NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

PUBLIC WORKSHOP BRIEFING BOOK

June 14, 2022 12:30 – 5PM ET

WEBCAST LINK:

https://www.nationalacademies.org/event/06-14-2022/use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research-meeting-3-and-public-workshop

***Questions for speakers can be submitted in the Slido box under the webcast**

NATIONAL ACADEMIES Sciences Engineering Medicine

Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

Public Workshop: June 14, 2022

Table of Contents

AGENDA

NATIONAL ACADEMIES Sciences Engineering Medicine

Board on Health Sciences Policy | Committee on Population

Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

Public Workshop: June 14, 2022

CLICK HERE TO REGISTER

TUESDAY, JUNE 14, 2022 12:30 – 5 PM ET

12:30–12:40 PM ET

Welcome and Goals for the Workshop

Charmaine Royal, *Committee Co-Chair*
Robert O. Keohane Professor of African & African American Studies, Biology, Global Health, and Family Medicine & Community Health
Director, Duke Center on Genomics, Race, Identity, Difference and Duke Center for Truth, Racial Healing & Transformation
Duke University

Aravinda Chakravarti, *Committee Co-Chair*
Director, Center for Human Genetics and Genomics Muriel G & George W Singer Professor of Neuroscience & Physiology
New York University Grossman School of Medicine

SESSION I: EXAMINING USE OF POPULATION DESCRIPTORS IN GENOMICS RESEARCH

Moderator: John Novembre, University of Chicago

Objectives

- To explore what types of population descriptors are needed for genetics and genomics studies
    - What is a genetics study trying to accomplish?
    - Who is sampled? Why are they sampled? What are participants called, and why?
- To examine how and why genetics studies should or should not incorporate social categories and environmental factors

12:40–12:45 PM

Brief Introduction to the Session

John Novembre
Professor, Department of Human Genetics, Department of Ecology & Evolution
University of Chicago

12:45–1:45 PM

Speakers' Talks

Gil McVean
Professor of Statistical Genetics
Director, Big Data Institute
Fellow of Linacre College
University of Oxford

Akinyemi Oni-Orisan
Assistant Professor, Department of Clinical Pharmacy
University of California, San Francisco School of Pharmacy

Nancy Cox
Director, Vanderbilt Genetics Institute
Director, Division of Genetic Medicine
Mary Phillips Edmonds Gray Professor of Genetics
Vanderbilt University

Graham Coop
Professor, Department of Evolution and Ecology and Center for Population Biology
University of California, Davis

1:45–2:25 PM **Q&A with speakers**

2:25–2:40 PM **Break**

SESSION II: USE OF POPULATION DESCRIPTORS BY BIOBANKS AND OTHER RESEARCH CONSORTIA

Moderator: Ann Morning, New York University

Objectives

- To examine how and why biobanks use taxonomies currently, especially in areas of large diversity
- To explore how legacy data might be managed and merged with future data
- To learn how large-scale data collection projects are designed

2:40–2:45 PM **Brief Introduction to the Session**
Ann Morning
Associate Professor, Department of Sociology
Academic Director, 19 Washington Square North (NYU Abu Dhabi in NY)
New York University

2:45–3:30 PM **Speakers' Talks**
Phil Tsao
Professor (Research), Medicine – Cardiovascular Medicine
Stanford University

Alice Popejoy
Assistant Professor, Department of Public Health Sciences
University of California, Davis

Mashaal Sohail
Associate Professor, Center for Genomic Sciences
National Autonomous University of Mexico

3:30–4:00 PM **Q&A with Speakers**

4:00–4:10 PM **Break**

SESSION III: COMMUNITY INPUT ON POPULATION DESCRIPTORS IN GENOMICS RESEARCH

Moderator: Charmaine Royal, Duke University

Objective

- To hear from a variety of stakeholders on the following topics:
  - What works and doesn't work about the current population descriptors used in genomics research?
  - What could be improved in current use of population descriptors in genomics research?

4:10–4:15 PM

Introduction to the Session

Charmaine Royal
Robert O. Keohane Professor of African & African American Studies, Biology, Global Health, and Family Medicine & Community Health
Director, Duke Center on Genomics, Race, Identity, Difference and Duke Center for Truth, Racial Healing & Transformation
Duke University

4:15–4:55 PM

Speakers' Comments

Jennifer Webster
Senior Director, Precision Medicine RWE Lead
Pfizer

Santiago Molina
Postdoctoral Fellow, Sociology/Science in Human Culture
Department of Sociology
Weinberg College of Arts & Sciences
Northwestern University

Norbert Tavares
Program Manager, Cell Biology
Chan-Zuckerberg Initiative

King Jordan
Professor, School of Biological Sciences
Director, Bioinformatics Graduate Program
Georgia Institute of Technology

Dianalee McKnight
Medical Affairs Director, Emerging Clinical Omics
Invitae

Ramya M. Rajagopalan
Associate Director, Training, Evaluation, and Qualitative Research
Center for Empathy and Technology, T. Denny Stanford Institute for Empathy and Compassion
University of California, San Diego

Stacy Christiansen
Managing Editor, *JAMA*
Chair, *AMA Manual of Styles*

Hannah Wand
Director, Preventive Genomics Program
Genetics Counselor
Stanford Health Care

4:55–5:00 PM **Concluding Remarks**

5:00 PM **Adjourn**

STUDY INFORMATION

NATIONAL ACADEMIES Sciences Engineering Medicine

Board on Health Sciences Policy
Committee on Population

Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

Statement of Task

An ad hoc committee under the auspices of the National Academies of Sciences, Engineering, and Medicine's Health and Medicine Division will convene to review and assess the existing methodologies, benefits, and challenges in the use of race and ethnicity and other population descriptors in genomics research. The committee work will focus on, but not be limited to the following tasks:

1. Document and evaluate the variety of population descriptors currently used in genomics research and the potential benefits and challenges of changing these descriptors.
2. Assess how race, ethnicity, and genetic ancestry are currently being used as population descriptors in health disparities research to study genetics and genomics.
3. Assess the appropriate use of race, ethnicity, and genetic ancestry as population descriptors in the determination of genetic risk scores and health risk.
4. Develop feasible and logical approaches to advance appropriate use of race and ethnicity and alternative population descriptors in published genomics research studies.
5. Examine the potential of new, culturally responsive methods and common data elements (CDEs) for advancing harmonization of population descriptors in large genomic studies in the United States and globally.
6. Assess when it is appropriate to use race and ethnicity as population descriptors in genetic and genomic research, and provide recommendations to scientists and researchers for future research.
7. Propose best practices for domestic and international harmonization of population group descriptors.
8. Assess the scientific knowledge of the relationships among race, ethnicity, and population genetic variation.
9. Identify and discuss potential obstacles to implementation of the new methods to describe populations.
10. Discuss potential implementation strategies to help enhance the adoption of best practices by the research community.
11. Identify obstacles and propose best practices in the use of population descriptors with legacy biological samples and associated data.

The final report should describe best practices on the use of race, ethnicity, and genetic ancestry and other population descriptors in genetics and genomics research, as formulated by the committee. Attention should be given to how these best practices could be used by biomedical and scientific communities to increase the robustness of study designs and methods for genetics and genomics research in the United States and globally.

These elements are beyond the scope of this consensus study:

1. Examining the use of race and ethnicity in clinical care
2. Examining racism in science and genomics
3. Examining the use of race and ethnicity in biomedical research generally (non-genetic and genomic research)
4. Providing policy recommendations to NIH and government agencies

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Committee Membership Roster

Aravinda Chakravarti, Ph.D. *(Co-Chair)*
Director, Center for Human Genetics and Genomics
Muriel G & George W Singer Professor of Neuroscience & Physiology
New York University Grossman School of Medicine

Charmaine Royal, Ph.D. *(Co-Chair)*
Robert O. Keohane Professor of African & African American Studies, Biology, Global Health, and Family Medicine & Community Health
Director, Duke Center on Genomics, Race, Identity, Difference and Duke Center for Truth, Racial Healing & Transformation
Duke University

Katrina Armstrong, M.D.
Executive Vice President for Health and Biomedical Sciences
Dean of the Faculties of Health Sciences and the Vagelos College of Physicians and Surgeons
Chief Executive Officer of Columbia University Irving Medical Center
Harold and Margaret Hatch Professor in the Faculty of Medicine
Vagelos College of Physicians and Surgeons
Columbia University Irving Medical Center

Michael Bamshad, M.D.
Professor and Chief, Division of Genetic Medicine
Allan and Phyllis Treuer Endowed Chair in Genetics and Development
University of Washington & Seattle Children's Hospital

Luisa Borrell, Ph.D., D.D.S., M.P.H.
Distinguished Professor, Department of Epidemiology & Biostatistics
Graduate School of Public Health & Health Policy
City University of New York, NY

Katrina Claw, Ph.D.
Assistant Professor, Division of Biomedical Informatics and Personalized Medicine, Department of Medicine
Faculty, Colorado Center for Personalized Medicine
University of Colorado Denver – Anschutz Medical Campus

Clarence Gravlee, Ph.D.
Associate Professor, Department of Anthropology
University of Florida

Mark Douglas Hayward, Ph.D.
Professor of Sociology
Centennial Commission Professor in the Liberal Arts
Faculty Research Associate, Population Research Center
The University of Texas at Austin

Rick Kittles, Ph.D.
Professor and Director of the Division of Health Equities
Department of Population Sciences
City of Hope

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION

BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION

COMMITTEE ON POPULATION

Sandra Soo-Jin Lee, Ph.D.
Professor of Medical Humanities & Ethics
Chief of the Division of Ethics
Department of Medical Humanities & Ethics (MHE)
Vagelos College of Physicians & Surgeons
Columbia University

Andrés Moreno-Estrada, M.D., Ph.D.
Professor
Advanced Genomics Unit
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV)

Ann Morning, Ph.D.
Associate Professor, Department of Sociology
Academic Director, NYU Abu Dhabi in NY
New York University

John Peter Novembre, Ph.D.
Professor, Department of Human Genetics, Department of Ecology & Evolution
University of Chicago

Molly Przeworski, Ph.D.
Professor, Department of Biological Sciences, Department of Systems Biology
Columbia University

Dorothy Roberts, J.D.
George A. Weiss University Professor of Law & Sociology
Raymond Pace & Sadie Tanner Mossell Alexander Professor of Civil Rights
University of Pennsylvania

Sarah A. Tishkoff, Ph.D
David and Ln Silfen University Professor
Departments of Genetics and Biology
Director, Center for Global Genomics & Health Equity
University of Pennsylvania

Genevieve Wojcik, Ph.D.
Assistant Professor of Epidemiology
Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Committee Member Biosketches

Aravinda Chakravarti, Ph.D. is the Director of the Center for Human Genetics & Genomics, and the Muriel G & George W Singer Professor of Neuroscience & Physiology and Professor of Medicine at the New York University Grossman School of Medicine. He has served on the faculty at the University of Pittsburgh (1980 – 1993), Case Western Reserve University (1994-2000), and Johns Hopkins University (2000-2018). He is one of the founding Editors-in-Chief of *Genome Research* and *Annual Reviews of Genomics & Human Genetics*, and is on the advisory boards of numerous national and international Institutes, charities, academic societies, the NIH and biotechnology companies. He has been a key participant in many genome projects, and now works on genome-scale analysis of the molecular basis of human disease. He was the 2008 President of the *American Society of Human Genetics* and been elected to the US National Academy of Science, the US National Academy of Medicine, the Indian National Academy of Science and the Indian Academy of Sciences. He was awarded the 2013 William Allan Award by the *American Society of Human Genetics* and the 2018 Chen Award by the *Human Genome Organization*. Dr. Chakravarti received his Ph.D. in human genetics in 1979.

Charmaine Royal, Ph.D. is the Robert O. Keohane Professor of African & African American Studies, Biology, Global Health, and Family Medicine & Community Health at Duke University. She directs the Duke Center on Genomics, Race, Identity, Difference and the Duke Center for Truth, Racial Healing & Transformation. She held previous faculty appointments at Howard University. Throughout her career, Dr. Royal has focused on ethical, social, scientific, and clinical implications of human genetics and genomics, particularly issues at the intersection of genetics and "race". Bringing expertise from her work in these areas, she has served as a chair or member of numerous national and international advisory boards and committees for government agencies, professional organizations, not-for-profit entities, and corporations, including the Board of Directors for the American Society of Human Genetics, the Independent Expert Committee for the Human Heredity and Health in Africa (H3Africa) Initiative, and the Ethics Advisory Board for Illumina, Inc. Dr. Royal obtained a bachelor's degree in microbiology, master's degree in genetic counseling, and doctorate in human genetics from Howard University. She completed postgraduate training in ethical, legal, and social implications (ELSI) research and bioethics at the National Human Genome Research Institute of the National Institutes of Health, and in epidemiology and behavioral medicine at Howard University Cancer Center. She was a member of the National Academies' committees that produced 'Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease' and 'Addressing Sickle Cell Disease: A Strategic Plan and Blueprint for Action'.

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION

BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION

COMMITTEE ON POPULATION

Katrina A. Armstrong, M.D. leads Columbia University's medical campus as the Executive Vice President for Health and Biomedical Sciences. She is Chief Executive Officer of the Columbia University Irving Medical Center and Dean of the Faculties of Health Sciences and Medicine, which includes Columbia's dental, medical, nursing and public health schools. She is an internationally recognized investigator in medical decision making, quality of care, and cancer prevention and outcomes, an award winning teacher, and a practicing primary care physician. She has served on multiple advisory panels for academic and federal organizations and has been elected to the National Academy of Medicine, the American Academy of Arts and Sciences, the Association of American Physicians, and the American Society for Clinical Investigation. Before joining Columbia, Dr. Armstrong was the Jackson Professor of Clinical Medicine at Harvard Medical School, Chair of the Department of Medicine and Physician-in-Chief of Massachusetts General Hospital, and Professor of Epidemiology at the Harvard T.H. Chan School of Public Health. Before joining Harvard, she was Chief of the Division of General Internal Medicine, Associate Director of the Abramson Cancer Center, and Co-Director of the Robert Wood Johnson Clinical Scholars Program at the University of Pennsylvania. She is a graduate of Yale University (BA degree in architecture), Johns Hopkins (MD degree), and the University of Pennsylvania (MS degree in clinical epidemiology). She completed her residency training in internal medicine at Johns Hopkins.

Mike Bamshad, M.D. is Professor and Chief of the Division of Genetic Medicine in the Department of Pediatrics at the University of Washington and Seattle Children's Hospital, and holds the Allan and Phyllis Treuer Endowed Chair in Genetics and Development. Dr. Bamshad is Editor-in-Chief of Human Genetics and Genomics Advances, published by the American Society of Human Genetics. His research focuses on understanding the impact of population structure and natural selection on human genetic variation; developing innovative ways to discover genetic variants underlying monogenic disorders, modifiers of monogenic traits and complex traits; and testing novel ways to translate genomic advances into the practice precision genetic medicine. He and his colleagues pioneered the use of exome and genome sequencing for discovery of genes underlying Mendelian conditions and has contributed to the identification of hundreds of genes for Mendelian disorders. He has also been a leader in understanding the relationship between genetic ancestry and notions of race, developing innovative ways to openly share phenotypic information and genetic data (e.g., MyGene2) and building platforms for self-guided return of genetic testing results (e.g., My46) from exome and whole genome sequencing in both research and clinical settings. He has published more than 300 scientific manuscripts as well as papers in periodicals such as Scientific American, and co-authors a popular textbook entitled Medical Genetics. He received his B.S. and M.D. at the University of Missouri in Kansas City and his M.A. at the University of Kansas.

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION

BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION

COMMITTEE ON POPULATION

Luisa N. Borrell, D.D.S., Ph.D. is a Distinguished Professor in the Department of Epidemiology and Biostatistics, City University of New York Graduate School of Public Health and Health Policy (CUNY SPH), New York, NY. She is a social epidemiologist with a research interest on the role of race/ethnicity, socioeconomic position, and neighborhood effects as social determinants of health. Her work on Hispanics'/Latinos' racial identity brings attention to the need for disaggregated analyses by race as Hispanics/Latinos are a heterogeneous group with a mix of European, Native American and African ancestry. She also has expertise in research methods and analyses of large and spatially-linked datasets. Dr. Borrell is a Fellow of the New York Academy of Medicine. She has a Doctor in Dental Surgery and a Master in Public Health, from Columbia University, New York, NY, as well as doctorate in Epidemiological Science from the University of Michigan, Ann Arbor, MI.

Katrina Claw, Ph.D. is an Assistant Professor in the Division of Biomedical Informatics and Personalized Medicine in the Department of Medicine at the University of Colorado Anschutz Medical Campus. Her research focuses broadly on personalizing medicine, using genetic information and biomarkers for tailored treatment, in relation to pharmacogenomics as well as understanding the ethical, cultural, and social implications of genomic research with populations historically underrepresented in health research. Her current research includes studying cytochrome P450 genetic variation in Indigenous communities (e.g., American Indian and Alaska Native peoples). Her other projects include exploring the perspectives of tribal members on genetic research with tribes and developing guidelines and policies in partnership with tribes. All of her projects strive to use community based participatory research approach and include cultural and Indigenous knowledge. She was awarded the Genomic Innovator Award from NHGRI in 2020 for her work on pharmacogenomics approaches to drug metabolism in American Indian/Alaska Native People. She received her B.S. and B.A. from Arizona State University and her Ph.D. from the University of Washington.

