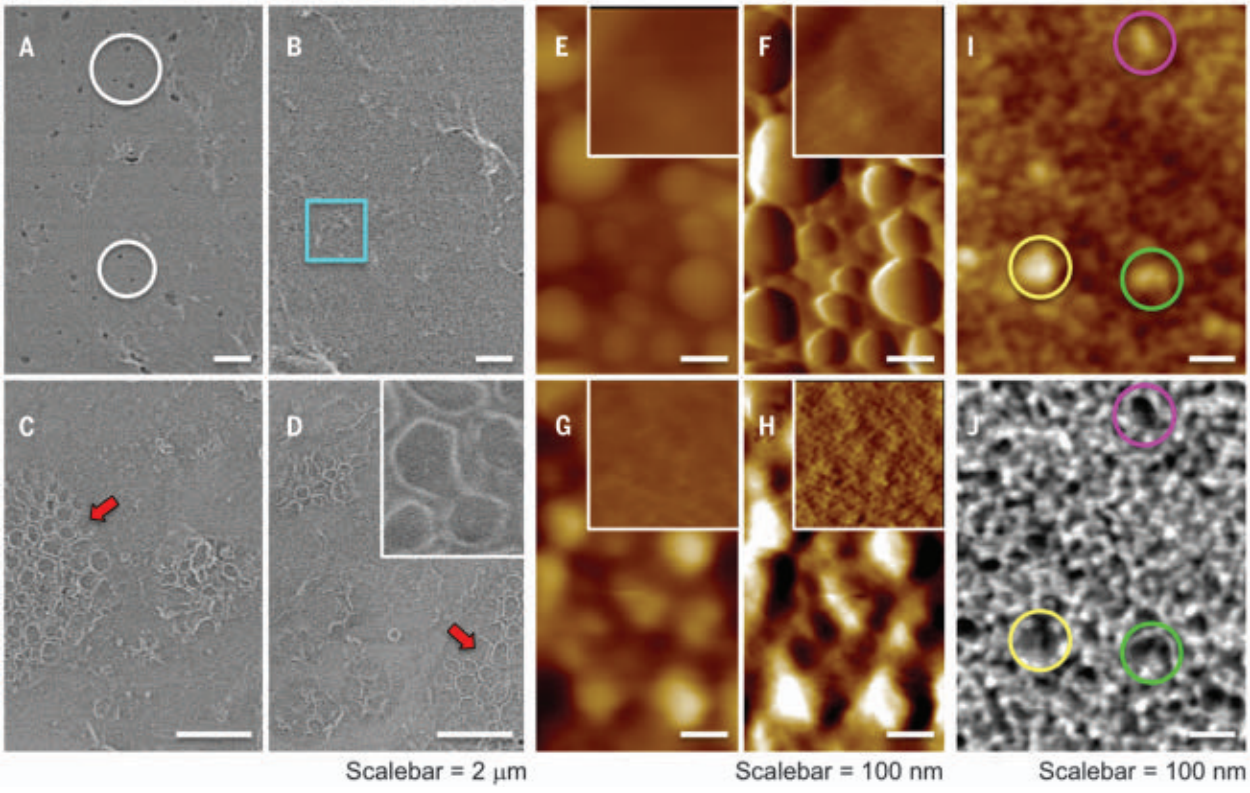


Fig. 3. Topological characteristics of parent and amine-modified PDMS and PTFE AF membranes. (**A** and **B**) Low-voltage planar SEM images of solvent-cast PTFE AF membranes before (A) and after (B) surface modification. Examples of pits measuring between ~80 and 450 nm in diameter are highlighted by circles in (A), whereas a region that is reminiscent of ridges such as those evident in (A) is identified by the square in (B). (**C** and **D**) SEM images reveal the surface features that develop on first-generation PDMS membranes upon surface modification. The inset in (D) is a 3× enlargement of one of the regions identified by arrows. (**E** to **H**) Complementary AFM height [(E) and (G)] and amplitude [(F) and (H)] images corresponding to PTFE AF [(E) and (F)] and PMDS [(G) and (H)] membranes after surface modification (the inset displayed at the same magnification in each frame is acquired from the parent, unmodified surface). (**I** and **J**) The paired height and phase AFM images of the surface of an am-PDMS-2 membrane in (I) and (J), respectively, confirm the existence of patchy amine-rich regions. The color-matched circles identify protrusions (light) in (I) that are mechanically soft (dark) in (J). Values of the surface RMS roughness measured from AFM images, such as these are provided in the text, and 3D topographical images of the am-PDMS-2 membrane are included in fig. S12.

is identified in Fig. 4A, which compares membranes that function according to the solution-diffusion mechanism. Included in this figure is a more recent (and shifted) upper bound proposed by Comesaña-Gándara *et al.* (*28*) in 2019 on the basis of polymers of intrinsic microporosity (PIMs). The PIM membranes afford very high CO_2 permeabilities at modest CO_2/N_2 selectivity levels (*29*, *30*), whereas a variety of other membranes also displayed in Fig. 4A range in permeability and selectivity from relatively low to relatively high at ~25° to 35°C. To provide an additional classification description of these performance metrics, we divide the Robeson plot into quadrants and refer to membranes exhibiting CO_2 permeabilities in excess of 10^3 Barrer [where 1 Barrer $= 10^{-10}$ $\frac{cm^3_{STP} \cdot cm}{cm^2 \cdot s \cdot cmHg}$ (cm^3_{STP}, standard cubic centimeter; cmHg, centimeter of mercury)] as ultrapermeable (uP) and membranes achieving CO_2/N_2 selectivities greater than 10^2 as ultraselective (uS). Although many of the membranes presented in Fig. 4A (including those prepared from unmodified PTFE AF and PDMS) are uP in the fourth quadrant, notably few are uS. In fact, two types of membranes evaluated for pilot-scale separation of CO_2 from flue gas are neither uP nor uS (in the third quadrant). This disparity suggests that membranes fabricated for CO_2 capture focus primarily on improved permeability (*6*, *28*, *31*).

Examples satisfying the criterion set forth above for uS membranes in the second quadrant of Fig. 4A are cellulose nanofiber/ionic liquid (CNF/IL) membranes (which rely on hydrated IL nanochannels) (*13*) as well as most of the am-PDMS and am-PDMS-2 HI membranes investigated in this work and several materials reported in the literature. However, only a few membranes including a bioinspired thin film (*32*) and several membranes incorporating graphene oxide nanosheets (*33*–*35*) can be classified as both uP and uS in the first quadrant of Fig. 4A. In marked contrast, all the am–PTFE AF membranes investigated here at temperatures from 25° to 55°C, as well as the am-PDMS-2 membrane evaluated at 55°C, are both uP and uS. Although surface amination of the PTFE AF membranes reduces their CO_2 permeability by 50 to 60%, the corresponding increase in CO_2/N_2 selectivity is ~150× greater. These results confirm that the patchy, low-density surface amine layer provides a highly hydrophilic and CO_2-philic environment that absorbs and transports CO_2 molecules by means of facilitated transport, whereas the highly permeable substrate provides contiguous pathways with little barrier to molecular diffusion. For added comparison, the results displayed in Fig. 4B confirm that the transport metrics reported here for the HI membranes have higher CO_2 selectivities than most purely facilitated-transport membranes, which primarily lie in the second quadrant (with two selected for pilot-scale evaluation).