Clarence C. Gravlee, Ph.D. is associate professor in the Department of Anthropology at the University of Florida, where he is also affiliated with the Center for Latin American Studies, the African American Studies Program, and the Genetics Institute. His research examines the genetic and environmental contributors to hypertension in the African Diaspora, with an emphasis on the biological consequences of systemic racism. His work, with collaborators, integrates methods and theory from the social and biological sciences, including ethnography, social network analysis, human biology, and genetics. Gravlee completed a B.A., M.A., and Ph.D. in anthropology at the University of Florida, a Fulbright graduate fellowship at the Universität zu Köln (Cologne, Germany), and postdoctoral training in community-based participatory research as a W.K. Kellogg Community Health Scholar at the University of Michigan School of Public Health.

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Mark D. Hayward, Ph.D. is a professor of sociology and Centennial Commission Professor in the Liberal Arts at the University of Texas at Austin. Hayward is a health demographer. Building on a long-standing interest in the developmental origins of adult health, his current work incorporates biosocial lenses (e.g., pathophysiological pathways and genetic risk) to better understand how social exposures from childhood through adulthood influence racial/ethnic disparities in dementia risk. Hayward is a recipient of the Matilda White Riley Award from the National Institutes of Health for his contributions to behavioral and social scientific knowledge relevant to mission of NIH. He has served on numerous major foundations (Robert Wood Johnson and Pew) and major federal agencies (e.g., the National Institutes of Health and the National Center for Health Statistics). Hayward is the current editor of his field's major journal, Demography, and President-elect of the Interdisciplinary Association of Population Health Science. He received his Ph.D. from Indiana University and his B.A. from Washington State University. He has served on scientific advisory boards at the NASEM including the Committee on Population and a Decadal Survey of Behavioral and Social Science Research on Alzheimer's Disease and Alzheimer's Disease-Related Dementias.

Rick Kittles, Ph.D. is Professor and founding Director of the Division of Health Equities within the Department of Population Sciences at the City of Hope (COH), Associate Director of Health Equities of COH Comprehensive Cancer Center, and Co-founder and Scientific Director of African Ancestry, Inc. His first faculty appointment was at Howard University where he helped establish the National Human Genome Center at Howard University. Dr. Kittles is well known for his research of prostate cancer and health disparities among African Americans, having published over 200 research articles. Dr. Kittles' research has focused on understanding the complex issues surrounding race, genetic ancestry, and health disparities. He has been at the forefront of the development of genetic markers for ancestry and how genetic ancestry can be used in genetic studies on disease risk and outcomes, showing the impact of genetic variation across populations. In 2010 Dr. Kittles was named in Ebony magazine's "The Ebony Power 100." Dr. Kittles presented the Keynote Address to the 2012 United Nations General Assembly, "International Day of Remembrance of Victims of Slavery and the Transatlantic Slave Trade." Recently he was named one of The Huffington Post's "50 Iconic Black Trailblazers Who Represent Every State In America." He received a Ph.D. in Biological Sciences from George Washington University in 1998.

Sandra Soo-Jin Lee, Ph.D. is Professor of Medical Humanities and Ethics and Chief of the Division of Ethics at Columbia University. Trained as a medical anthropologist, Dr. Lee leads interdisciplinary bioethics research on race, ancestry and equity in genomics, precision medicine and artificial intelligence, and publishes in the genomics, medical, bioethics, and social science literatures. Dr. Lee has investigated racial categorization in human genetics for over two decades and co-edited Revisiting Race in a Genomic Age

(2008). Her current NIH funded projects include the Ethics of Inclusion: Diversity in Precision Medicine Research. Dr. Lee is Co-Director of the Center for ELSI Resources and Analysis and the ELSI Congress. She is President-elect of the Association of Bioethics Program Directors and a Hastings Center Fellow. Dr. Lee serves on the US Health and Human Services Secretary's Advisory Committee on Human Research Protections, the Scientific Advisory Boards of the Kaiser Permanente National Research Biobank and the Human Pangenome Reference Consortium, and the editorial boards of the American Journal of Bioethics and Narrative Inquiry in Bioethics. Dr. Lee received her doctorate from the University of California, Berkeley/UCSF joint program in Medical Anthropology and her undergraduate degree in Human Biology from Stanford University.

Andrés Moreno-Estrada, Ph.D., M.D. is the Principal Investigator of the Human Evolutionary and Population Genomics Laboratory at the Advanced Genomics Unit (UGA-CINVESTAV), in Irapuato, Mexico. Previously, he was Research Associate of the Genetics Department at Stanford University until 2014. He is a Mexican population geneticist interested in human genetic diversity and its implications in population history and medical genomics. His work integrates genomics, evolution and precision medicine in projects involving large collections of understudied populations, in particular from the Americas and the Pacific. He authored the most detailed work so far of the genetic structure of the Mexican population, including the first genomic characterization of 20 diverse indigenous groups throughout Mexico, as well as fine-scale studies in the Caribbean region, South America, and Polynesia. He is leading the Human Cell Map of Latin American Diversity to increase the representation of diverse ancestry networks for the Human Cell Atlas project. For his work in Latin America he was awarded the "George Rosenkranz Prize for Health Care Research in Developing Countries" in 2012. He received his M.D. from University of Guadalajara in 2002 and Ph.D. in Evolutionary Genetics from Pompeu Fabra University in 2009. Dr. Moreno was a postdoctoral fellow until 2012 with Prof. Carlos Bustamante at Cornell University and Stanford University School of Medicine.

Ann Morning, Ph.D. is an Associate Professor of Sociology at New York University and the Academic Director of 19 Washington Square North, the home of NYU Abu Dhabi in New York. Trained in demography, her research focuses on race, ethnicity, and the sociology of science, especially as they pertain to census classification worldwide and to individuals' concepts of difference. She is the author of The Nature of Race: How Scientists Think and Teach about Human Difference (University of California Press 2011), and co-author of An Ugly Word: Rethinking Race in Italy and the United States (with Marcello Maneri, forthcoming in 2022 from Russell Sage Foundation). Morning was a 2008-09 Fulbright research fellow at the University of Milan-Bicocca and a 2014-15 Visiting Scholar at the

Russell Sage Foundation. She was a member of the U.S. Census Bureau's National Advisory Committee on Racial, Ethnic and Other Populations from 2013 to 2019 and has consulted on racial statistics for the European Commission, the United Nations, and Elsevier. Morning holds her B.A. in Economics and Political Science from Yale University, a Master's of International Affairs from Columbia University, and her Ph.D. in Sociology from Princeton University.

John Novembre, Ph.D. is a Professor at the University of Chicago in the Departments of Human Genetics and Ecology & Evolution. His research has developed computational methods to answer a diverse range of questions regarding genetic diversity. His work has especially had an impact on the understanding and analysis of geographic patterns in human genetic variation. He has been awarded as a MacArthur Fellow, Searle Scholar, and Sloan Research Fellow, and his research is supported by the National Institutes of Health. Dr. Novembre has authored more than 50 peer-reviewed publications in leading journals, including Nature, Science, Nature Genetics, and the American Journal of Human Genetics. He also serves as an academic editor for the journal Genetics, and previously served on the Scientific Advisory Board for AncestryDNA. He received his B.A. from The Colorado College and his Ph.D. from the University of California-Berkeley.

Molly Przeworski, Ph.D. is a Professor of Biological Sciences at Columbia University. Before moving to Columbia University, she was a faculty member at the University of Chicago as well as at Brown University and the Max Planck Institute for Evolutionary Anthropology in Germany. Her research aims to understand the genetic basis and evolutionary history of heritable differences among individuals; recent work focuses in part on genomic trait prediction in humans and implications. She is the recipient of the Rosalind Franklin Award from the Genetics Society of America, a Sloan Research Fellowship, and Howard Hughes Medical Institute Early Career Scientist Award, and is a member of the American Academy of Arts and Sciences and the National Academy of Sciences. She received a B.A. in Mathematics from Princeton University and a Ph.D. from the Committee on Evolutionary Biology at the University of Chicago, then conducted postdoctoral research in the Mathematical Genetics group of the University of Oxford in the United Kingdom.

Dorothy Roberts, J.D. is the George A. Weiss University Professor of Law & Sociology at University of Pennsylvania, with joint appointments in the Departments of Africana Studies and Sociology and the Law School, where she is the inaugural Raymond Pace and Sadie Tanner Mossell Alexander Professor of Civil Rights. She is also Founding Director of the Penn Program on Race, Science & Society. Author of Fatal Invention: How Science, Politics, and Big Business Re-create Race in the Twenty-First Century, Roberts is an expert

on structural racism in US science and medicine and the use of race as a variable in scientific research. Her research has been supported by the American Council of Learned Societies, National Science Foundation, Robert Wood Johnson Foundation, Fulbright Program, Harvard Program on Ethics & the Professions, and Stanford Center for the Comparative Studies in Race & Ethnicity. Recent honors include 2019 election as a College of Physicians of Philadelphia Fellow, 2017 election to the National Academy of Medicine, 2016 Society of Family Planning Lifetime Achievement Award, 2015 American Psychiatric Association Solomon Carter Fuller Award, and 2011 election as a Hastings Center Fellow. Professor Roberts serves on the advisory board for the Center for Genetics and Society. She received her J.D. from Harvard Law School and her B.A., magna cum laude, Phi Beta Kappa from Yale College.

Sarah Tishkoff, Ph.D. is the David and Lyn Silfen University Professor in Genetics and Biology at the University of Pennsylvania, holding appointments in the School of Medicine and the School of Arts and Sciences. She is also the Director of the Penn Center for Global Genomics & Health Equity. Dr. Tishkoff studies genomic and phenotypic variation in ethnically diverse Africans, using field work, laboratory research, and computational methods to examine African population history, the genetic basis of anthropometric, cardiovascular, and immune related traits, and how humans have adapted to diverse environments and diets. Dr. Tishkoff is a member of the National Academy of Sciences, the American Academy of Arts and Sciences, and the National Academy of Medicine. She is a recipient of an NIH Pioneer Award, a David and Lucile Packard Career Award, a Burroughs/Wellcome Fund Career Award, the ASHG Curt Stern Award, and a Penn Integrates Knowledge (PIK) endowed chair. She is on the NAS Board of Global Health and the Scientific Advisory Board for the Packard Fellowships in Science and Engineering, and is on the editorial boards at Cell, PLOS Genetics, and G3 (Genes, Genomes, and Genetics). She received her Ph.D. in Genetics and M.Phil in Human Genetics from Yale University and her B.S. in Anthropology & Genetics from University of California-Berkeley.

Genevieve L. Wojcik, Ph.D. is an Assistant Professor of Epidemiology at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. As a statistical geneticist and genetic epidemiologist, her research focuses on method development for diverse populations, specifically understanding the role of genetic ancestry and environment in genetic risk in admixed populations. Dr. Wojcik integrates epidemiology, sociology, and population genetics to better understand existing health disparities in minority populations, as well as underserved populations globally. In 2021, she was the recipient of one of NHGRI's Genomic Innovator Awards (R35) to do this work. She is a long-standing member of multiple NHGRI consortia focused on diverse populations, such as the Population

Architecture using Genomics and Epidemiology (PAGE) Study, which was formed by NHGRI over a decade ago to address the lack of genetics research in non-European ancestry populations, and the PRIMED consortium, which began this year to better conduct research around polygenic risk scores in diverse populations. Dr. Wojcik previously served as a consultant with Illumina, Inc. Prior to her faculty appointment, Dr. Wojcik was a postdoctoral research scholar at Stanford University in the Departments of Genetics and Biomedical Data Science. She received her Ph.D. in Epidemiology and M.H.S. in Human Genetics/Genetic Epidemiology from the Johns Hopkins Bloomberg School of Public Health and her B.A. in Biology from Cornell University.

WORKSHOP INFORMATION

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

Public Workshop on Use of Population Descriptors in Genomics Research

June 14, 2022

Speaker Biosketches

Graham Coop, Ph.D. is a professor in the Department of Evolution and Ecology and the Center for Population Biology at University of California, Davis. His research focuses on understanding the evolutionary forces that have shaped genetic differences between individuals, populations, and closely related species. A unifying principal of his work is the blending together of modeling and data analysis. His research interests include; the role of geography in adaption, the impact of Natural Selection on linked polymorphism, the causes and consequence of variation in recombination rates, and the inference of Demographic history for population genetic data. He received his Ph.D. from the University of Oxford in the mathematical genetics group in the Statistics Department.

Nancy Cox, Ph.D. is the director of Vanderbilt Genetics Institute and the Division of Genetic Medicine as well as the Mary Phillips Edmonds Gray Professor of Genetics at Vanderbilt University. Nancy Cox is a quantitative human geneticist with a long-standing research program in identifying and characterizing the genetic component to common human diseases; current research is focused on large-scale integration of genomic with other "-omics" data as well as biobank and electronic medical records data. Dr. Cox also has an active research program in data integration, particularly in the integration of functional genomic information to aid in discovery and interpretation of associations of genome variation with common disease. Her lab was the first to show that most of the common variant associations to common human diseases and complex human traits appear to be regulatory in function. Dr. Cox completed her Ph.D. at Yale University in 1982 and conducted postdoctoral research at Washington University and the University of Pennsylvania.

King Jordan, Ph.D. is a professor and Director of the Bioinformatics Graduate Program at Georgia Institute of Technology. He is broadly interested in the relationship between genome sequence variation and health outcomes. He studies this relationship through two main lines of investigation – human and microbial. In humans, he studies how genetic ancestry and population structure impact disease prevalence and drug response. Dr. Jordan's human genomics research is focused primarily on complex common disease and aims to characterize the genetic architecture of health disparities, in pursuit of their elimination. He received his B.A.

in Biology from the University of Colorado Boulder and his Ph.D. in Genetics from the University of Georgia.

Dianalee McKnight, Ph.D., FACMG is the Medical Affairs Director and Emerging Clinical Omics/Clinical Molecular Geneticist at Invitae. She is board-certified in clinical molecular genetics by the American Board of Medical Genetics and Genomics. She has more than 10 years of experience in the genetic diagnostics field. Prior to joining Invitae, Dr. McKnight was the director of the neurogenetics testing program at GeneDx, where she specialized in genetic testing for pediatric patients with epilepsy and intellectual disability. Dr. McKnight earned her doctorate degree at Penn State University and completed post-doctoral training at the NIH. While at the NIH, she was a research fellow at the National Institute of Dental and Craniofacial Research for 5 years, during which she also completed her board-accredited clinical molecular genetics training fellowship at the National Human Genome Research Institute.

Gil McVean, Ph.D. FMedSci, FRS is a Founder of and Chief Scientific Officer at Genomics plc. He has a background in statistical and population genetics and played leading roles in the International HapMap and 1000 Genomes Projects, as well as advisory roles in projects including the UK Biobank and Genomics England. He was the founding director of Oxford University's Big Data Institute and is a Fellow of both the Royal Society and the Academy of Medical Sciences.

Santiago J. Molina, Ph.D. is a Sociology/Science in Human Culture Postdoctoral Fellow at Northwestern University. Their work sits at the intersections of science and technology studies, political sociology, sociology of racial and ethnic relations, and bioethics. On a theoretical level, Santiago's work concerns the deeply entangled relationship between the production of knowledge and the production of social order. They are pursuing this line of research through two projects: The Biopolitics of Genome Editing and Categorical Heterogeneity in Population Genetics and Biomedicine. The second project is concerned with sampling practices and conventions of classification in human biology from the mid-Twentieth Century to the present. In collaboration with The Edmond J. Safra Center for Ethics at Harvard University, Santiago is researching how scientists in different disciplines conceptualize and operationalize human difference through "ancestry" and "population." Santiago's teaching has aimed to cultivate practical tools for thinking critically about the relationship between science and society. They received their Ph.D. in Sociology from the University of California, Berkeley and their B.A. from the University of Chicago.

Akinyemi Oni-Orisan, Pharm.D., Ph.D. is a pharmacist-scientist with an academic appointment as assistant professor at the University of California, San Francisco School of Pharmacy in the Department of Clinical Pharmacy and the Institute for Human Genetics. Dr.

Oni-Orisan has board certifications in applied pharmacology from the American Board of Clinical Pharmacology and in clinical lipidology from the Accreditation Council of Clinical Lipidology. He is a licensed clinician with practice experience in Cardiac Intensive Care Unit, cardiac stepdown, and outpatient (advanced dyslipidemia clinic) settings. His long-term research goal is to improve pharmacologic regimens for the prevention and treatment of cardiovascular disease using electronic health records linked to biorepositories. His current work involves the characterization of lipid-modifying agents for atherosclerotic cardiovascular disease through the utilization of electronic health records. Dr. Oni-Orisan teaches and mentors pharmacy students in the Discovery Projects research program including leading the inaugural Health Disparities Discovery Group. Dr. Oni-Orisan serves as Diversity Leader for the Department of Clinical Pharmacy to champion diversity, equity and inclusion (DEI) efforts in the department. He received his B.S. and Pharm.D. from University of Michigan and his Ph.D. from University of North Carolina at Chapel Hill.

Alice Popejoy, Ph.D. is an Assistant Professor in the Epidemiology Division of the Department of Public Health Sciences at the University of California, Davis (UC Davis Health). Dr. Popejoy's research program in public health genetics is situated at the intersections of evolutionary genomics, biomedical data science, statistical genetics, and the attending ethical, legal, and social implications (ELSI). She is currently focused on innovation and methods development to fundamentally shift the way human populations are categorized in biomedical research, epidemiology, and precision medicine. She received her B.A. from Hamilton College and her Ph.D. in public health genetics from the University of Washington.

Ramya Rajagopalan, Ph.D. conducts conceptual and applied research in the social and ethical implications of science, medicine, and technology, with a focus on the use of race and population descriptors in genomics. Her cross-disciplinary work has examined how concepts of race and ancestry circulate in the technology and practices of genomics and their intersections with social categories of identity, and has appeared in leading journals across the health sciences, sociology, and science and technology studies. Current projects investigate ethical issues in precision medicine and public health, the politics of genome editing, and algorithmic inequities in health care. Rajagopalan serves as Associate Director of Training, Evaluation, and Qualitative Research at the Center for Empathy and Technology, Sanford Institute for Empathy and Compassion, UC San Diego, where she leads the development of innovative curricula that bring a lens of justice, equity, inclusion, and compassion to health professional training in genomics and bioethics. She received her Ph.D. in Genome Sciences at MIT and was a postdoctoral fellow in sociology and bioethics at the University of Wisconsin-Madison.

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Mashaal Sohail, Ph.D. is an associate professor and computational genomics researcher in the Center for Genomic Sciences (CCG) at the National Autonomous University of Mexico. She has been developing a research program to study the genetic architecture and evolution of complex traits and disease in diverse humans using genomics, statistical approaches and deep learning. In the Sohail lab, they are interested in both method development using innovative approaches and statistics for reading genetic information, as well as in learning about the evolution and phenotypic variation of diverse humans. Her lab is also interested in the mechanisms of evolution, and in genome-wide patterns of selection and their relationship with gene expression patterns. A general goal is to use genomics data to learn about human evolutionary history, complex trait variation and their relationship. Dr. Sohail is working on methods for prediction of complex traits and disease in present-day and ancient humans for benefits of preventative and personalized medicine, and for understanding the evolution of complex disease. Her lab is also working on improving genomic resources for underserved groups to unravel their genetic history and complex trait architecture and are focusing these efforts on Mexico and Pakistan. She received her master's in History of Science and Ph.D. in Systems Biology from Harvard University.

Norbert Tavares, Ph.D. is a Science Program Manager at the Chan Zuckerberg Initiative, where he primarily manages single-cell biology research programs that support the international Human Cell Atlas consortium. Previously, he served at the National Cancer Institute, at the National Institutes of Health as an AAAS Science & Technology Policy Fellow, where he managed interdisciplinary trans-institute/agency research grant programs. Dr. Tavares is a microbiologist by training and was a Ruth Kirchstein Fellow at the University of Georgia, where he completed his Ph.D. investigating the bacterial biosynthesis of coenzyme B12 in the laboratory of Jorge Escalante-Semerena. Prior to his graduate work, he worked on large-scale protein drug production via bacterial fermentation in a biotech startup and in procurement for ESPN. Dr. Tavares has strong interests in advancing science by supporting basic research, open science, and the advancement of women and underrepresented individuals.

Phil Tsao, Ph.D. is Professor of Medicine (Cardiovascular Medicine), Stanford University School of Medicine and Associate Chief of Staff for R&D at the VA Palo Alto Health Care System. The primary interests in his laboratory include understanding the mechanisms regulating atherosclerosis and abdominal aortic aneurysm disease. While single genes can have dramatic effects in cellular biology, it is becoming increasingly clear that vascular disease (and health) is regulated by the coordinated expression of gene cassettes or pathways. By monitoring expression patterns of the entire genome simultaneously, his laboratory aims to identify networks of genes that work in concert to affect disease initiation and progression. This approach can often implicate specific nexus genes that are at the center of larger networks and/or participate in multiple pathways. Furthermore, his laboratory is investigating the role

microRNAs play in orchestrating the activity of multiple genes during the course of disease. He received his Ph.D. in Cardiovascular Physiology from Thomas Jefferson University.

Hannah Wand, MS, CGC (she/her) is director of the Preventive Genomics Program at Stanford Healthcare and a genetic counselor in preventive cardiology. Her research focuses on the socially responsible translation of population genomics through implementation science and community engagement. Her professional activities and teaching are in the areas of public health genetics, public education/engagement in genetics, increasing diversity in the workforce, and the use of REA in genetics.

Jennifer Webster, Ph.D. is a senior director of Precision Medicine Real-World Evidence Lead at Pfizer. Dr. Webster is a product lead and data scientist building curated real world evidence and population health applications. Dr. Webster received a B.A. in chemistry from Bryn Mawr College and Ph.D. in statistics from Texas A&M University.

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

Public Workshop on Use of Population Descriptors in Genomics Research

SPEAKER GUIDANCE: CONTEXT AND QUESTIONS

As a first step in the information gathering phase of their work, the Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research would like to better understand how and why individuals are identified in genomics research and how large consortia and biobanks use population descriptors. The goals for this workshop include examining the need for population descriptors in genetics and genomics research and learning how biobanks use population descriptors, manage legacy data, and integrate new data.

Session I: Examining Use of Population Descriptors in Genomics Research

Objectives

- To explore what types of population descriptors are needed for genetics and genomics studies.
- To examine how and why genetics studies should or should not incorporate social categories and environmental factors.

Key Questions for Speakers:

1. In your work, *who* is sampled? Why are they sampled? How are samples identified or categorized, and why?
2. How are population descriptors such as race, ethnicity, and ancestry used effectively in genetics and genomics research? How are they used ineffectively or inappropriately?
3. What types of genetics and genomics studies require the use of population descriptors such as race, ethnicity and ancestry?
4. Are population descriptors needed to capture social and environmental factors in genetics and genomics studies? If so, what descriptors are needed?

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

Public Workshop on Use of Population Descriptors in Genomics Research

SPEAKER GUIDANCE: CONTEXT AND QUESTIONS

As a first step in the information gathering phase of their work, the Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, would like to better understand how and why individuals are identified in genomics research and how large consortia and biobanks use population descriptors. The goals for this workshop include examining the need for population descriptors in genetics and genomics research and learning how biobanks use population descriptors, manage legacy data, and integrate new data.

Session II: Use of Population Descriptors by Biobanks and Other Research Consortia

Objectives

- To examine how and why biobanks and other large-scale programs use taxonomies currently, especially in areas of large diversity
- To learn how large-scale data collection projects are designed and managed
- To explore how legacy data might be managed and merged with future data

Key Questions for Speakers:

1. What taxonomies, classification schema, or population descriptors does your program use? How and why are these used?
2. Have your policies surrounding population descriptors changed since the project began? Why or why not?
3. How does your program approach legacy data? What challenges does your program face with managing or merging legacy data with current data? What special concerns might exist for merging legacy data with current and future datasets?
4. Beyond population descriptors, what attributes of individuals are important to collect for use in large datasets?
5. What are some lessons learned for the design of future large-scale data projects?

NATIONAL ACADEMIES Sciences Engineering Medicine

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCE AND EDUCATION
COMMITTEE ON POPULATION

Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

Public Workshop on Use of Population Descriptors in Genomics Research

SPEAKER GUIDANCE: CONTEXT AND QUESTIONS

As a first step in the information gathering phase of their work, the Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, would like to learn more about whom is studied in genomics research and how population descriptors are used within these studies. The goal for this workshop is to learn more about the historical context of population descriptors and how to standardize and use population descriptors in the future.

Session III: Community Input on Population Descriptors in Genomics Research

Objectives

- The committee requests public comments on the current use of population descriptors such as race, ethnicity, ancestry, etc. in genomics research and how the use of population descriptors could be improved upon in the future.

Key Questions for Speakers:

1. How do you identify yourself and how do you think that should be incorporated into genetics research studies?
2. How are population descriptors such as race, ethnicity, and ancestry being used or not used effectively in genomics research?
3. What population descriptors, if any, should not be used in genomics research?
4. Do all genetic studies need specific population and/or individual descriptors of their study subjects?
5. What aspects of the current use of population descriptors in genomics research need to be changed or improved?
6. How should population descriptors be used in genomics research moving forward?

BACKGROUND INFORMATION

Selected Readings in this Briefing Book

1. Mauro, M., D. S. Allen, B. Dauda, S. J. Molina, B. M. Neale, A. C. F. Lewis. (pre-print). A systematic review of guidelines for the use of race, ethnicity, and ancestry reveals widespread consensus but also points of ongoing disagreement. *arXiv* 2204.10672. https://arxiv.org/abs/2204.10672

2. Wang, T., et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature*. 604: 437-446. https://www.nature.com/articles/s41586-022-04601-8

3. Lewis, A. C., et al. 2022. Getting genetic ancestry right for science and society. *Science*: 376(6590):250-252. https://pubmed.ncbi.nlm.nih.gov/35420968/

4. Weale, M. E., et al. 2021. Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. *The American Journal of* Cardiology: 148:157-164. https://www.sciencedirect.com/science/article/pii/S0002914921002071?via%3Dihub

Links to Additional Resources

Session I: Examining Use of Population Descriptors in Genomics Research

- Edge, M.D., G. Coop. 2019. Reconstructing the History of Polygenic Scores Using Coalescent Trees. *Genetics* 211(1): 235-262. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6325695/

- Claussnitzer, M., J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis, M. E. Hurles, S. Kathiresan, E. E. Kenny, C. M. Lindgren, D. G. Macarthur, K. N. North, S. E. Plon, H. L. Rehm, N. Risch, C. N. Rotimi, J. Shendure, N. Soranzo, and M. I. McCarthy. 2020. A brief history of human disease genetics. *Nature* 577(7789):179-189. https://pubmed.ncbi.nlm.nih.gov/31915397/

- Oni-Orisan, A., Y. Mavura, Y. Banda, T. A. Thornton, and R. Sebro. 2021. Embracing genetic diversity to improve black health. *Mass Medical Soc.* https://pubmed.ncbi.nlm.nih.gov/33567186/

Session II: Use of Population Descriptors by Biobanks and Other Research Consortia

- Bycroft, C., et al. 2018. The uk biobank resource with deep phenotyping and genomic data. *Nature* 562(7726):203-209. https://pubmed.ncbi.nlm.nih.gov/30305743/

- Popejoy, A., et al. 2020. Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. *AJHG*. 107: 72-82. https://www.sciencedirect.com/science/article/pii/S000292972030152X

- Shedding Light on the Use of Population Descriptors in Clinical Genetics. Alice Popejoy. 2022. https://elsihub.org/video/shedding-light-use-population-descriptors-clinical-genetics

- How Race, Ethnicity, and Ancestry Information is Used in Genetic Diagnostics: One Lab's Perspective. Dee McKnight. 2022. https://elsihub.org/video/how-race-ethnicity-and-ancestry-information-used-genetic-diagnostics-one-labs-perspective

- Johnson, R. et al. (pre-print). Leveraging genomic diversity for discovery in an EHR-linked biobank: The UCLA ATLAS Community Health Initiative. *medRxiv* https://www.medrxiv.org/content/10.1101/2021.09.22.21263987v1.full.pdf

- Million Veterans Project publication list: https://www.research.va.gov/MVP/publications.pdf

A systematic review of guidelines for the use of race, ethnicity, and ancestry reveals widespread consensus but also points of ongoing disagreement

Madelyn Mauro[1], Danielle S. Allen[1], Bege Dauda[2,3], Santiago J. Molina[4], Benjamin M. Neale[5,6,7], Anna C. F. Lewis[1,8*]

1 Edmond J Safra Center for Ethics, Harvard University, Cambridge, MA, USA.
2 Center for Global Genomics and Health Equity, University of Pennsylvania, Philadelphia, PA, USA.
3 Institute of Clinical Bioethics, Saint Joseph's University, Philadelphia, PA, USA.
4 Department of Sociology, Northwestern University, Evanston, IL, USA
5 Broad Institute of Harvard and MIT, Cambridge, MA, USA.
6 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
7 Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.
8 Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA.
* Corresponding author

ABSTRACT

The use of population descriptors like race, ethnicity, and ancestry in science, medicine and public health has a long, complicated, and at times dark history, particularly for genetics, given the field's perceived importance for understanding between-group differences. The historical and potential harms that come with irresponsible use of these categories suggests a clear need for definitive guidance about when and how they can be used appropriately. However, while many prior authors have provided such guidance, no established consensus exists, and the extant literature has not been examined for implied consensus and sources of disagreement. Here we present the results of a systematic review of published normative recommendations regarding the use of population categories, particularly in genetics research. Following PRISMA guidelines, we extracted recommendations from n=121 articles matching inclusion criteria. Articles were published consistently throughout the time period examined and in a broad range of journals, demonstrating an ongoing and interdisciplinary perceived need for guidance. Examined recommendations fall under one of eight themes identified during analysis. Seven are characterized by broad agreement across articles; one, *Appropriate definitions of population categories and contexts for use*, revealed

substantial fundamental disagreement among articles. While many articles focus on the inappropriate use of race, none fundamentally problematize ancestry. This work can be a resource to researchers looking for normative guidance on the use of population descriptors, and can orient authors of future guidelines to this complex field, contributing to the development of more effective future guidelines for genetics research.

INTRODUCTION

Evidence of race-based health disparities has mounted in recent years, especially during the COVID-19 pandemic. [1–5] As researchers have stressed the role of structural racism in generating these disparities [6] and the importance of racially-stratified data to developing a better understanding of them[7], prominent institutions like the National Academies of Sciences, Engineering, and Medicine (NASEM) and the National Human Genome Research Institute have begun to wrestle with questions of the usefulness of population descriptors. [8,9] A recent series of commentaries in the New England Journal of Medicine illustrates the significant disagreement on the value of race as a variable in biomedicine, as well as a turn towards concepts from genetics as a potentially suitable alternative. Vyas et al. argued that the insertion of race into clinical tools relies on faulty assumptions about the genetic contribution to racial categories and can lead to interpretations of racial disparities as unavoidably genetically determined when they may be avoidably socially determined.[10] Borrell at al. raised the counterclaim that leaving awareness of race and ethnicity out of healthcare can exacerbate racial and ethnic disparities by failing to monitor or condemn them, and argued that race and ethnicity should therefore be used alongside genetic ancestry to understand health outcomes. [11] Oni-Orisan et al. highlighted a potential contribution of genetic ancestry to COVID-19 disease outcome, and after broadly reviewing the contribution of genetic difference to health disparities and the overlap between genetic and racial diversity, proposed that "the ultimate goal… would be to replace race with genetic ancestry in an

evidence-based manner." [12] But this turn to genetic ancestry as a more objective way to capture biological difference between groups has its own pitfalls.[13,14] Complex genetic ancestry information is most often smoothed into continental ancestry categories: a dangerous oversimplification due to the striking resemblance of continental ancestry categories to racial categories.

Given the complicated nature of this topic and this debate, a researcher considering employing population categories in their work might search for guidelines to consult in order to determine which population categories to use for which purposes. They would find a plethora of articles offering such guidelines, but with a wide range of focuses and varying levels of specificity; this phenomenon suggests a pervasive need for guidance in the scientific and medical community, but also highlights a continued lack of explicit, centralized normative guidance..

Prior works that have assessed this complicated body of normative literature have either examined only small sets of recommendations or have analyzed the impact of a single recommendation or set of recommendations on the practices of authors and/or clinicians. [15–17] A systematic review of all existing normative guidelines is lacking. In its absence, it is impossible to identify the recommendations that have been echoed by multiple authors, representing areas of normative consensus, or to pinpoint areas of consistent disagreement. The former is useful because commonly-provided recommendations point to noncontroversial areas of improvement for researchers, clinicians, and public health practitioners, and can also influence future guidelines by highlighting the topics and sentiments that have already been well-expressed in the field. The latter point to topics that merit further examination and likely hold some of the most pressing ethical concerns that underlie the employment of population categories in genetics and across scientific fields.

Here we present the results of a systematic review of normative guidelines relating to the use of population categories in science, medicine, and public health, thus providing an overdue clarification and categorization of normative works in this field. It focuses particularly on the

relevance of genetics to the use of these categories, given the perceived importance of genetics for understanding between-group differences. [10–12] This systematic review can provide a resource for consultation by researchers, funders, practitioners, regulatory bodies, and others who may find themselves lost in this otherwise disordered and overwhelming space. It also gives the authors of future guidelines — such as the recently convened National Academic committee on this topic — a much-needed orientation to existing work. This piece thus contributes to informing and improving future work and to pushing the field towards clearer and more effective future sets of guidelines.