In Fig. 4A, the CO_2/N_2 selectivity achieved at 1200 Barrer for the am–PTFE AF membrane is substantially higher than the Robeson (*14*)

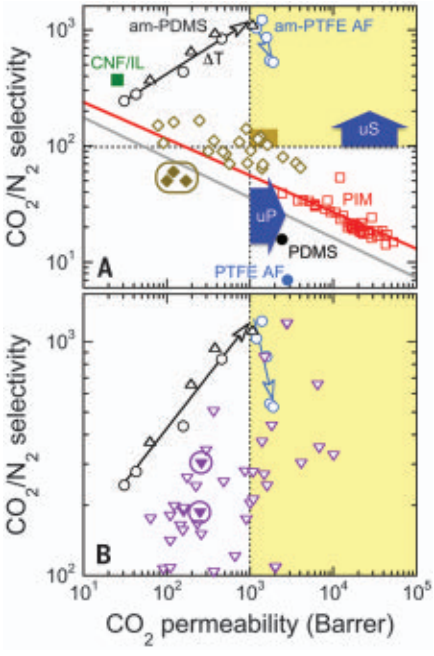


Fig. 4. Performance comparison of HI membranes relative to other polymer membranes developed for CO_2 capture. (**A**) Performance metrics of the HI membranes studied in this work (am-PDMS, open black circles; am-PDMS-2, open black triangles; am–PTFE AF, open blue circles) relative to their parent materials (PDMS, filled black circle; PTFE AF, filled blue circle) as well as other polymer membranes proposed for CO_2 capture on the basis of the solution-diffusion mechanism. According to the requirements proposed here, uP membranes exhibiting CO_2 permeabilities beyond 10^3 Barrer (in the fourth quadrant) include both PDMS and PTFE AF precursors in addition to several other promising polymers (open brown diamonds) and those in the PIM (open red squares) category. Membranes subjected to pilot-scale evaluation (filled brown diamonds) are identified in the third quadrant and listed in table S1. In the second quadrant, uS membranes, such as those based on CNF/IL membranes (*13*) (filled green square), reach CO_2/N_2 selectivities of more than 10^2 at lower CO_2 permeability levels. The Goldilocks (first) quadrant (shaded yellow) combines both uP and uS performance. The solid lines identify the permeability-selectivity trade-off proposed by Robeson (*14*) for various polymer membranes (gray) and more recently by Comesaña-Gándara *et al*. (*28*) for PIM membranes (red). (**B**) Performance metrics of the HI membranes relative to hydrated facilitated-transport membranes (open purple inverted triangles) in the first and second quadrants only. Promising membranes assessed in pilot-scale testing (closed purple inverted triangles) are circled and compiled in table S1. The color-coded arrows in (A) and (B) indicate the effect of increasing temperature (ΔT) on the HI membranes derived from PDMS (black) and PTFE AF (blue).

or PIM-based (*28*) upper-bound thresholds (by ~30× and ~20×, respectively). Equally important is our observation that the enhanced separation performance of am–PTFE AF at different temperatures in Fig. 4A varies less compared with that of the am-PDMS and am-PDMS-2 membranes. The CO_2 transport characteristics of the am–PTFE AF and am-PDMS-2 membranes converge as the amine surface layers exhibit less heterogeneities during polymerization (supplementary materials, section 5). Differences in separation performance accompanying variations in surface modification are considered in figs. S13 to S15, and the results reported here for CO_2/N_2 gas mixtures are extended to CO_2/CH_4 mixtures in fig. S16, which reveals that the am–PTFE AF membranes retain their uP and uS performance with regard to CO_2 in the case of two dissimilar gas mixtures. Additionally, aspects of durability and adaptability of the am-PDMS-2 and am–PTFE AF membranes are examined in figs. S17 and S18, respectively, and results collected from single- and mixed-gas tests are quantitatively compared in fig. S19 to further elucidate the molecular-level mechanism by which CO_2 permeates through the HI membranes studied in this work.

Although the results reported herein are highly encouraging, three other relevant issues should be recognized. First, the membranes developed in this work derive from polymers that are already used for commercial gas separations, in which case they are widely available and the processes required to manufacture and upscale gas-separation modules for industrial operations would accommodate the surface-functionalized analogs. Second, because the PTFE AF and PDMS membranes are already in commercial use, the added expense of surface functionalization is most likely much less than introducing a new designer membrane. Third, the methodology described in this study is not restricted to PTFE AF and PDMS with a patchy, amine-rich surface layer. We have selected the systems discussed here as exemplars to provide proof of concept. By marrying the facilitated transport of a molecularly thin amine surface layer with a high-permeability polymer film, we have successfully demonstrated that HI membranes can be nanofabricated to take advantage of this high-solubility, quick-diffusion combination and produce next-generation membranes that, instead of being limited by a permeability-selectivity trade-off, are both uP and uS.

REFERENCES AND NOTES

1. M. E. Boot-Handford *et al*., *Energy Environ. Sci.* **7**, 130–189 (2014).
2. D. S. Sholl, R. P. Lively, *Nature* **532**, 435–437 (2016).
3. J. M. S. Henis, M. K. Tripodi, *Science* **220**, 11–17 (1983).
4. N. P. Patel, A. C. Miller, R. J. Spontak, *Adv. Mater.* **15**, 729–733 (2003).
5. H. Lin, E. Van Wagner, B. D. Freeman, L. G. Toy, R. P. Gupta, *Science* **311**, 639–642 (2006).
6. S. Wang *et al*., *Energy Environ. Sci.* **9**, 1863–1890 (2016).
7. J. Liu *et al*., *Joule* **3**, 1881–1894 (2019).
8. M. D. Guiver, Y. M. Lee, *Science* **339**, 284–285 (2013).
9. L. M. Robeson, Z. P. Smith, B. D. Freeman, D. R. Paul, *J. Membr. Sci.* **453**, 71–83 (2014).
10. M. Galizia *et al*., *Macromolecules* **50**, 7809–7843 (2017).
11. Z. Dai, L. Ansaloni, J. J. Ryan, R. J. Spontak, L. Deng, *Green Chem.* **20**, 1391–1404 (2018).
12. Z. Dai *et al*., *NPG Asia Mater.* **11**, 53 (2019).
13. S. Janakiram *et al*., *Green Chem.* **22**, 3546–3557 (2020).
14. L. M. Robeson, *J. Membr. Sci.* **320**, 390–400 (2008).
15. R. W. Baker, B. T. Low, *Macromolecules* **47**, 6999–7013 (2014).
16. H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech, B. D. Freeman, *Science* **356**, eaab0530 (2017).
17. R. D. Noble, *J. Membr. Sci.* **50**, 207–214 (1990).
18. W. J. Koros, R. Mahajan, *J. Membr. Sci.* **175**, 181–196 (2000).
19. L. Deng, T.-J. Kim, M.-B. Hägg, *J. Membr. Sci.* **340**, 154–163 (2009).
20. M. Sandru, S. H. Haukebø, M.-B. Hägg, *J. Membr. Sci.* **346**, 172–186 (2010).
21. J. Liao *et al*., *Chem. Sci.* **5**, 2843–2849 (2014).
22. B. Wang *et al*., *Adv. Mater.* **32**, e1907701 (2020).
23. S. Edmondson, V. L. Osborne, W. T. S. Huck, *Chem. Soc. Rev.* **33**, 14–22 (2004).
24. M. A. C. Stuart *et al*., *Nat. Mater.* **9**, 101–113 (2010).
25. J. R. Werber, C. O. Osuji, M. Elimelech, *Nat. Rev. Mater.* **1**, 16018 (2016).
26. J. Deng, L. Wang, L. Liu, W. Yang, *Prog. Polym. Sci.* **34**, 156–193 (2009).
27. L. M. Robeson, *J. Membr. Sci.* **62**, 165–185 (1991).
28. B. Comesaña-Gándara *et al*., *Energy Environ. Sci.* **12**, 2733–2740 (2019).
29. Z.-X. Low, P. M. Budd, N. B. McKeown, D. A. Patterson, *Chem. Rev.* **118**, 5871–5911 (2018).
30. A. A. Shamsabadi *et al*., *Prog. Energy Combust. Sci.* **84**, 100903 (2021).
31. W. J. Koros, C. Zhang, *Nat. Mater.* **16**, 289–297 (2017).
32. Z. Wang, Y. Zhang, J. Wang, Y. Zhang, *Sustain. Energy Fuels* **4**, 1191–1198 (2020).
33. Q. Xin *et al*., *J. Mater. Chem. A* **3**, 6629–6641 (2015).
34. X. Li *et al*., *ACS Appl. Mater. Interfaces* **7**, 5528–5537 (2015).
35. M. Chen *et al*., *J. Ind. Eng. Chem.* **63**, 296–302 (2018).