METHODS

Search Method

This is a systematic review of literature containing normative recommendations for the use of race, ethnicity, and ancestry in science, medicine and public health, with particular focus on genetics. We conducted the review in accordance with guidelines established by Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[18] In December of 2021, two different search methods were applied in parallel to identify and collect articles of interest. The PubMED electronic database was searched using the following string: "population groups AND genetics AND human research AND bioethics[sb]." The [sb] qualifier limits the set of articles returned by the search to only include those labeled by PubMed as pertaining to bioethics. "Population groups" was used rather than "race AND ethnicity AND ancestry" because the phrase is a MeSH term—a database-specific phrase that more thoroughly and specifically refines searched articles. At the same time, a search was conducted using the Google search engine with the string "race, ethnicity, and ancestry in genetics". The first 100 articles of the results of this search were included. The Google search was designed and included to ensure that articles accessed by authors conducting a preliminary search for normative recommendations to guide their work would be included. Given the transformative impact of the Human Genome Project, which was completed in

the first few years of the 21st century, both searches were refined to only include articles published since January 1st, 2000.

This combination of search terms yielded ten of thirteen articles previously identified by one reviewer as relevant to the topic of this systematic review. Although other tested combinations yielded more of these articles, they raised the total number of returned articles to the tens or hundreds of thousands—thus, in an effort to balance the aim of including these articles with the aim of maintaining precise focus within this review, the above-mentioned search term combination was chosen.

Preliminary screening criteria

In order to refine the results of this search to yield articles with a normative focus, the titles and full abstracts of all returned articles were read. Each article was screened for how likely it seemed to contain normative recommendations on the use of race, ethnicity, and ancestry in science, medicine, or public health. Gray literature, such as transcribed presentations and speeches, as well as pieces published in popular media were not excluded. Standards for inclusion were established by comparing the labeling of the first 100 PubMED articles between two reviewers .The rest of the articles were labeled by one reviewer. Figure 1 illustrates the full process of inclusion and exclusion.

Extraction

Normative recommendations from these articles were extracted using a three stage process. First, summarized normative recommendations were extracted from the abstract and conclusion. Second, if articles contained tables or sections explicitly allocated to reporting recommendations, the contents of those tables and sections were extracted. Finally, each article was searched for the following words and phrases to identify further normative recommendations for extraction: *recommendations*, *guidelines*, *should*, *must*, *need*, *ought*, and *consider*. This strategy was determined

to be sufficient by comparing, for 10 articles, recommendations identified using this strategy to recommendations extracted by another reviewer based on an examination of the whole text. Along with text of each recommendation, any accompanying text pertaining to the justification of the recommendation was extracted.

Additionally, the following basic data about each article were collected: authors, all countries of author affiliation and country of majority author affiliation, journal, country of journal publication, publication year, and number of citations as recorded by PubMed. For the purpose of later categorization, journals were sorted into six categories—science, medicine, public health, law, ethics, or other. Finally, articles were tagged if their explicit aim was to provide normative recommendations, and if they pertained to ancestry.

After all the recommendations were extracted, thematic content analysis was conducted by examining each recommendation and identifying emerging themes. In order to establish the list of themes, two reviewers examined the recommendations from several articles. Theme assignments were all made by one reviewer. Initial themes were grouped into broader themes. Another reviewer assessed theme assignments and any disagreements were discussed and resolved.

RESULTS

Study Selection and Characteristics

As shown in Figure 1, this search yielded 1,073 articles in total after duplicates were removed, 1,035 of which emerged from the PubMED search and 38 of which emerged uniquely from the Google search. 218 articles remained after preliminary screening, and 121 were included in the final analysis. 384 normative recommendations were extracted from these 121 articles.

There is a clear dominance of articles published in the first decade of the 21st century, surrounding the completion and publication of the Human Genome Project, as shown in Figure 2a—however, Figure 2a also shows that there have been sustained publications on this topic

throughout the examined time period. The articles examined are broadly distributed across journals with different focuses, although most articles hail from scientific or medical journals, see Figure 2b.

Most if not all contributing authors for 102 articles (84% of total article yield) were affiliated with the USA. Four articles hailed from Canadian authors, three from authors in Germany, two from authors in the U.K., two from Italian authors, and one each from authors in Austria, Iceland, Japan, Scotland, Singapore, and South Africa. The country of majority authorship remains undefined for two of the included articles, one of which is an editorial with no explicit author, and one of which was written by four authors each from different countries.

Although many recommendations called for the formation of working groups to publish research or normative commentary on the use of race, ethnicity, and ancestry in the sciences, only three of the included articles themselves were written on behalf of a group. The articles have a median of 27.5 citations, with an interquartile range of 10.75 to 65.25, and with 23 of the articles having over 100 citations and two having over 500. Finally, 34 of the 121 examined articles contain a total of 60 recommendations that mention ancestry.

Thematic Discussion

Twenty-eight distinct themes (referred to as sub-themes from here on) were identified and grouped further into eight broader themes. Table 1 lists these themes and their associated sub-themes, as well as the articles in which recommendations pertaining to these themes and sub-themes were found. Recommendations that did not fall under any of these themes were marked "Other". Not all sub-themes will be fully addressed in the discussion of each theme: refer to Table 1 and to the data listed in Supplementary Materials for the complete list of recommendations and the sub-themes assigned to them.

1. The need for transparency

Forty-seven (12.2%) recommendations extracted from the analyzed articles propose a need for transparency from researchers conducting studies that employ population categories. Almost half of those urge investigators and authors to be transparent about *how* any population categories used in their studies are defined. This includes how participants were categorized, and if they were categorized by the investigators or by the participants themselves. [19–27] Many recommendations also encourage authors to provide their own definitions of the broad population categories they use (e.g., of the term "race"), often citing the confusion that results from leaving population categories undefined and the fact that many population categories are assumed to be potentially explanatory variables without justification.[15,24,28–31] Five of the recommendations in this sub-category pertain specifically to ancestry. The conflation of population categories with biological similarity is addressed by two of them, which urge researchers to acknowledge when they are using categories as a proxy for ancestry and to note explicitly how much biology is assumed to contribute to the categories they employ. [24,32] The other three all belong to one article, which urges investigators to clearly differentiate race, ethnicity, and ancestry when using more than one of these in a study and to fully describe the methods by which genetic ancestry was determined or inferred. It also calls for investigators not just to use the term ancestry, but to specify "genetic ancestry" or "inferred genetic ancestry" based on the methods used to assign ancestry labels. [15]

The second most-popular sub-category of recommendations within this theme is made up of recommendations that urge investigators and authors to be transparent about *why* they have employed population categories in their research. These mainly encourage investigators to make explicit the relevance of any employed population categories to their research question(s), especially when the category employed is race. [22–24,26,32,33] Five of the recommendations within this subcategory more specifically urge authors to justify the use of non-genetic, non-biological population categories either as additional variables in genetic studies or in place of genetically-inferred variables. [22,24,30,32]

The rest of the recommendations that fall under this theme relate to transparency about data, collection methods, and the link between results and conclusions. Six recommendations call for investigators to make their collection methods and all of their data publicly accessible, so that other researchers can further investigate the validity of any claims made on the basis of this data and so that investigated communities can access study results.[33–38] Six more urge investigators to fully analyze and explain their findings: all of these recommendations share the goal of discouraging unsubstantiated inferences from data and treating every finding carefully, either by proposing follow-up studies to investigate the finding further, exploring the potential contribution of non-population factors, or simply making sure not to report observed associations without any comment on potential nuance.[16,23,39]

2. Awareness of impact on particular communities

Seventy-five (29.4%) of all extracted recommendations fell under this theme. Most of these encourage awareness within the scientific and medical communities of the past and present impacts of racism in science and medicine. Besides compelling researchers to be generally wary of group harms, misconceptions, and mistrust that can be seeded by irresponsible race-based research [40], many of these recommendations advise researchers not to repeat the grave mistakes made and harms done throughout the racist history of science and medicine. [25,28,32,41,42,42–46] All of these articles encourage developing awareness of these historical lessons through the education of investigators and authors. However, while some identify past mistakes and the lessons that should be learned from them—these include Nazism and the Tuskegee Syphilis Study to caution against "eugenic temptation" [42] and the use of racial labels to insinuate inferiority or superiority [32]—many simply broadly advise that research today should be historically informed. [28,43,45] A separate class of recommendations within this sub-theme encourages education of clinicians and researchers about the ways in which racism still extends its hand into science and medicine today. [25,32,41,44,46] Finally,

several recommendations urge clinicians to acknowledge the impact of racism on their practices, so that they can root it out and develop more sophisticated understandings of race. [23,47–50]

Thirty-three recommendations within this theme encourage investigators to respect the study populations examined in their work. Most of these address the concern that population-specific studies are often conducted without proof that a population-specific focus has tangible scientific or medical benefit, and so urge either review boards, the FDA, or society as a whole to ensure that this is true for all research involving only particular populations. [11,51–53] Others advocate for improving the sensitivity of the informed consent process for minority populations by implementing measures to ensure all participants have a holistic understanding of the research question and scope, and of all of the potential uses of any collected material.[51,54–56] Thirteen more encourage researchers to foster a respectful partnership with studied communities by releasing their eventual conclusions to the studied communities [23,38,52,54,56,57] and maintaining respectful communication with them throughout and beyond the study [37,53,56,58–60]. Relatedly, twelve recommendations within this theme urge investigators to consult their studied populations throughout the research design and implementation process in order to establish a respectful partnership, develop trust, ensure those populations are appropriately named in the study, and identify risks that may not be obvious to the researchers themselves or to review boards. [41,53,55,56,58,61–63]

3. The use of appropriate statistical methodologies

Forty (10.4%) of the extracted recommendations fell under this theme, and many of these urge investigators to be aware of and account for the contribution of factors other than race or ethnicity to health outcomes when designing and conducting research. These factors include racism [64], socioeconomic status [21,26,51,64–67], environmental exposures [51,64,68], education [21,26,51,64], place of residence [26,64], ancestry [51,65,67,69] and cultural practices [65,66,68,69]. One of these articles argues that an

examination of the effects of ancestry (as opposed to race or ethnicity) on health should also include a consideration of other factors [68]. Recommendations from one article emphasized that the hypothesis being tested should dictate which of race, ethnicity, or ancestry should be used [23].

Another subset of recommendations within this theme discourages investigators from making causal claims based on associations found in their data. These recommendations share the central message that correlation does not and should not imply causation, and that researchers should interpret their data carefully before using them to make and support claims. [11,24,25,36,43,66] Three of these have a more specific focus: one denounces the use of genetic data to enforce between-group differences in any context [25], and two others caution against overemphasizing the impact of ancestry on health [66,70].

Finally, some recommendations pertaining to statistical methodologies set out specific guidelines for the statistical interpretation of genetic data. These include establishing sufficient statistical significance in any investigated associations between racial, ethnic, genetic, or other categorization and phenotype [24], conducting pooled rather than stratified analysis to avoid "arbitrary clustering decisions" [32], how to select the number of principal components needed in principal components analysis (PCA) to account for population stratification [32], and incorporating socio-historical considerations into understanding population-specific genetic variation [71].

4. Public reactions and public engagement

Most of the thirty (8.9%) recommendations within this theme encourage anyone investigating links between population categories and health to be aware of all possible interpretations of their conclusions, and to present findings in a way that minimizes dangerous misinterpretations. This category shares a concern about racism with the recommendations in the category *Awareness of impact on particular communities*, but this category places emphasis on public misconceptions rather than specific harms done to communities. Those aimed at the

practices of investigators urge them to note the limitations of their methods [62], avoid overstatement or generalization in their presentation of results [72,73], make accurate versions of their conclusions accessible to popular media to avoid sensationalization [51,74,75], and correct any observed misinterpretations of their results [36]. Five recommendations specifically state that researchers and reviewers should anticipate any potential socioethical problems raised by the associations such research makes or could make, such as the reinforcing or creation of social stigma, and address them within the research or use them as a basis for rejection of funding or publication. [15,36,51,76,77]

The recommendations aimed at healthcare advertising and direct-to-consumer ancestry testing companies also encourage those formulating public messaging to be aware of and attempt to prevent potential misinterpretations of their claims [78–80]. They recommend that direct-to-consumer ancestry companies demonstrate how consumers should interpret their results and explain to consumers the difference between ancestry and race, ethnicity, and other group membership. [80] Messaging from healthcare advertising, meanwhile, should be careful not to suggest links between race, ethnicity, ancestry, and health that could legitimize racism. [79] One recommendation explicitly states that this messaging should avoid insinuating any biological basis for racial categories.[78]

The rest of the recommendations within this theme encourage investigators and institutions to increase the involvement of the public in the refinement of future population-based research. The two main ways in which they recommend this be accomplished are through consultation of the public throughout the research process [23,40,81] and through education efforts that raise awareness of racism and its impacts and of the pitfalls of current research involving population categories and health. [36,82–84]

5. The need for diverse samples and practitioners

This theme consists of nineteen (5%) extracted recommendations that bring attention to the need for diversity both in research studies and in the medical and scientific community as a

whole. Recommendations that encourage more diverse recruitment of study participants mostly advocate for increased recruitment of minority populations, either to better represent global diversity [21,67,85–87] or, in the context of ancestry, to improve the applicability of clinical genomics as a whole to those of non-European genetic ancestry [68,88,89]. Three recommendations lie tangential to this sub-theme, as they share the motivation of refining the true applicability of population-based research: they urge researchers to note explicitly how representative their study participant cohorts are of the larger population(s) being studied. [22,24,86]

The last seven recommendations within this theme call for the makeup of and discussions within the scientific and medical community to be diversified: according to these recommendations, doing so further incorporates those with an interest in improving population-focused research into medicine and science and develops a wider understanding of and appreciation for diversity. [38,50,59,74,86,90]

6. The need for an appreciation of nuance

Twenty-seven (7%) extracted recommendations fell under this theme, which consists of recommendations that encourage researchers and/or clinicians to consider and use race, ethnicity, ancestry, and/or population in a more nuanced fashion.

Twelve of these specifically pertain to race and ethnicity. Five encourage researchers and clinicians to acknowledge that race and ethnicity are not strict categories and are influenced by multiple circumstances. [23,28,48,81,91,92] Two others offer specific variables that may provide useful replacements for the category of race specifically—ethnicity [21] and geographic origin [93]. The last three simply urge scientists and clinicians to acknowledge the bounds of the relevance of racial and ethnic categories. [27,94] Four of these touch on conceptions of ancestry and population; they broadly propose that population categorization should not be immediately assumed to be legitimate, and that its justification, the types of categories that currently exist, and the ways in which people may

be assigned to categories should consistently be questioned. [43,93,95]

Finally, nine of these encourage those in the scientific and medical community to consider the relationship between race, ethnicity, ancestry, and population with more nuance. Seven recommendations in this category encourage that any associations between biology and existing categories be made only upon findings of substantial correlation, clear justification, and unbiased presentation. [24,52,90,96–98] Another calls for a reexamination of the historical contributions to genetic variation, so that we may understand why it is sometimes customary to assign biological significance to arbitrary categories. [95] The final recommendation within this theme calls for researchers to use a specific algorithm developed by the authors that attempts to refine self-identified race and ethnicity. [99]

7. Appropriate definitions of population categories and contexts for use

A subset of the eighty-three (24.9%) recommendations within this theme set out how certain population categories should or should not be defined. Unlike the other identified themes, this theme comprises many mutually incompatible recommendations. Many of these encourage researchers to define race not as a biological phenotype, but as a complicated social phenomenon, or to categorize populations by socio-environmental variables instead of by race. [23,48,78,81,100,101] However, one of these introduces the social concept of race not as a replacement for a biological concept of race, but as a definition of race that should be considered in concert with a biological definition of race [100]. Others are concerned with how strictly population categories should be defined: while one states that populations should only be defined by genetic variation [93], another considers rare genetic variants to be too small and strict of a basis for population categorization [43]. Another suggests that populations should be defined in multiple ways [95], while still another maintains that population categories should not and cannot be broadly defined, and urges researchers to instead circumstantially define them [28]. Finally, some provide their own definitions of

race, ethnicity, and/or ancestry rather than commenting on how they should be defined [52,67]. For example, in 2015 Mersha and Abebe proposed, *"To better understand human genetic variation in the context of health disparities, we suggest using "ancestry" (or biogeographical ancestry) to describe actual genetic variation, "race" to describe health disparity in societies characterized by racial categories, and "ethnicity" to describe traditions, lifestyle, diet, and values."*

A second subset unconditionally approves of the scientific and medical communities investigating links between population categories and health. Some claim that self-identified racial categories overlap with genetic groupings and so provide a more feasible way to establish ancestry than empirical genotyping [102]; others justify this position by claiming that race- and ethnicity-based disparities still exist, and that it is fruitless to try to eliminate them and understand their causes without examining them [23,28,96,103–105]. Three more recommendations focus on the value of using race or ethnicity to examine the effects of racism on health [95,104,106]. One powerful recommendation among these argues that the potential to uncover the health effects of racism is unsacrificable. *"To truly get to the bottom of racism and its negative impact on persons of African ancestry,"* wrote Dawson in 2003, *"we cannot solve the problem by pretending that we are not what we are, or who we are, because some, be they friend or foe, are ashamed of their actions, if possible, or that of others. Because in effect, to argue that Blacks should somehow desist from categorizing ourselves as we see fit; in our intellectual exchanges is in and of itself, in the final analysis, a pernicious form of racism and self-hatred if that is the source of calls for us to 'change.'"* [104]