ACKNOWLEDGMENTS

We thank L. E. Parnas, Ø. Dahl, M. T. Guzmán-Gutiérrez, and C. B. Mooney for technical assistance. **Funding:** This study was supported by the Research Council of Norway through the POLYMEM project (grant no. 254791) in the CLIMIT program; UEFSCDI Romania through the CO2Hybrid project (grant no. 13/2020); the National Science Foundation through the North Carolina Research Triangle Nanotechnology Network (grant no. ECCS-2025064), which is part of the National Nanotechnology Coordinated Infrastructure; and the Kraton Corporation. **Author contributions:** M.S. and P.M.S. conceived the idea; E.M.S. conducted the surface polymerizations; J.D. and L.D. conducted the initial surface characterization and data interpretation; W.F.I. and R.J.S. conducted the final surface characterization and correlated surface topology with membrane performance; M.S. and R.J.S. wrote the manuscript and prepared the manuscript figures; and all authors contributed to preparing manuscript revisions, ensuring accurate interpretation, and incorporating reference data. **Competing interests:** E.M.S., M.S., and P.M.S. have filed a patent on this work entitled "Surface modified membranes for gas separation, and a method for preparing thereof" (US patent application #17/051,706; filed 5 August 2021). The remaining authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj9351
Supplementary Text
Figs. S1 to S19
Tables S1 and S2
References (*36*–*64*)

10 June 2021; resubmitted 28 November 2021
Accepted 7 March 2022
10.1126/science.abj9351

COGNITIVE SCIENCE

Complex cognitive algorithms preserved by selective social learning in experimental populations

B. Thompson[1]*†, B. van Opheusden[2]†, T. Sumers[2], T. L. Griffiths[2,3]

Many human abilities rely on cognitive algorithms discovered by previous generations. Cultural accumulation of innovative algorithms is hard to explain because complex concepts are difficult to pass on. We found that selective social learning preserved rare discoveries of exceptional algorithms in a large experimental simulation of cultural evolution. Participants (N = 3450) faced a difficult sequential decision problem (sorting an unknown sequence of numbers) and transmitted solutions across 12 generations in 20 populations. Several known sorting algorithms were discovered. Complex algorithms persisted when participants could choose who to learn from but frequently became extinct in populations lacking this selection process, converging on highly transmissible lower-performance algorithms. These results provide experimental evidence for hypothesized links between sociality and cognitive function in humans.

Reading, counting, cooking, and sailing are just some of the human abilities passed from generation to generation through social learning (*1*, *2*). Complex abilities like these often depend on learned cognitive algorithms: procedural representations of a problem that coordinate memory, attention, and perception into sequences of useful computations and actions. Accumulation of complex algorithms—from ancient tool-making techniques to bread making, boat building, or horticulture—is central to human adaptation (*3*–*5*) yet challenging to explain because algorithmic concepts can be difficult to discover, communicate, and learn from observation (*6*), making them vulnerable to loss (*7*–*10*). Theories of cultural evolution suggest that human social learning may help overcome this fragility (*11*, *12*). For example, mathematical models (*13*) predict that choosing to learn from successful or prestigious individuals can prevent the loss of rare innovations. However, this potential link between sociality and complex abilities (*14*) is challenging to establish.

We conducted large-scale simulations of cultural evolution with human participants to assess how selective social learning influenced the evolution of cognitive algorithms. Prior research (*15*–*22*) shows that social learning can improve decisions in multiple-choice tasks (*23*), perceptual judgments (*24*), and search problems (*25*) and can improve artifacts such as physical structures (*26*) or computer programs (*27*). However, the evolution of cognitive algorithms at the population level has been difficult to study (*28*). We developed custom software to recruit large numbers of participants online and organize them into evolving societies facing a common problem. Twenty populations tackled a sequential decision problem (Fig. 1A and materials and methods 1.1). Presented with six images, participants attempted to establish hidden arbitrary orderings using pairwise comparisons. Out-of-order pairs swapped positions when compared. Participants were rewarded for establishing the ordering using fewer comparisons. This task poses a sorting problem, requiring a strategy for executing appropriate sequences of actions, analogous to culturally evolved strategies for making tools or food.

Participants transmitted their strategies across 12 generations (Fig. 1B). Initial asocial generations completed the task individually. Participants in generations 2 to 12 were randomly assigned to either the experimental selective social learning (SSL) treatment (participants could choose demonstrators from the previous generation based on demonstrator performance) or to a control random mixing (RM) treatment (demonstrator performance was not visible). Participants were incentivized to improve on strategies that they observed and provide a helpful description and demonstration of their own strategy (materials and methods 1.1.13).

Figure 2 shows task performance. Asocial participants achieved the lowest scores [mean over scored trials = 0.33, SD = 0.19, relative to a theoretical maximum of 1; supplementary materials (SM) 3; mean number of successful trials = 6.45 of 10 scored trials, SD = 2.89; SM 2.1.1]. Asocial participants improved over trials, confirming within-task learning {regression coefficient posterior mean (β) = 0.14, 97% highest density interval (HDI) = [0.11, 0.16]; SM 2.1.1}. Participants who learned socially (generations 2 to 12) achieved higher performance [mean (M) = 0.446, SD = 0.199] than asocial participants (β = 0.114, 97% HDI = [0.083, 0.145]; SM 2.1.2). RM participants (M = 0.399, SD = 0.167) outperformed asocial participants (β = 0.067, 97% HDI = [0.036, 0.098]; SM 2.1.3). SSL participants performed highest (M = 0.493, SD = 0.216), outperforming asocial (β = 0.162, 97% HDI = [0.13, 0.193]) and RM (β = 0.095, 97% HDI = [0.095, 0.107]; SM 2.1.3) participants.

By the final generations (9 to 12), performance among SSL (M = 0.514, SD = 0.216) and RM (M = 0.433, SD = 0.163) participants increased substantially relative to asocial participants (SSL: β = 0.181, 97% HDI = [0.147, 0.221]; RM: β = 0.101, 97% HDI = [0.071, 0.132]; SM 2.1.5 and 2.1.6). This was not just a consequence of accumulating more knowledge of the task: In a follow-up experiment, asocial participants completed 52 trials (equivalent to four generations) yet performed worse than SSL participants (but better than RM participants; SM 1.3). In the SSL group, but not the RM group, a distinct population of high performers emerged (performance > 0.6; Fig. 2B). Most of these participants (85.2%) learned from a high-performing demonstrator (Fig. 2C). However, most people who learned from

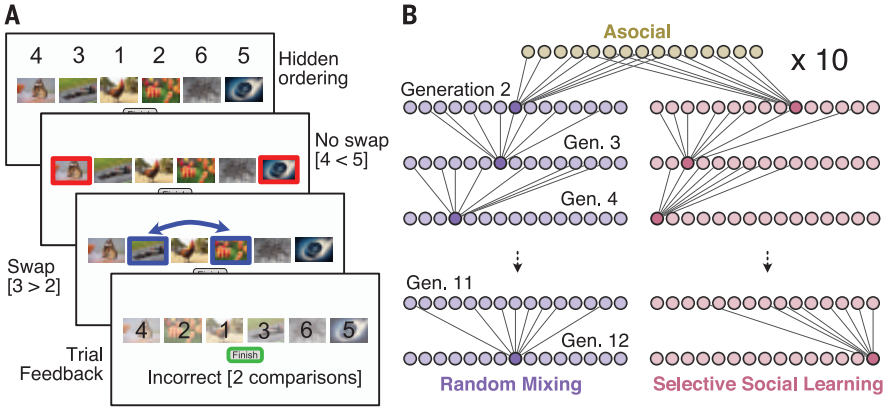


Fig. 1. Problem-solving task and study design. (**A**) Sorting task. (**B**) Participants chose up to three demonstrators from a sample of eight (out of 15) members of the preceding generation. Ten SSL networks were paralleled by 10 RM networks yoked to shared initial generations.