There are some other recommendations that justify the use of population categories in research, but only when researchers take additional measures to minimize confusion, unsupported conclusions, and arbitrary employment of population categories [23,24,62,78,107,108]. These measures include collecting data on socioeconomic status and genetics to supplement analysis, precisely defining employed population categories, and establishing a specific health or scientific interest

furthered by the use of categorization, but some define these measures with sweeping statements like "tak[ing] great care." [107] Others justify the use of population categories only for research with certain purposes, such as to survey incidence of disease, to examine the interface between social and biological variation, or to lay the ground for further research into more specific risk factors. [39,52,95,106,109] The last four recommendations then urge investigators and authors to be sensitive with the terminology used to describe populations in such research, in order to avoid confusion and the insinuation of a hierarchy between groups. [21,32,44]

However, some recommendations completely discourage the employment of certain population categories in research. Overall, these pertain entirely to race and ethnicity, and none broadly call for ancestry not to be employed in research. Two recommendations within this sub-theme call for the refusal of approval and/or funding to genetic studies using race as a variable. [101,108] Other recommendations urge clinicians and researchers to minimize or eliminate their use of race and ethnicity [21,25,36,68,110], to focus their efforts on genetic variation [111], or to focus on ethnic rather than racial categorization [29]. Finally, ten more recommendations within this theme discourage researchers from using race as a proxy for other variables, most often ancestry, in their research. [27,32,39,70,112–115] Six more recommendations urge investigators to collect ancestry data instead of race or ethnicity, in order to better understand the variation captured by their study participant cohorts or to investigate a genetic basis for any observed disparities. [25,98,101,102,108] Only one article fundamentally problematizes the use of ancestry in medicine, but only in the context of developing individualized treatments: the authors state that race and ancestry are both *"imperfect measures that will both slow researchers' progress toward individualized genetic knowledge and provide clinicians with incomplete, potentially misleading information."* [116]

8. The need for further research and guidelines

Fifty-seven (17%) extracted recommendations fall under this theme. Most of these

encourage further research into several different aspects of and important considerations regarding the use of race, ethnicity, and ancestry in health and sciences research. This includes research to elucidate where population categories lie along the social-biological axis [28,35,41,55,81,117,118] and further investigation of existing disparities between populations to identify their causes and to address them [84,92,119]. However, this subcategory is also marked by a high incidence of ancestry-related recommendations, which state the need to optimize the utility and robustness of genetic ancestry estimation [11,43,120,121] and to learn how to best characterize human genetic variation, whether that be by continental ancestry categories or not. [118,122]

The rest of the recommendations within this theme call for the development of further guidelines and regulations for future research involving race, ethnicity, and/or ancestry. Many of these are in vague terms, calling for the formation of policies or new definitions that "mitigate risks" or "address challenges". [15,19,35,74,77,81,109,117,123] However, many recommendations call for journals or centers to adopt the specific standards or guidelines suggested by other articles in this collection, such as those that suggest a need for transparency. [15,33,46,56,89,122,124] One article contains the ancestry-specific recommendations that pertain to this category; it urges leaders in the human genetics field to convene to develop standards of practice and communication of ancestry-testing and research, including standards of representing statistical confidence in ancestry estimation and guidelines for appropriate terminology. [120]

DISCUSSION

This systematic review analyzed a total of 121 articles published since 2000 that provide normative recommendations on the use of race, ethnicity, and ancestry in science, medicine, and public health, with particular focus on genomic research. Recommendations were consistently published throughout the 21-year period analyzed (see Figure 2a), demonstrating an ongoing perception of the need for such normative criteria. Articles appeared in journals from a wide variety

of domains (see Figure 2b), demonstrating that this perceived need extends across several disciplines.

Most if not all contributing authors for 84% of these articles were affiliated with the USA. The dominance of articles from the United States may be a result of its particularly dark history of systemic racism motivating exceptional concern for rooting out notions of prejudice and hierarchy from customs of population categorization. Further, a cross-national survey of the 2000 global census, published in 2008, revealed that of all 147 national census questionnaires examined, only that of the United States measured race and ethnicity separately. [125] As the report states: *"In this view, which is extremely unusual in international perspective, ethnic groups are different from races because they are rooted in sociohistorical contexts; races thus appear to be grounded in something other than social processes."* This practice may contribute to general confusion within the United States about the significance of these categories, and by extension the context for their usage in medicine and science.

Seven out of eight themes identified during analysis were marked by significant consensus amongst the recommendations within them, indicating that there are certain uncontroversial sources of concern but that these have remained unresolved throughout the analyzed time period. Even agreeing recommendations, however, broadly give a sense of confusion and frustration. Different recommendations pertaining to the same theme often drew attention to the same issue but proposed different approaches. For example, some articles that drew attention to the need for transparency urged researchers to be transparent about different aspects of their research—while some focused on the need for clear definitions of terms like "race" and ethnicity", others encouraged transparency about research methods, like how and by whom study participants are categorized by race or ethnicity. Furthermore, the entire set of recommendations pertaining to transparency demonstrated pervasive frustration with vague circulating definitions and practices that can lead to subpar or nonexistent justifications for the employment of population categories in research.

Finally, the third most popular theme calls for further research to be conducted and further guidelines to be established about the use of race, ethnicity, and ancestry in science, medicine, and public health. The prevalence of recommendations within this theme illustrates a clear desire to pave the way for future normative criteria that are more informed and organized than the existing set of recommendations.

In contrast to the seven themes with broad agreement, the theme with the most recommendations, *Appropriate definitions of population categories and contexts for use*, contained recommendations that significantly disagreed with one another. Many recommendations defined "race" at wildly different places along the social-biological axis, and many put forth differing opinions on how strictly population categories in general should be defined, ranging from proposals strictly delineating the definitions of race and ethnicity [67], to the proposal that they be jointly be defined as *"any classification, whether genetic or self-reported, cognizable and supported by evidence in medical, legal, or social disciplines."* [52]

The recommendations specifically pertaining to the *use* of population categories then argued one of three disparate positions: that it is always acceptable to use certain population categories in research, that it is only acceptable to use them under certain conditions, and that it is never acceptable to use them. The considerable disagreement amongst these recommendations indicates that a lack of clear, centralized definitions of population categories goes hand in hand with confusion and disagreement about when they should be used. The large volume of recommendations within this theme then arguably indicates that answering questions about the definition and appropriate use of population categories will be foundational to approaching the problems identified throughout all themes.

Thirty-four of the 121 examined articles and 60 of the 334 extracted recommendations pertain to ancestry. Most of the recommendations about ancestry relate to establishing its relationship with the concepts of race and ethnicity. Many argue for an active differentiation of race

and ethnicity from ancestry, especially by direct-to-consumer ancestry companies, in order to avoid misinterpretation of the meaning and significance of ancestral differences by the public. However, some disagree about the nature of the relationship between ancestry data and racial data. For example, while certain recommendations encourage researchers not to use racial data as a proxy for ancestral data or even support using ancestral data in place of racial data, others argue that there is significant overlap between racial and ancestral categories, so racial data should be collected when ancestral data is too costly to obtain. This disagreement, along with the many recommendations in *The need for further research and guidelines* theme that encourage further research into how to accurately estimate and describe genetic ancestry, illustrate that a clear role for and understanding of the value of ancestry in science and medicine is yet to be established. This imperfect understanding of ancestry, its significance, and how it should be used is also reflected in the many recommendations that call for investigators to adjust their research methods in order to avoid overestimating the impact of genetic ancestry on their results. Several of these urge researchers and care providers to consistently include an acknowledgement of other variables, such as socioeconomic status, when investigating or considering ancestry.

However, few of the extracted recommendations focus on the potential dangers of using ancestry categories. Only one fundamentally problematizes the use of ancestry, and only in the context of precision medicine.[116] Recommendations in *The use of appropriate statistical methodologies* theme that are critical of the use of ancestry propose amendments to the statistical methodologies currently used to infer ancestry, and these only problematize the way ancestry is estimated, not the contexts in which it is used. Recommendations in *The need for diverse samples and practitioners* theme only criticize the overrepresentation of European samples in genomic research in particular, and urge researchers to improve the applicability of such research by collecting samples from a broader range of individuals. Furthermore, one of the most-cited articles in this collection, “The meanings of "race" in the new genomics: implications for health disparities

research" by Lee et al. in 2001, calls ancestry a *"neutral word"* that can be used to *"avoid potentially misleading terms."*[78] Overall, the extracted body of recommendations does not demonstrate the same examination of the acceptability of the use of ancestry that it does for race and ethnicity. This begs the question: is ancestry being touted as a cure-all to the problems raised by the use of race and ethnicity in science and medicine? The objective value and potential dangers of the use of ancestry thus provide potential areas of examination for future normative research that will rely heavily on the engagement of the field of genetics in questions about the value of population descriptors.

CONCLUSION

We have conducted the first systematic review of normative guidelines pertaining to the use of population categories in science, medicine, and public health, with particular focus on genomic research. Examined recommendations were published consistently throughout the examined time period and in a wide range of journals, illustrating a persistent and interdisciplinary recognition of the need for normative guidance on this topic.

We identified seven themes with broad agreement across recommendations: *The need for transparency; Awareness of impact on particular communities; The use of appropriate statistical methodologies; The role of public reactions and public engagement; The need for diverse samples and practitioners; The need for an appreciation of nuance; The need for further research and guidelines*. Despite broadly agreeing, however, recommendations pertaining to these themes reflected a sense of confusion, uncertainty, and persistent dissatisfaction. Many used nonspecific language to identify pertinent issues without proposing concrete solutions; others proposed different solutions to the same issues. Recommendations under the theme of *The need for transparency* conveyed particular frustration at the current lack of explicit definitions for many of the population categories in circulation. We identified significant fundamental disagreement within one theme: *Appropriate definitions of population categories and contexts for use.* Finally, we identified that although many

recommendations focus on the inappropriate use of race and ethnicity, and some even condemn their usage entirely, none fundamentally problematize the use of ancestry, and criticism of ancestry is limited to calls to refine the statistical strength of ancestral estimations and to improve the applicability of genetic research.

This review demonstrates a strong need for new, clear, and effective guidelines on the use of population categories in science, medicine, and public health, and a particular need to investigate the appropriate definition and use of ancestry within these disciplines. The applicability of these guidelines should extend across disciplinary and national borders, given the global and interdisciplinary concern about this topic represented in this examined pool of recommendations, and should particularly engage the genetics field as they develop new standards for describing between-group differences.[8]

SUPPLEMENTARY INFORMATION

One excel file containing information on all the articles meeting inclusion criteria, all extracted recommendations, and theme assignments to these recommendations. (Please email annalewis@fas.harvard.edu for a copy).

ACKNOWLEDGEMENTS

We thank the members of the E J Safra working group on the concepts of ancestry, genetic ancestry, and population for their input on this work.

DECLARATIONS OF INTEREST

B.M.N. is a member of the scientific advisory board at Deep Genomics and RBNC Therapeutics, a member of the scientific advisory committee at Milken, and a consultant for Camp4 Therapeutics and Merck. A.C.F.L. owns stock in Fabric Genomics.

FIGURES

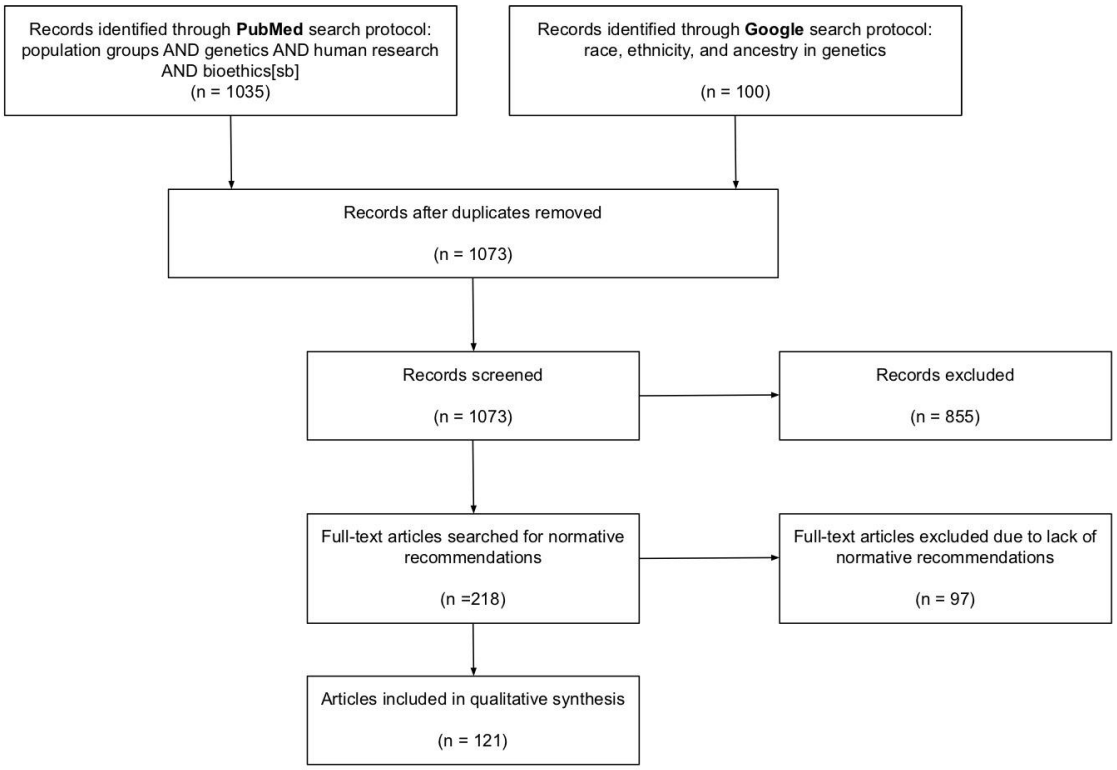


Figure 1: PRISMA Diagram, demonstrating the process of article inclusion and exclusion. 121 articles returned by the complete search strategy yielded 334 normative recommendations on the use of race, ethnicity, and ancestry in science, medicine, and public health.

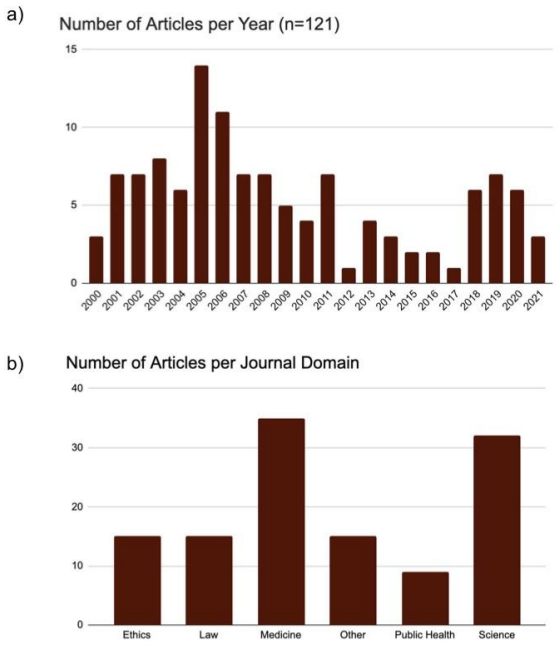


Figure 2: Articles continuing normative recommendations on the use of population categories have been published consistently since the year 2000, and across a broad variety of domains. a) Although there is a clear clustering of articles around the first decade of the 21st century, articles pertaining to this topic have been published consistently throughout the 21 examined years. b) While most articles hail from journals with a majorly scientific or medical focus, the other listed domains are also well-represented, suggesting that this topic is truly an interdisciplinary one. Journals categorized within the "Other" domain have main focuses ranging from history to philosophy (see Supplementary Materials).

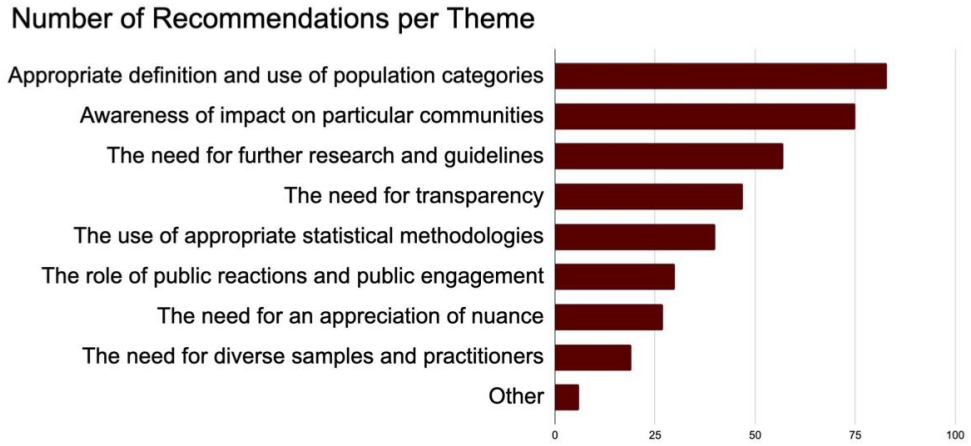


Figure 3: Extracted recommendations were distributed across eight themes. Before these eight themes were identified, each recommendation was examined and tagged based on its content—these twenty-nine tagged "sub-themes" were then sorted into eight larger themes. Those which did not correspond to a sub-theme were labeled "Other". Table 1 reports the sub-themes that correspond to each theme.