[1]Department of Psychology, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Computer Science, Princeton University, Princeton, NJ, USA. [3]Department of Psychology, Princeton University, Princeton, NJ, USA.
*Corresponding author. Email: wdt@berkeley.edu
†These authors contributed equally to this work.

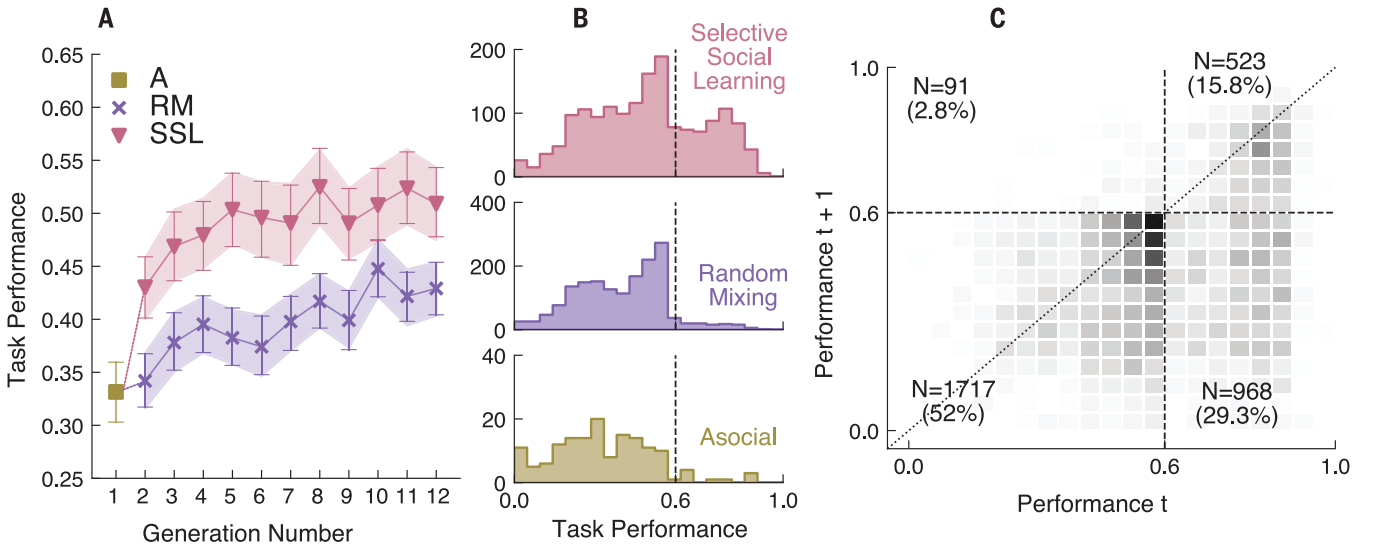


Fig. 2. Performance in the sorting task. (**A**) Performance over generations. Error bars show 95% confidence intervals. A, asocial. (**B**) Performance among SSL (top), RM (middle), and asocial (bottom) participants. The vertical axis shows the number of participants. The vertical dashed line indicates the high-performance threshold (0.6). (**C**) Participant performance relative to their demonstrators. t, time (generation number).

a high-performing demonstrator were not themselves high performing (64.9%). This asymmetry indicated an exceptional strategy that is rarely discovered and difficult to pass on.

To deepen our understanding of these strategies, we analyzed the algorithmic structure of participant responses. Using recurrent neural network models (SM 4), we identified eight classes of algorithms (SM 4.3). Participants discovered several known sorting algorithms. Two algorithms were more frequent than all others. The first algorithm is known in computer science as selection sort. Selection sort generates 15 specific comparisons (1-2, 1-3, 1-4, 1-5, 1-6, 2-3, 2-4, 2-5, 2-6, 3-4, 3-5, 3-6, 4-5, 4-6, and 5-6; Fig. 3A) that are guaranteed to succeed (performance = 0.55). The second algorithm is an even more efficient solution, known as gnome sort. The comparisons implied by gnome sort depend on the outcome of earlier comparisons, so its behavior varies. Participants using gnome sort made specific sweeps of adjacent comparisons until a pair failed to swap, remembering where each sweep started (Fig. 3B).

Overall, 80% of participants who used an identifiable algorithm used selection sort or gnome sort (SM 4.2.1). Figure 4 shows major independent lineages of algorithms in four networks (all networks are shown in the SM). Only 44% of asocial participants (66 of 150) used an identifiable algorithm: Among asocial participants, 16% (26) used selection sort and just 13% (20) used gnome sort. All other algorithms were even rarer. In early generations, selection sort began to spread in both the RM and SSL groups (SM 2.2.2), independently becoming the most frequent algorithm in 7 of 10 network pairs. In the RM group, selection sort continued to spread, dominating 9 of

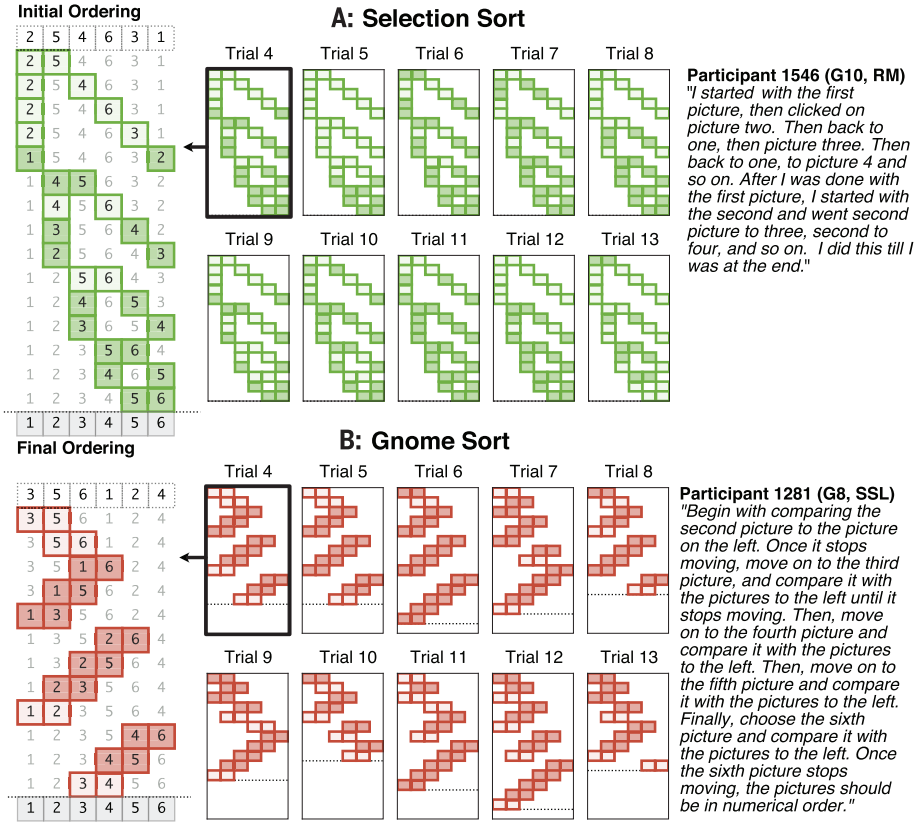


Fig. 3. Selection sort and gnome sort. (**A** and **B**) Strategy descriptions and executions by participants using selection sort [(A); participant 1546] and gnome sort [(B); participant 1281]. Squares show swap (solid) or no swap (transparent). G, generation.