TABLES

Table 1: Common themes and sub-themes identified among included articles

Theme	**Sub-themes**
The need for transparency	a) Investigators should be transparent about why they are using population categories. b) Investigators should be transparent about how any population categories used are defined. c) Investigators should make their data and collection methods publicly accessible. d) Investigators should fully analyze and explain their findings.
Awareness of	a) Investigators should respect study populations.

impact on particular communities	b) Awareness of past and present impacts of racism in science and medicine should be encouraged/increased. c) Investigators should consult specific communities during the research process.
The use of appropriate statistical methodologies	a) Investigators should not make causal claims based on associations found in their data. b) Relates to the statistical interpretation of genetic data. c) Investigators should be aware of the contribution of factors other than race/ethnicity, such as socioeconomic circumstances or ancestry, to health outcomes. d) Investigators should match collected variables to their study question(s).
The role of public reactions and public engagement	a) Investigators should be aware of the potential social impact of research involving population categories, and sensitively release results of or information about this research to the public. b) Investigators and institutions should consult and educate the public. c) The scientific community should increase visibility of and education about important considerations for research involving race, ethnicity, and/or ancestry.
The need for diverse samples and practitioners	a) Investigators should recruit more diverse cohorts of study participants. b) The makeup of and discussions within the medical community should be diversified. c) Researchers should ensure that study populations are representative of the investigated population.
The need for an appreciation of nuance	a) Those in the scientific or medical field need to use and consider race/ethnicity in a more nuanced fashion. b) Those in the scientific or medical field need to use and consider ancestry/population in a more nuanced fashion. c) Those in the scientific or medical field need to consider the relationship between race, ethnicity, ancestry, and population in a more nuanced fashion.

Appropriate definitions of population categories and contexts for use	a) Certain population categories should not be employed in research. b) The scientific/medical community should continue collecting race and/or ethnicity data and investigating race- or ethnicity-based disparities. c) It is appropriate for investigators to employ population categories under certain conditions. d) Researchers should adopt or avoid certain definitions of population categories. e) Investigators should collect ancestry data instead of race or ethnicity. f) Race should not be used as a proxy for other variables in research. g) Investigators and authors should be sensitive with terminology.
The need for further research and guidelines	a) Guidelines should be developed and regulation tightened for future research involving race, ethnicity, or ancestry. b) Further research into important considerations and current pitfalls of research involving race, ethnicity, and/or ancestry should be conducted.

REFERENCES

1. Lopez, L., III, Hart, L.H., III, and Katz, M.H. (2021). Racial and Ethnic Health Disparities Related to COVID-19. JAMA *325*, 719–720.
2. Zavala, V.A., Bracci, P.M., Carethers, J.M., Carvajal-Carmona, L., Coggins, N.B., Cruz-Correa, M.R., Davis, M., de Smith, A.J., Dutil, J., Figueiredo, J.C., et al. (2021). Cancer health disparities in racial/ethnic minorities in the United States. Br. J. Cancer *124*, 315–332.
3. Burton, D.C., Flannery, B., Bennett, N.M., Farley, M.M., Gershman, K., Harrison, L.H., Lynfield, R., Petit, S., Reingold, A.L., Schaffner, W., et al. (2010). Socioeconomic and Racial/Ethnic Disparities in the Incidence of Bacteremic Pneumonia Among US Adults. Am. J. Public Health *100*, 1904–1911.
4. Forno, E., and Celedón, J.C. (2012). Health Disparities in Asthma. Am. J. Respir. Crit. Care Med. *185*, 1033–1035.
5. Graham, G. (2015). Disparities in Cardiovascular Disease Risk in the United States. Curr. Cardiol. Rev. *11*, 238–245.
6. Khazanchi, R., Evans, C.T., and Marcelin, J.R. (2020). Racism, Not Race, Drives Inequity Across the COVID-19 Continuum. JAMA Netw. Open *3*, e2019933.
7. Raghav, K., Anand, S., Gothwal, A., Singh, P., Dasari, A., Overman, M.J., and Loree, J.M.

(2021). Underreporting of race/ethnicity in COVID-19 research. Int. J. Infect. Dis. *108*, 419–421.
8. Use of Race Ethnicity and Ancestry as Population Descriptors in Genomics Research | National Academies.
9. Byeon, Y.J.J., Islamaj, R., Yeganova, L., Wilbur, W.J., Lu, Z., Brody, L.C., and Bonham, V.L. (2021). Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades. Am. J. Hum. Genet. *108*, 2215–2223.
10. Vyas, D.A., Eisenstein, L.G., and Jones, D.S. (2020). Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. N. Engl. J. Med. *383*, 874–882.
11. Borrell, L.N., Elhawary, J.R., Fuentes-Afflick, E., Witonsky, J., Bhakta, N., Wu, A.H.B., Bibbins-Domingo, K., Rodríguez-Santana, J.R., Lenoir, M.A., Gavin, J.R., et al. (2021). Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism. N. Engl. J. Med. *384*, 474–480.
12. Oni-Orisan, A., Mavura, Y., Banda, Y., Thornton, T.A., and Sebro, R. (2021). Embracing Genetic Diversity to Improve Black Health. N. Engl. J. Med. *384*, 1163–1167.
13. Getting genetic ancestry right for science and society.
14. Bliss, C. (2020). Conceptualizing Race in the Genomic Age. Hastings Cent. Rep. *50 Suppl 1*, S15–S22.
15. Ali-Khan, S.E., Krakowski, T., Tahir, R., and Daar, A.S. (2011). The use of race, ethnicity and ancestry in human genetic research. HUGO J. *5*, 47–63.
16. Sankar, P., Cho, M.K., and Mountain, J. (2007). Race and Ethnicity in Genetic Research. Am. J. Med. Genet. A. *143*, 961–970.
17. Smart, A., Tutton, R., Martin, P., Ellison, G.T.H., and Ashcroft, R. (2008). The standardization of race and ethnicity in biomedical science editorials and UK biobanks. Soc. Stud. Sci. *38*, 407–423.
18. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ n71.
19. Callier, S.L. (2019). The Use of Racial Categories in Precision Medicine Research. Ethn. Dis. *29*, 651–658.
20. (2000). Census, race and science. Nat. Genet. *24*, 97–98.
21. Duggan, C.P., Kurpad, A., Stanford, F.C., Sunguya, B., and Wells, J.C. (2020). Race, ethnicity, and racism in the nutrition literature: an update for 2020. Am. J. Clin. Nutr. *112*, 1409–1414.
22. Foster, M.W. (2009). Looking for race in all the wrong places: analyzing the lack of productivity in the ongoing debate about race and genetics. Hum. Genet. *126*, 355–362.
23. Jones, C.P. (2001). Invited Commentary: “Race,” Racism, and the Practice of Epidemiology. Am. J. Epidemiol. *154*, 299–304.
24. Kahn, J. (2006). Genes, Race, and Population: Avoiding a Collision of Categories. Am. J. Public Health *96*, 1965–1970.
25. Lee, S.S.-J., Mountain, J., Koenig, B., Altman, R., Brown, M., Camarillo, A., Cavalli-Sforza, L., Cho, M., Eberhardt, J., Feldman, M., et al. (2008). The ethics of characterizing difference: guiding principles on using racial categories in human genetics. Genome Biol. *9*, 404.
26. Winker, M.A. (2004). Measuring race and ethnicity: why and how? JAMA *292*, 1612–1614.
27. Winker, M.A. (2006). Race and Ethnicity in Medical Research: Requirements Meet Reality. J. Law. Med. Ethics *34*, 520–525.
28. Bhopal, R. (2006). Race and Ethnicity: Responsible Use from Epidemiological and Public Health Perspectives. J. Law. Med. Ethics *34*, 500–507.
29. Rothstein, M.A., and Epps, P.G. (2001). Pharmacogenomics and the (ir)relevance of race. Pharmacogenomics J. *1*, 104–108.
30. Sankar, P., and Cho, M.K. (2002). Toward a New Vocabulary of Human Genetic Variation.

Science *298*, 1337–1338.
31. Terry, S.F., Christensen, K.D., Metosky, S., Rudofsky, G., Deignan, K.P., Martinez, H., Johnson-Moore, P., and Citrin, T. (2012). Community Engagement about Genetic Variation Research. Popul. Health Manag. *15*, 78–89.
32. Khan, A.T., Gogarten, S.M., McHugh, C.P., Stilp, A.M., Sofer, T., Bowers, M., Wong, Q., Cupples, L.A., Hidalgo, B., Johnson, A.D., et al. (2021). Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: experiences from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. ArXiv210807858 Q-Bio.
33. Reverby, S.M. (2010). Invoking "Tuskegee": problems in health disparities, genetic assumptions, and history. J. Health Care Poor Underserved *21*, 26–34.
34. Lee, S.S. (2009). Pharmacogenomics and the challenge of health disparities. Public Health Genomics *12*, 170–179.
35. Lee, S.S.-J. (2005). Racializing Drug Design: Implications of Pharmacogenomics for Health Disparities. Am. J. Public Health *95*, 2133–2138.
36. Takezawa, Y., Kato, K., Oota, H., Caulfield, T., Fujimoto, A., Honda, S., Kamatani, N., Kawamura, S., Kawashima, K., Kimura, R., et al. (2014). Human genetic research, race, ethnicity and the labeling of populations: recommendations based on an interdisciplinary workshop in Japan. BMC Med. Ethics *15*, 33.
37. Capocasa, M., and Volpi, L. (2019). The ethics of investigating cultural and genetic diversity of minority groups. Homo Int. Z. Vgl. Forsch. Am Menschen *70*, 233–244.
38. Claw, K.G., Anderson, M.Z., Begay, R.L., Tsosie, K.S., Fox, K., and Garrison, N.A. (2018). A framework for enhancing ethical genomic research with Indigenous communities. Nat. Commun. *9*, 2957.
39. Ossorio, P., and Duster, T. (2005). Race and genetics: Controversies in biomedical, behavioral, and forensic sciences. Am. Psychol. *60*, 115–128.
40. Clayton, E.W. (2002). Complex Relationship of Genetics, Groups, and Health: What It Means for Public Health Symposium Article - Part III: Salient Issues in Public Health Law. J. Law. Med. Ethics *30*, 290–297.
41. Brewer, R.M. (2006). Thinking Critically about Race and Genetics. J. Law. Med. Ethics *34*, 513–519.
42. Francis, C.K. (2001). The medical ethos and social responsibility in clinical medicine. J. Natl. Med. Assoc. *93*, 157–169.
43. Kittles, R.A., and Weiss, K.M. (2003). Race, ancestry, and genes: implications for defining disease risk. Annu. Rev. Genomics Hum. Genet. *4*, 33–67.
44. Rambachan, A. (2018). Overcoming the Racial Hierarchy: the History and Medical Consequences of "Caucasian." J. Racial Ethn. Health Disparities *5*, 907–912.
45. Rusert, B.M., and Royal, C.D.M. (2011). Grassroots Marketing in a Global Era: More Lessons from BiDil. J. Law. Med. Ethics *39*, 79–90.
46. Schwartz, R.S. (2001). Racial profiling in medical research. N. Engl. J. Med. *344*, 1392–1393.
47. Bhopal, R. (2009). Medicine and public health in a multiethnic world. J. Public Health Oxf. Engl. *31*, 315–321.
48. Bonham, V.L., and Knerr, S. (2008). Social and ethical implications of genomics, race, ethnicity, and health inequities. Semin. Oncol. Nurs. *24*, 254–261.
49. Tashiro, C.J. (2005). The meaning of race in health care and research--part 1: the impact of history. Pediatr. Nurs. *31*, 208–210.
50. Tashiro, C.J. (2005). The meaning of race in healthcare and research--part 2. Current controversies and emerging research. Pediatr. Nurs. *31*, 305–308.
51. Hoffman, S. (2005). "Racially-Tailored" Medicine Unraveled (Rochester, NY: Social Science Research Network).
52. Ruel, M.D. (2006). Using race in clinical research to develop tailored medications. Is the

FDA encouraging discrimination or eliminating traditional disparities in health care for African Americans? J. Leg. Med. *27*, 225–241.
53. Sharp, R.R., and Foster, M.W. (2007). Grappling with groups: protecting collective interests in biomedical research. J. Med. Philos. *32*, 321–337.
54. Hausman, D. (2008). Protecting groups from genetic research. Bioethics *22*, 157–165.
55. Ilkilic, I., and Paul, N.W. (2009). Ethical aspects of genome diversity research: genome research into cultural diversity or cultural diversity in genome research? Med. Health Care Philos. *12*, 25–34.
56. Sharp, R.R., and Foster, M.W. (2002). An analysis of research guidelines on the collection and use of human biological materials from American Indian and Alaskan Native communities. Jurimetrics *42*, 165–186.
57. Bankoff, R.J., and Perry, G.H. (2016). Hunter-gatherer genomics: Evolutionary insights and ethical considerations. Curr. Opin. Genet. Dev. *41*, 1–7.
58. Boyer, B.B., Dillard, D., Woodahl, E.L., Whitener, R., Thummel, K.E., and Burke, W. (2011). Ethical issues in developing pharmacogenetic research partnerships with American Indigenous communities. Clin. Pharmacol. Ther. *89*, 343–345.
59. Hiratsuka, V.Y., Hahn, M.J., Woodbury, R.B., Hull, S.C., Wilson, D.R., Bonham, V.L., Dillard, D.A., Avey, J.P., Beckel-Mitchener, A.C., Blome, J., et al. (2020). Alaska Native genomic research: perspectives from Alaska Native leaders, federal staff, and biomedical researchers. Genet. Med. *22*, 1935–1943.
60. McGregor, J.L. (2007). Population genomics and research ethics with socially identifable groups. J. Law Med. Ethics J. Am. Soc. Law Med. Ethics *35*, 356–370.
61. Foster, M.W., and Sharp, R.R. (2000). Genetic research and culturally specific risks: one size does not fit all. Trends Genet. TIG *16*, 93–95.
62. Foster, M.W., and Sharp, R.R. (2002). Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. Genome Res. *12*, 844–850.
63. Fullerton, S.M., Yu, J.-H., Crouch, J., Fryer-Edwards, K., and Burke, W. (2010). Population description and its role in the interpretation of genetic association. Hum. Genet. *127*, 563–572.
64. Race, Ethnicity, and Genetics Working Group (2005). The use of racial, ethnic, and ancestral categories in human genetics research. Am. J. Hum. Genet. *77*, 519–532.
65. Barr, D.A. (2005). The Practitioner's Dilemma: Can We Use a Patient's Race To Predict Genetics, Ancestry, and the Expected Outcomes of Treatment? Ann. Intern. Med. *143*, 809–815.
66. Matthews-Juarez, P., and Juarez, P.D. (2011). Cultural competency, human genomics, and the elimination of health disparities. Soc. Work Public Health *26*, 349–365.
67. Mersha, T.B., and Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum. Genomics *9*, 1.
68. Bonham, V.L., Green, E.D., and Pérez-Stable, E.J. (2018). Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. JAMA *320*, 1533–1534.
69. Whittle, P.M. (2010). Health, inequality and the politics of genes. N. Z. Med. J. *123*, 67–75.
70. Batai, K., Hooker, S., and Kittles, R.A. (2021). Leveraging genetic ancestry to study health disparities. Am. J. Phys. Anthropol. *175*, 363–375.
71. Braun, L. (2002). Race, Ethnicity, and Health: Can Genetics Explain Disparities? Perspect. Biol. Med. *45*, 159–174.
72. Egalité, N., Ozdemir, V., and Godard, B. (2007). Pharmacogenomics research involving racial classification: qualitative research findings on researchers' views, perceptions and attitudes towards socioethical responsibilities. Pharmacogenomics *8*, 1115–1126.
73. Lewontin, R. (2005). The fallacy of racial medicine: confusions about human races. Genewatch Bull. Comm. Responsible Genet.
74. Martschenko, D.O., and Smith, M. (2021). Genes do not operate in a vacuum, and neither should our research. Nat. Genet. *53*, 255–256.

75. Palsson, G., and Helgason, A. (2003). Blondes, lost and found: representations of genes, identity, and history. Dev. World Bioeth. *3*, 159–169.
76. Balaban, E. (2005). The new racial economy: making a silk purse out of the sow's ear of racial distinctions. Genewatch Bull. Comm. Responsible Genet. *18*, 8–10.
77. Lee, S.S.-J. (2003). Race, Distributive Justice and the Promise of Pharmacogenomics. Am. J. Pharmacogenomics *3*, 385–392.
78. Lee, S.S., Mountain, J., and Koenig, B.A. (2001). The meanings of "race" in the new genomics: implications for health disparities research. Yale J. Health Policy Law Ethics *1*, 33–75.
79. Parrott, R.L., Silk, K.J., Dillow, M.R., Krieger, J.L., Harris, T.M., and Condit, C.M. (2005). Development and validation of tools to assess genetic discrimination and genetically based racism. J. Natl. Med. Assoc. *97*, 980–990.
80. Walajahi, H., Wilson, D.R., and Hull, S.C. (2019). Constructing identities: the implications of DTC ancestry testing for tribal communities. Genet. Med. Off. J. Am. Coll. Med. Genet. *21*, 1744–1750.
81. Ozdemir, V., Graham, J.E., and Godard, B. (2008). Race as a variable in pharmacogenomics science: from empirical ethics to publication standards. Pharmacogenet. Genomics *18*, 837–841.
82. Bloche, M.G. (2004). Race-Based Therapeutics. N. Engl. J. Med. *351*, 2035–2037.
83. Popejoy, A.B., Crooks, K.R., Fullerton, S.M., Hindorff, L.A., Hooker, G.W., Koenig, B.A., Pino, N., Ramos, E.M., Ritter, D.I., Wand, H., et al. (2020). Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. Am. J. Hum. Genet. *107*, 72–82.
84. Smart, A., Martin, P., and Parker, M. (2004). Tailored medicine: whom will it fit? The ethics of patient and disease stratification. Bioethics *18*, 322–342.
85. Aldhous, P. (2002). Geneticist fears "race-neutral" studies will fail ethnic groups. Nature *418*, 355–356.
86. Burchard, E.G. (2014). Medical research: Missing patients. Nature *513*, 301–302.
87. Stevens, J. (2003). Racial Meanings and Scientific Methods: Changing Policies for NIH-Sponsored Publications Reporting Human Variation. J. Health Polit. Policy Law *28*, 1033–1088.
88. Nugent, A., Conatser, K.R., Turner, L.L., Nugent, J.T., Sarino, E.M.B., and Ricks-Santi, L.J. (2019). Reporting of race in genome and exome sequencing studies of cancer: a scoping review of the literature. Genet. Med. Off. J. Am. Coll. Med. Genet. *21*, 2676–2680.
89. Popejoy, A.B., Ritter, D.I., Crooks, K., Currey, E., Fullerton, S.M., Hindorff, L.A., Koenig, B., Ramos, E.M., Sorokin, E.P., Wand, H., et al. (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum. Mutat. *39*, 1713–1720.
90. Peterson-Iyer, K. (2008). Pharmacogenomics, ethics, and public policy. Kennedy Inst. Ethics J. *18*, 35–56.
91. Angel, R.J. (2011). Agency versus structure: genetics, group membership, and a new twist on an old debate. Soc. Sci. Med. 1982 *73*, 632–635.
92. Winkelmann, B.R. (2003). Pharmacogenomics, genetic testing and ethnic variability: tackling the ethical questions. Pharmacogenomics *4*, 531–535.
93. Garte, S. (2002). The racial genetics paradox in biomedical research and public health. Public Health Rep. *117*, 421–425.
94. Anomaly, J. (2017). Race Research and the Ethics of Belief. J. Bioethical Inq. *14*, 287–297.
95. Foster, M.W., and Sharp, R.R. (2004). Beyond race: towards a whole-genome perspective on human populations and genetic variation. Nat. Rev. Genet. *5*, 790–796.
96. Feller, L., Ballyram, R., Meyerov, R., Lemmer, J., and Ayo-Yusuf, O.A. (2014). Race/ethnicity in biomedical research and clinical practice. SADJ J. South Afr. Dent. Assoc. Tydskr. Van Suid-Afr. Tandheelkd. Ver. *69*, 272–274.

97. Frank, R. (2008). Functional or futile?: the (in)utility of methodological critiques of genetic research on racial disparities in health. A commentary on Kaufman's "Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example." Soc. Sci. Med. 1982 *66*, 1670–1674.
98. Sade, R.M. (2007). What's right (and wrong) with racially stratified research and therapies. J. Natl. Med. Assoc. *99*, 693–696.
99. Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T.L., Huang, J., Vujkovic, M., Damrauer, S.M., Pyarajan, S., Gaziano, J.M., et al. (2019). Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. Am. J. Hum. Genet. *105*, 763–772.
100. Hardimon, M.O. (2013). Race concepts in medicine. J. Med. Philos. *38*, 6–31.
101. Payne, P.W., and Royal, C. (2007). The Role of Genetic and Sociopolitical Definitions of Race in Clinical Trials. J. Am. Acad. Orthop. Surg.
102. Shields, A.E., Fortun, M., Hammonds, E.M., King, P.A., Lerman, C., Rapp, R., and Sullivan, P.F. (2005). The use of race variables in genetic studies of complex traits and the goal of reducing health disparities: A transdisciplinary perspective. Am. Psychol. *60*, 77–103.
103. Burchard, E.G., Ziv, E., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Pérez-Stable, E.J., Sheppard, D., and Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. N. Engl. J. Med. *348*, 1170–1175.
104. Dawson, G. (2003). Human genome, race and medicine. J. Natl. Med. Assoc. *95*, 309–312.
105. Shanawani, H., Dame, L., Schwartz, D.A., and Cook-Deegan, R. (2006). Non-reporting and inconsistent reporting of race and ethnicity in articles that claim associations among genotype, outcome, and race or ethnicity. J. Med. Ethics *32*, 724–728.
106. Kaufman, J.S., and Cooper, R.S. (2001). Commentary: Considerations for Use of Racial/Ethnic Classification in Etiologic Research. Am. J. Epidemiol. *154*, 291–298.
107. Kahn, J. (2005). Ethnic drugs. Hastings Cent. Rep. *35*, 1 p following 48.
108. Lillquist, E., and Sullivan, C.A. (2006). Legal regulation of the use of race in medical research. J. Law Med. Ethics J. Am. Soc. Law Med. Ethics *34*, 535–551, 480.
109. Jaja, C., Gibson, R., and Quarles, S. (2013). Advancing Genomic Research and Reducing Health Disparities: What Can Nurse Scholars Do? J. Nurs. Scholarsh. *45*, 202–209.
110. Umek, W., and Fischer, B. (2020). We Should Abandon "Race" as a Biological Category in Biomedical Research. Female Pelvic Med. Reconstr. Surg. *26*, 719–720.
111. Lorusso, L. (2011). The justification of race in biological explanation. J. Med. Ethics *37*, 535–539.
112. Eichelberger, K.Y., Alson, J.G., and Doll, K.M. (2018). Should Race Be Used as a Variable in Research on Preterm Birth? AMA J. Ethics *20*, 296–302.
113. Hunt, L.M., Truesdell, N.D., and Kreiner, M.J. (2013). Genes, race, and culture in clinical care: racial profiling in the management of chronic illness. Med. Anthropol. Q. *27*, 253–271.
114. Jones, T., and Roberts, J.L. (2020). GENETIC RACE? DNA ANCESTRY TESTS, RACIAL IDENTITY, AND THE LAW. Columbia Law Rev. *120*, 1929–2016.
115. Schaefer, G.O., Tai, E.S., and Sun, S.H.-L. (2020). Navigating conflicts of justice in the use of race and ethnicity in precision medicine. Bioethics *34*, 849–856.
116. Jones, D.S., and Perlis, R.H. (2006). Pharmacogenetics, race, and psychiatry: prospects and challenges. Harv. Rev. Psychiatry *14*, 92–108.
117. Craddock Lee, S.J. (2005). The Risks of Race in Addressing Health Disparities. Hastings Cent. Rep. *35*, 1p–48.
118. Lee, S.S.-J. (2007). The ethical implications of stratifying by race in pharmacogenomics. Clin. Pharmacol. Ther. *81*, 122–125.
119. Cohn, J.N. (2006). The Use of Race and Ethnicity in Medicine: Lessons from the African-American Heart Failure Trial. J. Law. Med. Ethics *34*, 552–554.

120. Royal, C.D., Novembre, J., Fullerton, S.M., Goldstein, D.B., Long, J.C., Bamshad, M.J., and Clark, A.G. (2010). Inferring genetic ancestry: opportunities, challenges, and implications. Am. J. Hum. Genet. *86*, 661–673.
121. Slabbert, N., and Heathfield, L.J. (2018). Ethical, legal and social implications of forensic molecular phenotyping in South Africa. Dev. World Bioeth. *18*, 171–181.
122. Zhang, F., and Finkelstein, J. (2019). Inconsistency in race and ethnic classification in pharmacogenetics studies and its potential clinical implications. Pharmacogenomics Pers. Med. *12*, 107–123.
123. Lee, S.S.-J., Soo-Jin Lee, S., Bolnick, D.A., Duster, T., Ossorio, P., and Tallbear, K. (2009). Genetics. The illusive gold standard in genetic ancestry testing. Science *325*, 38–39.
124. Bamshad, M. (2007). Lost in translation: Meaningful policies for writing about genetics and race. Am. J. Med. Genet. A.
125. Morning, A. (2015). Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round. In Social Statistics and Ethnic Diversity: Cross-National Perspectives in Classifications and Identity Politics, P. Simon, V. Piché, and A.A. Gagnon, eds. (Cham: Springer International Publishing), pp. 17–37.

Perspective

The Human Pangenome Project: a global resource to map genomic diversity

https://doi.org/10.1038/s41586-022-04601-8

Received: 30 August 2021

Accepted: 1 March 2022

Published online: 20 April 2022

Check for updates

Ting Wang[1,2,3]✉, Lucinda Antonacci-Fulton[3], Kerstin Howe[4], Heather A. Lawson[1], Julian K. Lucas[5], Adam M. Phillippy[6], Alice B. Popejoy[7], Mobin Asri[5], Caryn Carson[1,2,3], Mark J. P. Chaisson[8], Xian Chang[5], Robert Cook-Deegan[9], Adam L. Felsenfeld[10], Robert S. Fulton[3], Erik P. Garrison[11], Nanibaa' A. Garrison[12,13,14], Tina A. Graves-Lindsay[3], Hanlee Ji[15], Eimear E. Kenny[16,17,18], Barbara A. Koenig[19], Daofeng Li[1,2,3], Tobias Marschall[20], Joshua F. McMichael[3], Adam M. Novak[5], Deepak Purushotham[1,2,3], Valerie A. Schneider[21], Baergen I. Schultz[10], Michael W. Smith[10], Heidi J. Sofia[10], Tsachy Weissman[22], Paul Flicek[23]✉, Heng Li[24,25]✉, Karen H. Miga[5]✉, Benedict Paten[5]✉, Erich D. Jarvis[26,27]✉, Ira M. Hall[28]✉, Evan E. Eichler[29,30]✉, David Haussler[5,31]✉ & the Human Pangenome Reference Consortium*

The human reference genome is the most widely used resource in human genetics and is due for a major update. Its current structure is a linear composite of merged haplotypes from more than 20 people, with a single individual comprising most of the sequence. It contains biases and errors within a framework that does not represent global human genomic variation. A high-quality reference with global representation of common variants, including single-nucleotide variants, structural variants and functional elements, is needed. The Human Pangenome Reference Consortium aims to create a more sophisticated and complete human reference genome with a graph-based, telomere-to-telomere representation of global genomic diversity. Here we leverage innovations in technology, study design and global partnerships with the goal of constructing the highest-possible quality human pangenome reference. Our goal is to improve data representation and streamline analyses to enable routine assembly of complete diploid genomes. With attention to ethical frameworks, the human pangenome reference will contain a more accurate and diverse representation of global genomic variation, improve gene–disease association studies across populations, expand the scope of genomics research to the most repetitive and polymorphic regions of the genome, and serve as the ultimate genetic resource for future biomedical research and precision medicine.

The human reference genome is the foundational open-access resource of modern human genetics and genomics, providing a centralized coordinate system for reporting and comparing results across studies[1–4]. Its release set the bar for genomic data sharing, essential for nearly all human genomics applications, including alignments, variant detection and interpretation, functional annotations, population genetics and epigenomic analyses. The current human reference (GRCh38.p13) is a mosaic of genomic data assembled from more than 20 individuals, with approximately 70% of the sequence contributed by a single individual[5,6,7]. Dependence on a single mosaic assembly (which does not

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [2]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA. [3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. [4]Wellcome Sanger Institute, Cambridge, UK. [5]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [6]Genome Informatics Section, National Human Genome Research Institute, Bethesda, MD, USA. [7]Epidemiology Division, Department of Public Health Sciences, University of California, Davis, CA, USA. [8]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA. [9]Arizona State University, Barrett & O'Connor Washington Center, Washington DC, USA. [10]National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD, USA. [11]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA. [12]Institute for Society & Genetics, College of Letters and Science, University of California, Los Angeles, Los Angeles, CA, USA. [13]Institute for Precision Health, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [14]Division of General Internal Medicine & Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [15]Department of Medicine, Stanford University, School of Medicine, Stanford, CA, USA. [16]Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [17]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [18]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [19]Program in Bioethics and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [20]Heinrich Heine University, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany. [21]National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD, USA. [22]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [23]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK. [24]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [25]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [26]Vertebrate Genome Lab and and Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA. [27]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [28]Yale School of Medicine, New Haven, CT, USA. [29]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [30]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [31]Howard Hughes Medical Institute, University of California, Santa Cruz, CA, USA. *A list of members and their affiliations appears in the Supplementary Information. ✉e-mail: twang@wustl.edu; flicek@ebi.ac.uk; hli@jimmy.harvard.edu; khmiga@ucsc.edu; bpaten@ucsc.edu; ejarvis@rockefeller.edu; ira.hall@yale.edu; eee@gs.washington.edu; haussler@ucsc.edu

represent the sequence of any one person) creates reference biases, adversely affecting variant discovery, gene–disease association studies and the accuracy of genetic analyses[8,9]. More than two decades after the first human genome reference sequences were released, the current reference genome still contains errors, rare structural configurations that do not exist in most human genomes, and gaps in regions that have been difficult to assemble[7,10] because of their repetitive and highly polymorphic nature. The human reference genome, like most technology-driven resources, is overdue for an upgrade[11].

For years, the Genome Reference Consortium has updated the linear reference by fixing errors, filling in gaps and adding newly discovered variants[1,4,7,12]. When enough changes accumulate, new builds are generated and released. Although this process has served the community well, shortcomings have been identified along the way. Segments of genome sequences sampled from individuals may differ considerably from the reference genome, leading to errors in read mapping to the reference and reducing the accuracy of variant calls[13,14]. Identification of structural variants (more than 50-bp deletions, insertions, tandem duplications, inversions and translocations) relies on detecting patterns of discordant read pairs or split read alignments, which in turn depend on the accuracy of read mapping[15,16]. Assembling and detecting these structural variants are challenging when the reads are too short to cover long, repetitive regions of the genome[8]. This is because short reads (50–300 bp) from different repeats may be identical and/or overlapping with one another such that it is impossible to determine where they should map. Both the limitations of short reads and reference biases mean that we may have missed more than 70% of structural variants in traditional whole-genome sequencing studies[17,18].

Advances in sequencing technologies and a greater appreciation for the importance of genetic diversity make improving the human reference sequence both timely and practical. First, the development of long-read (more than 10 kb) sequencing technologies has enabled the assembly of large, repeat-rich regions, facilitated phasing and assembly of maternal and paternal haplotypes, and improved representation of GC-rich regions of the genome that are often missing in short-read assemblies[8,19–22]. Second, growing recognition of the importance of diversity and inclusion in human genomics[23] has led to widespread calls to improve representation and methods for detecting and presenting global variation.

In this Perspective article, we outline the goals, strategies, challenges and opportunities for the Human Pangenome Reference Consortium (HPRC). We will engage scientists and bioethicists in creating a human pangenome reference and resource that represents genomic diversity across human populations, as well as improving technology for assembly and developing an ecosystem of tools for analyses of graph-based genome sequences. This new reference will maintain essential ties to the original reference for continuity, even as we strive to develop complete and error-free telomere-to-telomere (T2T) assemblies of all chromosomes of individual human genomes, referred to here as 'haplotypes'.

Goals and strategies of the HPRC

A 'pangenome' is the collective whole-genome sequences of multiple individuals representing the genetic diversity of the species. Originally popularized in the context of highly dynamic bacterial genomes[24], the concept has been adapted to the field of human genomics, in which the full extent of human genomic variation is expected to be much broader than has thus far been revealed. The pangenome data infrastructure depends on the high-throughput production of high-quality, phased haplotypes (segments of a chromosome identified as being maternally or paternally inherited) that improve upon the current human reference genome. Highly accurate and complete haplotype-phased genome assemblies will be organized into a graph-based data structure for the pangenome reference that compresses and indexes information[25–27]. This data structure will contain a coordinate system with a simple, intuitive framework for referring to genomic variants, as well as preserving backward compatibility with GRCh38 and previous linear reference builds. Managing and interpreting these data require transdisciplinary collaboration and innovation, focused on the development of novel conceptual frameworks and analytic methods to construct the pangenome infrastructure and tools for downstream analyses and visualization. The goals of the HPRC are laid out in Box 1.

The HPRC functions through multidisciplinary collaborations, convening cross-institutional and multinational working groups dedicated to sample collection and consent, population genetic diversity, technology and production, phasing and assembly, approaches to construction of a pangenome reference, resource improvement and maintenance, and resource sharing and outreach (Fig. 1). The HPRC has begun the process of engaging international partnerships with the Australian National Centre for Indigenous Genomics (NCIG; https://ncig.anu.edu.au), the US Food and Drug Administration (FDA)-recognized Clinical Genome Resource (ClinGen)[28], the National Institutes of Health (NIH)-funded Human Heredity and Health in Africa (H3Africa; https://h3africa.org) Consortium, the Personal Genome Project (PGP; https://www.personalgenomes.org), the Vertebrate Genomes Project[8] and the Global Alliance for Genomics and Health (GA4GH; https://www.ga4gh.org). The HPRC will integrate perspectives from the international scientific community through these collaborators and others yet to be identified to inform the development of HPRC references, methods and standards.