10 populations by generations 9 to 12 (67% of identifiable algorithms). However, in the SSL group, selection sort began to decline after generation 3 (Fig. 4). In the SSL group, but not the RM group (SM 2.2.3), gnome sort dominated 8 of 10 populations. By the final two generations, more than half of all SSL participants used gnome sort.

Participants successfully transmitted selection sort and gnome sort from person to

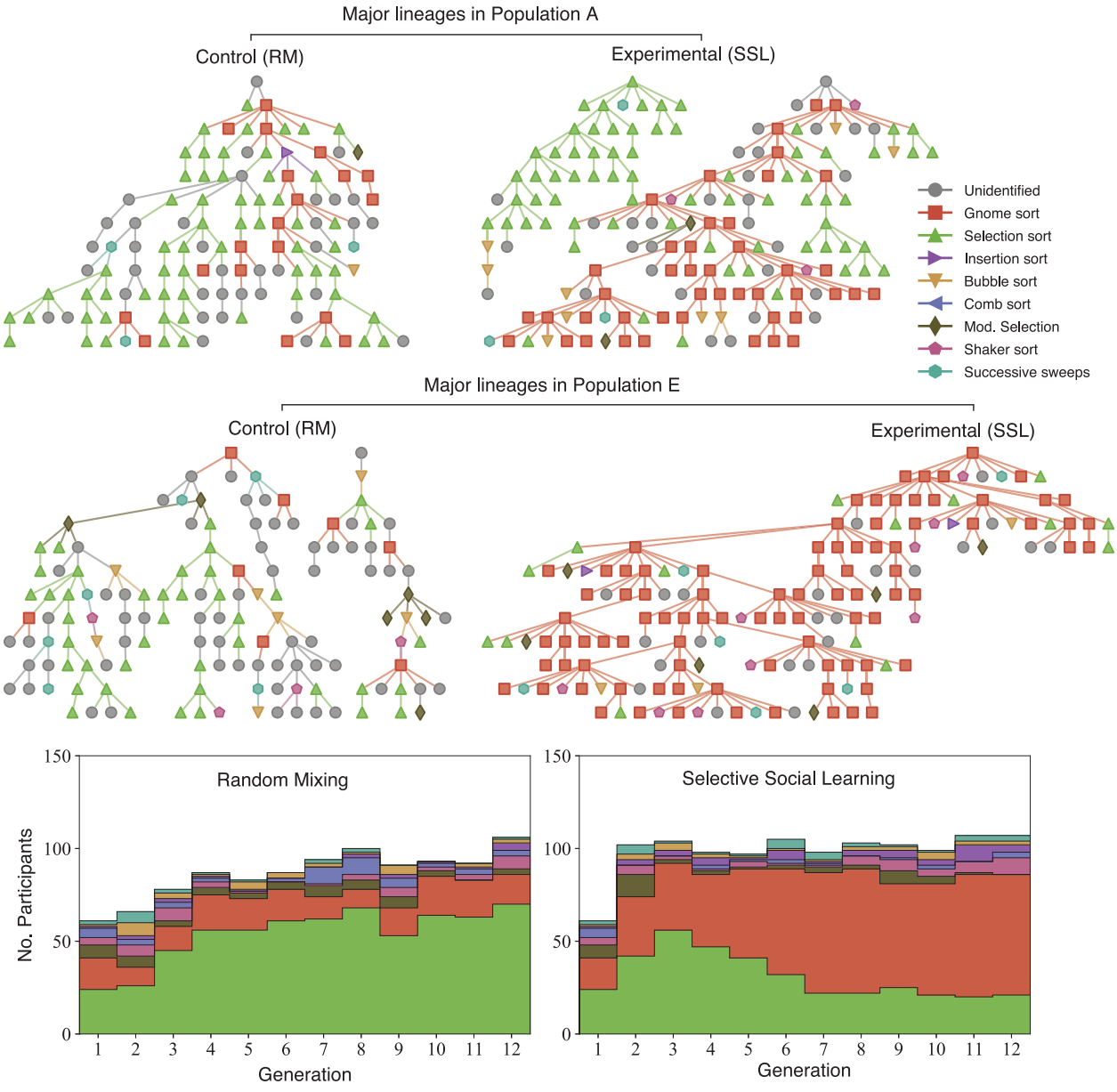


Fig. 4. Algorithm lineages. (**Top**) Major lineages (connected subcomponents exceeding 20 participants) in two network pairs. (**Bottom**) Algorithm frequencies under RM (left) and SSL (right).

person. However, transmission fidelity differed (SM 2.3). Selection sort uses a fixed comparison sequence, so it is easier to describe and learn. Several algorithms (e.g., comb sort, bubble sort, insertion sort; SM 4.3) share this property but were rare, suggesting that selection sort is particularly intuitive or memorable [consistent with sequence-learning studies (*29*) and mathematical analyses of attractors (*30*)]. Gnome sort was harder to pass on. Its conditional logic improved performance (sometimes needing only five comparisons) but made the algorithm harder to describe, learn, and use (SM 7). A follow-up study (SM 1.4) showed that transmission of complex algorithms in this task was primarily facilitated by participant demonstrations, not written descriptions (though descriptions do improve transmission; SM 2.1.8). Finally, we found that selective social learning counteracted differences in algorithm transmission fidelity. SSL participants chose higher-performing demonstrators and therefore encountered gnome sort more often (SM 2.4), offsetting the complexity disadvantage at the population level. Numerical simulations support this dynamic (SM 8).

Our study offers insight into one potential mechanism linking sociality and the transmission of complex discoveries, supporting assumptions of mathematical models (*31*). Random mixing led to an evolutionary process shaped by how difficult the algorithms were to learn and convey, favoring highly transmissible algorithms (e.g., selection sort). By contrast, selective social learning led to an evolutionary process heavily influenced by how efficiently participants could solve the underlying problem, favoring more complex algorithms (e.g., gnome sort).

Sequential decision-making problems arise in many higher cognitive functions such as planning, social interaction, navigation, and tool use. Like the sorting problem, these challenges call for procedural strategies that we can follow to reliably achieve a goal when interacting with a dynamic task. Our study showed that a simple form of selective social learning helped populations establish such

solutions by increasing the take-up of rare but innovative algorithms. Alternative routes to increased uptake, such as forms of content-biased learning (*13*), may have analogous evolutionary consequences.

Our study does not address potential consequences of population size, demographic profiles, or more complex forms of demonstrator selection that may arise in richer contexts (e.g., strategies that account for potential relationships between expertise and teaching skills). Our findings highlight interactions between reconstructive and preservative aspects of cultural transmission (*30*). Choosing who to learn from helps complex ideas be preserved and built on by future generations in evolving populations.

REFERENCES AND NOTES

1. J. Henrich, *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter* (Princeton Univ. Press, 2015).
2. K. N. Laland, *Darwin's Unfinished Symphony* (Princeton Univ. Press, 2017).
3. M. Derex, M.-P. Beugin, B. Godelle, M. Raymond, *Nature* **503**, 389–391 (2013).
4. A. Mesoudi, A. Thornton, *Proc. Biol. Sci.* **285**, 20180712 (2018).
5. C. A. Caldwell, E. Renner, M. Atkinson, *Rev. Phil. Psychol.* **9**, 751–770 (2018).
6. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, *Science* **350**, 1332–1338 (2015).
7. D. Sperber, N. Claidière, *Biol. Philos.* **23**, 283–292 (2008).
8. J. Henrich, *Am. Antiq.* **69**, 197–214 (2004).
9. M. Muthukrishna, B. W. Shulman, V. Vasilescu, J. Henrich, *Proc. Biol. Sci.* **281**, 20132511 (2013).
10. H. M. Lewis, K. N. Laland, *Philos. Trans. R. Soc. London Ser. B* **367**, 2171–2180 (2012).
11. A. Mesoudi, *Philos. Trans. R. Soc. London Ser. B* **376**, 20200053 (2021).
12. D. Nettle, *Philos. Trans. R. Soc. London Ser. B* **375**, 20190358 (2020).
13. R. Boyd, P. J. Richerson, *Culture and the Evolutionary Process* (Univ. Chicago Press, 1985).
14. C. Heyes, *Cognitive Gadgets* (Harvard Univ. Press, 2018).
15. R. L. Kendal *et al*., *Trends Cogn. Sci.* **22**, 651–665 (2018).
16. L. A. Scanlon, A. Lobb, J. J. Tehrani, J. R. Kendal, *Evol. Human Sci.* **1**, E17 (2019).
17. T. J. H. Morgan *et al*., *Nat. Commun.* **6**, 6029 (2015).
18. P. E. Smaldino, *Trends Cogn. Sci.* **23**, 818–820 (2019).
19. S. Kirby, M. Tamariz, H. Cornish, K. Smith, *Cognition* **141**, 87–102 (2015).
20. D. J. Yeh, L. Fogarty, A. Kandler, *Proc. Biol. Sci.* **286**, 20191951 (2019).
21. J. Winters, *Palgrave Commun.* **5**, 1–13 (2019).
22. C. H. Legare, M. Nielsen, *Trends Cogn. Sci.* **19**, 688–699 (2015).
23. R. McElreath *et al*., *Philos. Trans. R. Soc. London Ser. B* **363**, 3515–3528 (2008).
24. A. Almaatouq *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11379–11386 (2020).
25. C. M. Wu, E. Schulz, M. Speekenbrink, J. D. Nelson, B. Meder, *Nat. Hum. Behav.* **2**, 915–924 (2018).
26. C. A. Caldwell, A. E. Millen, *Evol. Hum. Behav.* **29**, 165–171 (2008).
27. F. Khatib *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18949–18953 (2011).
28. H. Miton, M. Charbonneau, *Proc. Biol. Sci.* **285**, 20180677 (2018).
29. N. Claidière, K. Smith, S. Kirby, J. Fagot, *Proc. Biol. Sci.* **281**, 20141541 (2014).
30. R. Boyd, J. Henrich, *J. Cogn. Cult.* **2**, 87–112 (2002).
31. A. Powell, S. Shennan, M. G. Thomas, *Science* **324**, 1298–1301 (2009).

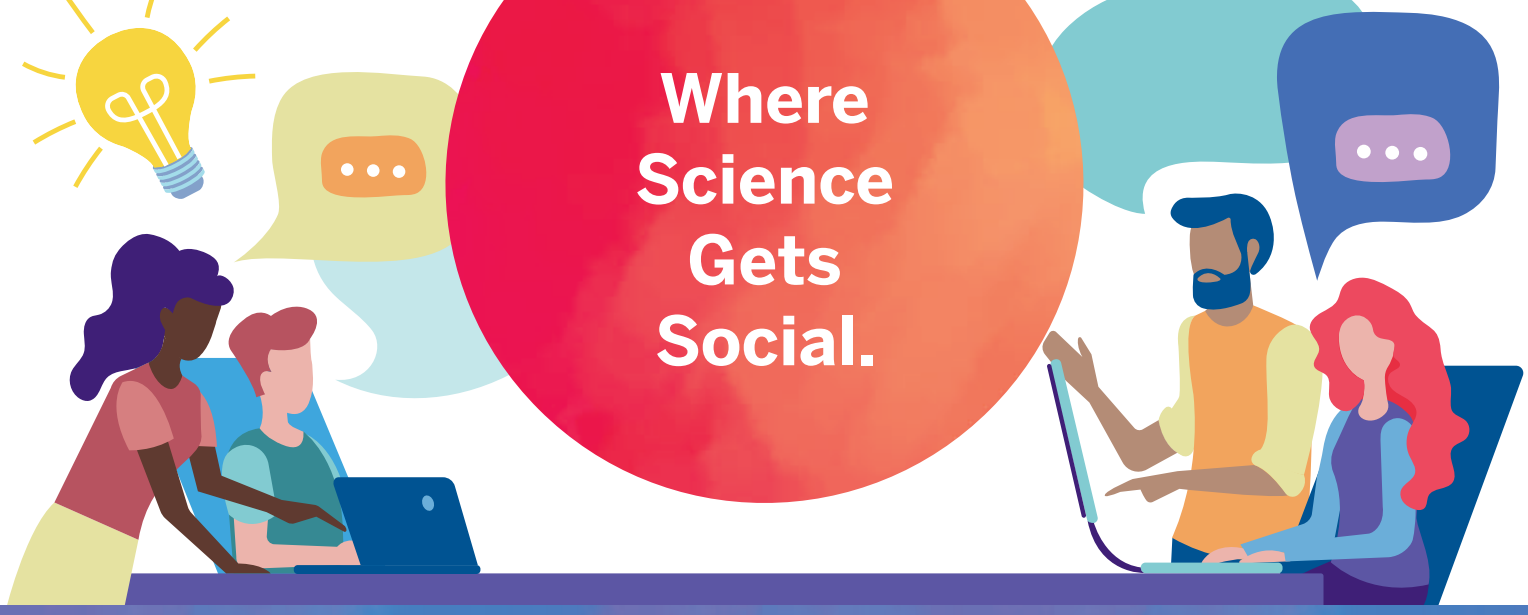
ACKNOWLEDGMENTS

Funding: This work was supported by Defense Advanced Research Projects Agency (DARPA) agreement D17AC00004 and the Schmidt Family Foundation (T.L.G.) and a National Defense Science and Engineering Graduate Fellowship (T.S.). **Author contributions:** Conceptualization: B.T., B.v.O., T.L.G.; Methodology: B.T., B.v.O., T.S., T.L.G.; Investigation: B.T., B.v.O., T.S., T.L.G.; Visualization: B.T., B.v.O., T.S., T.L.G.; Writing: B.T., B.v.O., T.S., T.L.G. **Competing interests:** None declared. **Data and materials availability:** Data, materials, and study registrations are available through the Open Science Framework at https://osf.io/nbc7a/. All other data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn0915
Materials and Methods
Figs. S1 to S55
Tables S1 to S15
References (*32*, *33*)

3 November 2021; accepted 25 February 2022
10.1126/science.abn0915



AAAS.ORG/COMMUNITY







AAAS' Member Community is a one-stop destination for scientists and STEM enthusiasts alike. It's "Where Science Gets Social": a community where facts matter, ideas are big and there's always a reason to come hang out, share, discuss and explore.

Member COMMUNITY
AAAS

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

ScienceCareers online @sciencecareers.org

USF HEALTH

CARDIOVASCULAR RESEARCH FACULTY – OPEN RANK

The Department of Pathology and Cell Biology at the University of South Florida, Morsani College of Medicine is seeking applicants for a **full-time, tenured or tenure track, Open Rank faculty position** in a broad area of cardiovascular research. We are particularly interested in researchers utilizing animal models to study cardiovascular development and disease. This position will be academically appointed to the Department of Pathology and Cell Biology and will include opportunities to interact with researchers both on the main USF campus and the USF Health Heart Institute. Opened in 2020 in downtown Tampa, the USF Health Heart Institute is a modern research facility within four floors totaling 100,000 square feet dedicated to cardiovascular and related research, including clinical trials targeting new ways to manage, prevent, and cure cardiovascular disease.

DUTIES: The primary duties of this position are research-related, including performing basic and translational research related to heart disease, publishing in major peer-reviewed journals and maintaining NIH and other grant funding. Administrative duties may include serving on Institute, College or University committees and councils.

REQUIREMENTS: Applicants are required to have a Doctoral degree from an accredited institution or the highest degree appropriate in the field of specialization. Evidence of scholarly activity, sustained independent research funding, and professional research service are also preferred as is experience utilizing animal models.

Postdoctoral research experience and at least one externally funded award (NIH R01 or equivalent, K99/R00 awards are acceptable for applicants at the Assistant Professor level) are required. Must meet university criteria for appointment to the rank of Assistant, Associate, or Full Professor. For Associate/Full Professor evidence of creative work, professional writing or research in refereed and other professional journals, and recognized authority in the field of specialization is required.

HOW TO APPLY: Applications will only be accepted through the USF Careers System. To apply, please visit: **https://gems.usf.edu/ Job Posting ID 29861**. In addition to completing the required online application, please include a cover letter, a description of research interests (limit to 3 pages or less), current curriculum vitae, and contact information for three references. These documents must be consolidated into one attachment.

WORKING AT USF: With more than 16,000 employees in the USF system, the University of South Florida is one of the largest employers in the Tampa Bay region. At USF you will find opportunities to excel in a rich academic environment that fosters the development and advancement of our employees. We believe in creating a talented, engaged and driven workforce through on-going development and career opportunities. We also offer a first class benefit package that includes medical, dental and life insurance plans, retirement plan options, tuition program and generous leave programs and more.

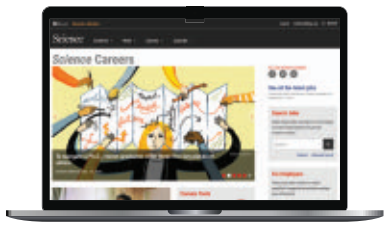
The Tampa metropolitan area is rapidly expanding and provides a culturally diverse environment that offers plenty of sunshine, world famous beaches, arts and entertainment, and an annual average temperature of 72 degrees.



Science Careers helps you advance your career. Learn how!

- Register for a free online account on **ScienceCareers.org**.
- Search hundreds of job postings and find your perfect job.
- Sign up to receive e-mail alerts about job postings that match your criteria.
- Upload your resume into our database and connect with employers.
- Watch one of our many webinars on different career topics such as job searching, networking, and more.
- Download our career booklets, including Career Basics, Careers Beyond the Bench, and Developing Your Skills.
- Complete an interactive, personalized career plan at "my IDP."
- Visit our Employer Profiles to learn more about prospective employers.
- Read relevant career advice articles from our library of thousands.

Visit **ScienceCareers.org** today — all resources are free



SCIENCECAREERS.ORG

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

Tulane University
SCHOOL OF MEDICINE

PROFESSOR OF THE PRACTICE IN PHARMACOLOGY ON THE NON-TENURE TRACK

The Department of Pharmacology at Tulane University School of Medicine invites applications for a full-time, 12-month position as **Professor of the Practice in Pharmacology** on the non-tenure track. The successful candidate will serve as a Principal Teacher and Coordinator of Medical and Graduate Teaching. This position will begin no later than **August 1, 2022**.

The Department of Pharmacology has a great history of excellence in research and education in biomedical and medical pharmacology and is committed to providing high-quality programming at the graduate level. The Department of Pharmacology has responsibility for teaching courses in the medical and graduate degree programs focusing on Pharmacology. Current instructional methodologies used in teaching these courses include lectures, small group sessions, team-based learning (TBL), flipped classrooms, and simulation.

The appointed candidate will serve as a Principal Teacher and Coordinator of Medical and Graduate teaching and is expected to participate in the following: curriculum development and implementation, graduate and undergraduate academic advising and mentoring, departmental and school committees, and scholarly activities.

The requirements for this position include having either a Ph.D., M.D., D.M.D., D.O degree, or equivalent. Additionally, prior experience in teaching Pharmacology in a health professional program is required. Preference will be given to candidates who are outstanding teachers.

Applicants are expected to submit a full application that includes: a cover letter; a curriculum vitae; a statement of teaching experience and teaching philosophy; and previous teaching evaluations or other evidence of teaching excellence, a summary of research/scholarly accomplishments, interests, and future plans. Also required is the contact information for 3-5 professional references including e-mail addresses. Please apply electronically in Interfolio at **http://apply.interfolio.com/103640**. Informal inquiries and further information about the position may be directed to the Department Chair (Dr. David Busija, **dbusija@tulane.edu**) or Director of Graduate Studies (Dr. Craig Clarkson, **cclarks@tulane.edu**). Review of applications will begin immediately and will continue until the position is filled.

Employment Equity Statement: Tulane University is an Affirmative Action and Equal Opportunity Employer committed to excellence through diversity, Tulane University is strongly committed to employment equity within its community and to recruiting diverse faculty and staff. The university encourages applications from all qualified candidates including women, under-represented minorities, such as BIPOC (Black, Indigenous, and People of Color), LGBTQ, persons with disabilities and protected veterans.

Who's the top employer for 2021?

Science Careers' annual survey reveals the top companies in biotech & pharma voted on by *Science* readers.

Read the article and employer profiles at sciencecareers.org/topemployers

Science 2021 TOP EMPLOYER



Shenzhen Institute of Advanced Technology
Chinese Academy of Sciences



Established in partnership between the Chinese Academy of Sciences and the Shenzhen Municipal Government, the Shenzhen Institute of Advanced Technology (SIAT) is a newly-created university with an objective to become the world's preeminent institute for emerging science and engineering programs. SIAT is equipped with state-of-art teaching and research facilities and is dedicated to cultivating international, visionary, and interdisciplinary talents while delivering research support to pursue innovation-driven development.

SIAT is located in Shenzhen, also known as the "Silicon Valley of China," a modern, clean, and green city, well-known for its stunning architecture, vibrant economy, and its status as a leading global technology hub. SIAT is seeking applications for faculty positions of all ranks in the following academic programs: Computer Science and Engineering, Bioinformatics, Robotics, Life Sciences, Material Science and Engineering, Biomedical Engineering, Pharmaceutical Sciences, Synthetic Biology, Neurosciences, etc. SIAT seeks individuals with a strong record of scholarship who possess the ability to develop and lead high-quality teaching and research programs. SIAT offers a comprehensive benefits package and is committed to faculty success throughout the academic career trajectory, providing support for ambitious and world-class research projects and innovative, interactive teaching methods.

Further information:
https://www.siat-sz.edu.cn/

By **Kyra H. Kim**

Bucking tradition

I was born in 1990—the year of the white horse. In the lunar zodiac system, the white horse reigns once every 60 years and is believed to be much stronger and faster than all other stallions. For boys born that year, it is considered a good sign. Not so for girls. A girl born under this zodiac is believed to have a wild, steedlike spirit that will cause her to buck away her fortune, destining her family to trouble. These beliefs were indoctrinated into me as a girl growing up in South Korea. I was repeatedly told to fight against my steedlike spirit by striving to be still, gentle, and quiet. It's an attitude that nearly prevented me from pursuing a career in the geosciences.

When I was young, I loved exploring the mountainous trails behind our house, examining the various soil layers in nearby ditches, and generally running wild outdoors. I often returned home with ripped tights and soiled clothes. "Maybe it's the white horse in you," my mother would say resignedly.

My family elders didn't see that side of me as much, but I frequently heard about the white horse from them as well. They boasted how "open minded" our family was for defying the omen and not terminating the pregnancy that led to my birth. "Many girls did not make it to birth that year," they'd say. "But we don't believe in that myth—our women are disciplined and gentle."



"My wild, steedlike spirit wasn't something to suppress."

I came to believe that my curious nature and desire to explore were not positive attributes—that they needed to be tamed. It didn't help that Korean culture frowned on women pursuing strenuous activities; anything that involved heavy lifting, running, or even sweating was best left to men, traditional views held. I lived in unknowing submission to these ideologies for a long time, struggling with my desire to roam, in body and spirit.

In high school, I settled on a quiet career choice: I would become a lawyer specializing in human rights and environmental issues. My elders praised me for a decision they imagined would lead me to a respectable desk job. However, when I was accepted into a prelaw program and saw the curriculum, I felt a pang of disappointment: The courses struck me as mundane. That's when a mentor nudged me in a different direction. "You could always get a science degree," she said, "and go back into law."

I took her advice and landed in the United States to study geology. Shortly thereafter, though, culture shock set in. I was not ready for the intense physical requirements of my field courses. One 6-week summer course required strenuous hikes, camping in extreme heat, and heavy lifting. My cultural upbringing had discouraged such "wild" activities and I had never gone on extended hikes or camped before.

I blundered a lot in the beginning. One night, I nearly set my tent on fire. But over time my abilities grew, along with my confidence. The other women in my program were an inspiration. I saw beauty in their strength and fitness, and I wanted to emulate them. It was freeing to know I had found an environment where I could participate in physically demanding activities and not feel ashamed of it—but rather, feel empowered.

By the end of my undergraduate degree, I had fully embraced my love of being a rugged, hand-chapped, rock hammer–wielding girl. And I'd given up on the idea of becoming a lawyer. My parents didn't quite know what to make of my adventurous field life. But they were supportive when I told them about the change in my career direction. "So, a Ph.D., right?" they asked, trying to reassure themselves that my activities were leading me down an academic path. "Yes, that's the plan," I responded.

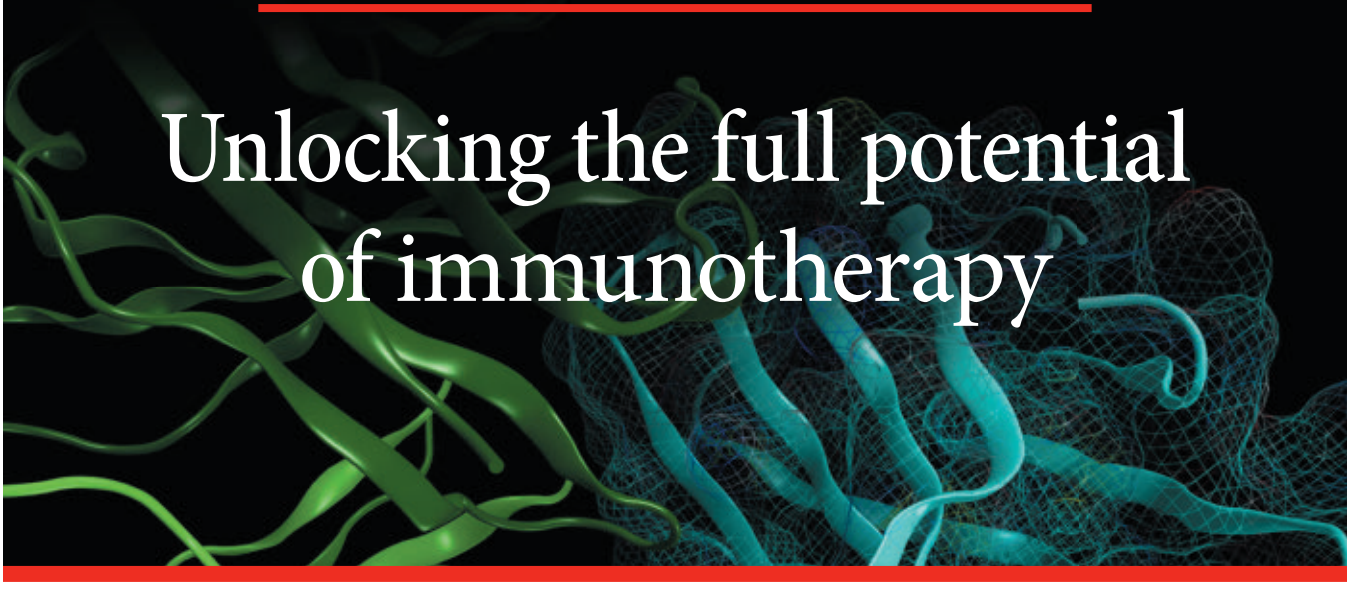
I'm now a postdoc with years of experience collecting samples in challenging field environments, and I'm thankful I bucked against the cultural expectations placed on me. My wild, steedlike spirit wasn't something to suppress. Instead, it led me to a career that's a perfect fit for me. ■

Kyra H. Kim is a postdoctoral researcher at NASA's Jet Propulsion Laboratory at the California Institute of Technology. Do you have an interesting career story to share? Send it to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER

The next breakthrough in **Making Cancer History®**

Unlocking the full potential of immunotherapy



MD Anderson Cancer Center proudly announces the launch of our global immunotherapy research and innovation hub, the James P. Allison Institute.

Named for our Nobel laureate and world-renowned researcher Dr. Jim Allison, the institute will advance exceptional discovery, translational and clinical research to integrate immunobiology across disciplines and bring the benefits of this novel therapy to more cancer patients.

Immunotherapy is a pioneering treatment that unleashes the patient's own immune system to attack and eliminate cancer. The Allison Institute will bridge all research areas, forging new breakthroughs that continue to transform cancer care.

With the launch of the Allison Institute, the world can begin to realize an end to cancer, which has been MD Anderson's singular mission for more than 80 years.

James P. Allison Institute | THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

MDAnderson.org/AllisonInstitute

CALL FOR PAPERS

SPJ
SCIENCE PARTNER JOURNALS

PEKING UNIVERSITY 1898



Health Data Science

Health Data Science is an online-only Open Access Science Partner Journal published in affiliation with **Peking University (PKU)** and distributed by the **American Association for the Advancement of Science (AAAS)**. This journal is committed to promoting the appropriate application of cutting-edge technologies and analytic approaches for the health field, translating the value of health data into practice for the benefit of patients' health and populations' well-being and advancing health data science. *Health Data Science* publishes innovative, scientifically-rigorous original research, comprehensive reviews, insightful perspectives, viewpoints and editorials. The topics of interest include but are not limited to: data management, governance and provenance; comparative effectiveness research; ethics issues; public health and health policies; application of cutting-edge technologies in healthcare and more.

Submit your research to *Health Data Science* today!

Learn more at **spj.sciencemag.org/hds**

The Science Partner Journal (SPJ) program was established by the American Association for the Advancement of Science (AAAS), the nonprofit publisher of the *Science* family of journals. The SPJ program features high-quality, online-only, Open Access publications produced in collaboration with international research institutions, foundations, funders and societies. Through these collaborations, AAAS furthers its mission to communicate science broadly and for the benefit of all people by providing top-tier international research organizations with the technology, visibility, and publishing expertise that AAAS is uniquely positioned to offer as the world's largest general science membership society.

Visit us at **spj.sciencemag.org**

@SPJournals **@SPJournals**

OPEN ACCESS

ARTICLE PROCESSING CHARGES WAIVED THROUGH DECEMBER 2024