Inclusion criteria

For initial inclusion, the HPRC selected individual genomes for high-quality sequencing among existing cell lines established by the 1000

Box 1

Goals of the HPRC

- Identify individuals from diverse genomic and biogeographical backgrounds to include in the pangenome reference, with at least 350 reference quality haplotype-phased human diploid genomes (700 haplotypes in total).
- Integrate ethical, legal and social implications (ELSI) scholarship in the development of recommended policies and protocols for inclusion, data acquisition and stewardship from study recruitment to publication of findings.
- Prioritize the use of long-read and long-range technologies for assemblies, with haplotype-aware algorithms to generate the highest quality phased genomes possible.
- Create methods to finish diploid genomes from T2T across complex regions, closing gaps and ensuring hard-to-measure variants are identified.
- Foster an ecosystem of pangenome reference tools to facilitate the annotation of genes and other genomic features.
- Implement an iterative design–development–engagement process to understand and respond to user community needs.
- Develop communication strategies that will assure understanding of the pangenome reference resource, including the ability of the community to fix and report errors.
- Enable appropriately controlled access to data through genomics platforms such as the INSDC[59,71], the NCBI, the UCSC Genome Browser[72,73], Ensembl[74,75], the WashU Epigenome Browser[76,77] and NHGRI's cloud-based analysis platform AnVIL[46,78].
- Foster an international human pangenome reference alliance that actively engages the diverse populations it seeks to represent.

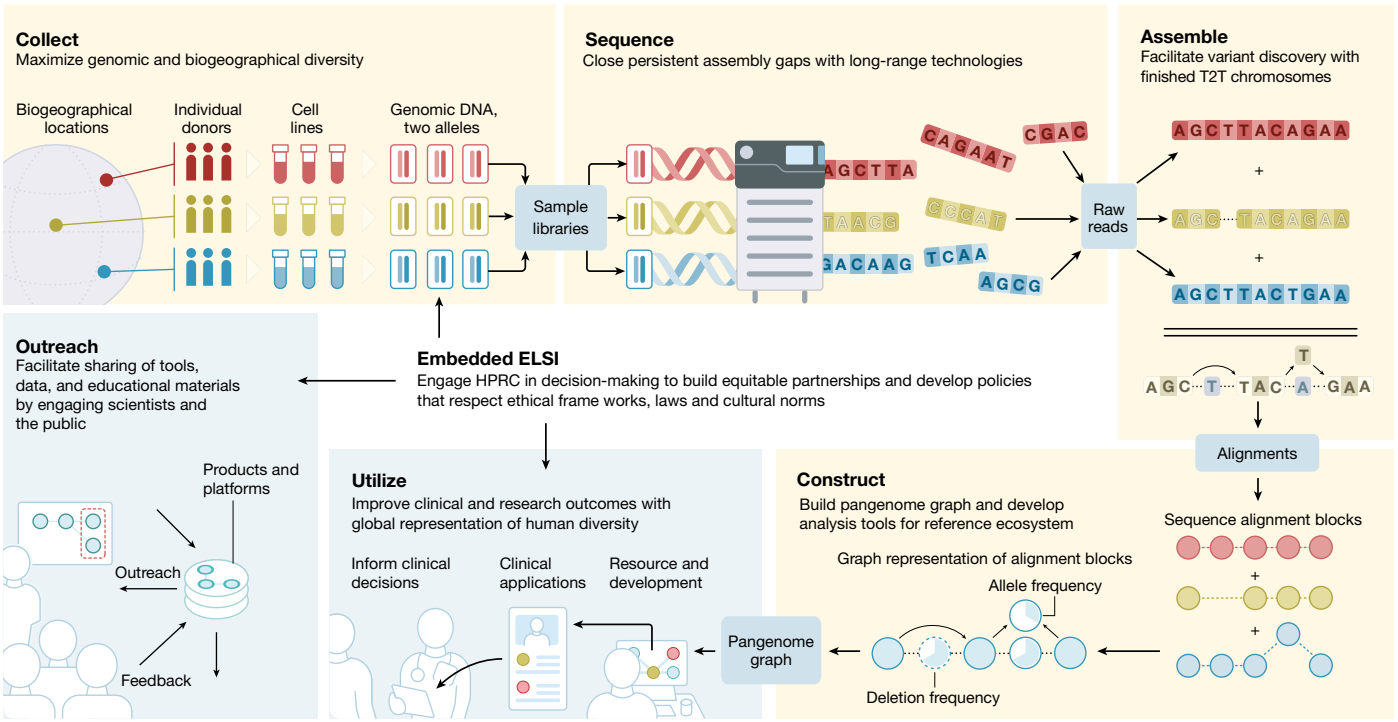


Fig. 1 | The HPRC. An overview of several components of the HPRC. Collect: 1,000 Genomes samples start the project and will be followed by additional samples collected through community engagement and recruitment. Sample selection efforts will ensure that the graph-based reference captures global human genomic diversity. Sequence: long-read and long-range technologies are used to generate genome graphs and bridge gaps in difficult-to-assemble genomic regions. Assemble: T2T finished diploid genomes will foster variant discovery, especially in complex, difficult-to-assemble genomic regions. Construct: scalable bioinformatics approaches assemble, quality control, call variants and benchmark graph assembly accuracy. The graph is annotated with gene descriptions and transcriptome data, making it more accessible and interpretable. Utilize: collaboration across scientific and stakeholder communities will create a new ecosystem of analysis tools. Clinical applications and research use will involve analysis, validation, interpretation and publication of results. Outreach: members of the HPRC outreach community engage and educate the user community and broadly share all genomic products and informatics platforms. ELSI: ELSI scholars will develop selection processes and policy frameworks that meet investigator needs as well as respect research partner autonomy and cultural norms.

Genomes Project (1KGP), which offers a deep catalogue of human variation from 26 populations[29]. These cells were originally collected from volunteer donors using consent procedures designed for unrestricted data use, and the cell lines are available in the National Human Genome Research Institute (NHGRI) Biorepository at Coriell (https://www.coriell.org). The selected cell lines were prioritized on the basis of a combination of criteria, ranging from genetic and geographical diversity of the donors, to the availability of relevant parental data (for haplotype phasing), and limited time in cell culture (to minimize the accumulation of de novo mutations).

Differences between individuals were initially identified using clustering and visualization techniques (uniform manifold approximation and projection clusters generated from 1KGP data) and observed allelic diversity (heterozygosity), and then selected for inclusion in the first phase of the HPRC ($n = 100$). Our inclusion criteria and recruitment strategies are evolving with the project, and we recognize that there are inherent limitations to clustering algorithms and using only the 1KGP dataset.

Although useful for the first phase of the HPRC, genomes selected from the 1KGP data represent a limited scope of geographical and genomic diversity. One reason is that the resource was developed by sampling in 26 geographical locations across the globe, and the discrete number of individuals included from each location limits the amount of genomic variation representing those regions, especially regarding rare variants that are less likely to be observed in small sample sizes. The genomes of individuals sampled from each 1KGP location cannot be assumed to have sufficient variation to be comprehensive of the genomic diversity in the natural population of the region, let alone to represent an entire continent. Furthermore, 1KGP populations were often selected by asking potential study participants questions about their racial, ethnic or ancestral identities, assigning ancestry on the basis of geographical location, or some combination, which would not necessarily produce a representative sampling of any natural population. As population descriptors can be inconsistent on clinical forms[30] and are fluid across cultural contexts[31–33], there are many unknown layers of diversity within each geographical sampling cohort of the 1KGP data.

Because the 1KGP data are insufficient to support the ambitious sampling and genetic diversity goals of the HPRC, the consortium will include additional genomes, including from participants identified through the BioMe Biobank at Mount Sinai and a cohort of African American individuals recruited by Washington University. All participants will give informed consent, and their sequence data will be deindentified in open access. Some will have cell lines generated at Coriell. In later phases, the HPRC will foster additional domestic and international partnerships to explore additional avenues to broaden diversity and enhance inclusion (Box 2).

Embedded ELSI scholarship

Most human genomics has been based on individuals of European ancestry, and the datasets available for analyses are thus biased. As a result, current precision medicine is based on genomic variation found in populations with primarily European ancestry. Much of the global genetic diversity that contributes to clinical phenotypes is missing from clinical genetic tests. Many ethical, legal and social challenges arise in efforts to include previously excluded populations, communities or groups.

Box 2

Commitment to diversity and inclusion

There are many aspects of diversity to consider for broad inclusion, and the first step is to assess current gaps in diversity. Researchers have demonstrated a lack of diversity in genomics research using biogeographical ancestry groupings at the continental level, as well as sociocultural categories such as racial and ethnic identities[79]. It is important to distinguish biological and sociocultural diversity, as sociocultural labels are not derived from genotype data, and vice versa. Owing to gaps in genomic sampling worldwide, the distribution of allelic variants appears to be correlated with continental-level biogeographical 'ancestry'. However, variants are rarely unique to a single biogeographical 'population', and factors such as effective population size, founder events and genetic drift are responsible for differences in allele frequencies between such groups.

Capturing the full range of human genomic diversity is a daunting task: some gaps are understood and predictable, but we also face 'unknown' unknowns. The initial HPRC dataset cannot be comprehensive of global genomic variation, but it can set a foundation to build on. The HPRC will initially produce high-quality genome data for 350 individuals (700 haploid genomes) selected to maximize global representation within the logistical constraints of the initial HPRC efforts. Strategic partnerships with organizations such as the GA4GH and H3Africa are underway, which we anticipate will help to facilitate international engagement and broaden our understanding of the cultural, ethical, legal, social and political considerations of the HPRC. However, further partnerships will be needed to include populations that are underrepresented or entirely missing from current data resources. The HPRC actively welcomes additional partners and collaborators to join us in rising to this challenge.

Box 3

Sequencing and assembly

Notable improvements in long-read technologies have resulted in complete chromosome assemblies[20–22] and have demonstrated the ability to broaden variant analysis to span large, complex human structural variants[51]. Use of highly accurate consensus reads (99.9%, or Q30) of moderate length (for example, 10–20 kb), such as high fidelity (HiFi) reads from PacBio, routinely resolves long tandem repeats, or satellite arrays, and large segmental duplications[20,21,63]. In parallel, the nanopore-based sequencing platform (Oxford Nanopore Technologies) offers long-read data that routinely generate substantial coverage of reads that are hundreds of kilobytes in length (or 'ultra-long' data) with an increasing number of reported reads greater than 1 million bases. Like the HiFi data, ultra-long data is used to close large and persistent assembly gaps, including in human centromeres, subtelomeric regions and large segmental duplications[22,63]. Furthermore, chromosome conformation capture methods produce long-range data for both short-read (Hi-C[80]) and long-read (Pore-C[81]) sequencing. Such chromatin crosslinking protocols generate chimeric DNA fragments from interacting chromosomal regions that are covalently linked together. These ligated DNA molecules are sequenced to help determine phasing and spatial organization at the level of an entire chromosome. With continuing gains in both HiFi read length and nanopore single-read base-level quality, and improved methods for the use of chromosome conformation capture methods to guide phased haplotype assembly, we are entering into a new era of routine complete chromosome-level assemblies[20–22,78]. In collaboration with the T2T Consortium, which aims to use long-read sequencing and cutting-edge algorithmic approaches to close the hundreds of gaps persisting in the human reference genome and that of other species[22], the HPRC will generate accurate assemblies of entire chromosomes. These assemblies will empower us to characterize variations in large, repeat-rich regions that have historically been out of reach for standard genetic analysis and interpretation.

The HPRC has formed an 'embedded' team of scholars to address ethical, legal and social implications (ELSI) of its work, with expertise at the intersections of genomics with biomedical ethics, law, social sciences, demography, community engagement and population genetics. The main objective of the HPRC-ELSI team is to identify, investigate and ultimately offer consortium investigators advice about the issues they face, which must be addressed if the HPRC is to meet its goals. In the embedded model, with ELSI scholars participating in key meetings during which decisions are made, investigators can engage these colleagues in discussions that deepen their understanding and appreciation of what is at stake as we seek to improve the human reference genome.

Large-scale human population genetics projects aimed at broadening the diversity in genomic datasets and analyses have often missed the mark in demonstrating respect for individuals and communities. The Human Genome Diversity Project encountered strong opposition three decades ago[34], facing objections that its approach was extractive and its goals benefitted scientists and institutions in rich countries, but did not match the priorities of Indigenous peoples or people in resource-poor regions who were asked to donate their samples and data. The Havasupai Tribe of northern Arizona sued the Arizona Board of Regents in 2002 when they learned that samples donated for diabetes research were shared with other researchers and re-used for studies of schizophrenia and population origins, to which tribal members did not agree. That case was settled in 2010 (refs. [35,36]), but the effect it had on relations between tribal communities and genomics research has persisted.

More recently, the Wellcome Sanger Institute was criticized for licensing access to data arising from southern African samples, despite institutions in Africa asserting terms of informed consent that did not permit commercial uses. The NIH was also criticized for inadequate tribal engagement and consultation in the All of Us programme[37,38]. With a keen awareness of this history, the HPRC has initiated a process to consult with, engage and genuinely include groups who are currently not well represented in the genomic database. Indigenous scholars have spearheaded a movement for Indigenous data sovereignty[39], for example, including the development of the CARE (collective benefit, authority to control, responsibility, and ethics) principles for Indigenous data governance[40] to layer onto the FAIR (findable, accessible, interoperable, reusable) principles that support the open-science approach that the HPRC and similar projects take[41]. The HPRC is reaching out to Indigenous geneticists, leaders and community members to engage and collaboratively develop a truly global and inclusive reference resource, taking into account FAIR and CARE principles. Furthermore, similar efforts will be made for other diverse populations that the HPRC will work with.

Some groups who we seek to engage with may develop sampling and sequencing efforts parallel to, rather than directly participating in, the

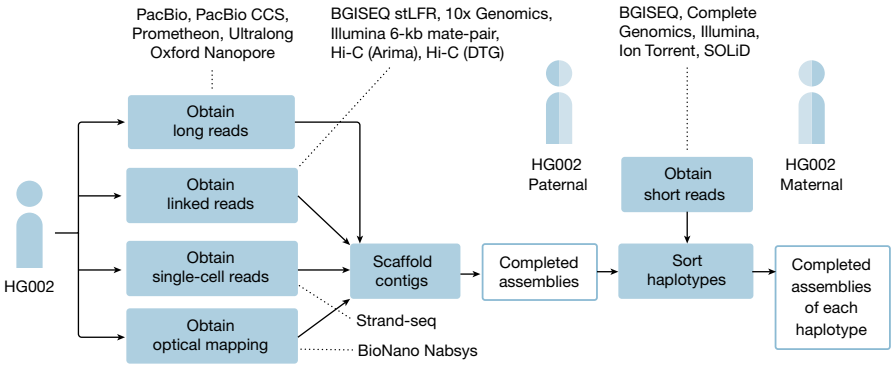


Fig. 2 | Standards were developed through a pilot benchmark study of one individual. Multiple long-read and long-range technologies and computational methods were evaluated to develop the combination of platforms and an automated pipeline that provides the most complete and accurate genome graph. CCS, circular consensus sequencing; stLFR, single-tube long fragment read.

HPRC. In developing a state-of-the-art pangenome reference sequence, the HPRC will continue to disseminate standards for accuracy and completeness of sequencing, as well as emphasizing the importance of ELSI considerations. It is a priority for the HPRC to actively communicate with parallel genome-wide sequencing efforts to ensure compatibility between efforts, thus enabling integration into a global pangenome reference resource. The HPRC is committed to assessing local policies and promoting broad sharing of resources developed through interdisciplinary engagement, scholarship and innovative technical solutions. The HPRC will establish procedures for navigating potential tensions among its technical, research and resource-generating objectives with local customs, laws and data sharing policies for the groups within the HPRC as well as for those in the parallel projects.

Initial data generation and release

Technological advances in genomics enable sequencing long repeats, physical mapping to chromosomes, and phasing maternally and paternally inherited haplotypes (Box 3).

For the initial phase of the project, we sequenced a single individual, HG002, whose genomic sequence has been thoroughly characterized by the Genome in a Bottle (GIAB) Consortium[42]. We evaluated multiple sequencing technologies and assembly algorithms to identify the optimal combination of platforms and develop an automated pipeline that generated the most complete and accurate genome representation[43] (Fig. 2). We began with the now well-established assumption that long reads (more than 10 kb) yield more complete genome assemblies than short reads alone[8]. The technologies tested included Pacific Biosciences (PacBio) and/or ONT long reads for generating contigs, 10x Genomics linked reads, Hi-C paired reads, Strand-seq long reads, and/or BioNano optical maps for scaffolding contigs into chromosomes. This pilot benchmark study created the standards for sequencing technologies and computational methodologies that are critical to the success of the HPRC.

We found that the trio approaches using parental short-read sequence data to sort haplotypes of the long-read data of offspring gave the most complete assemblies of each haplotype with the fewest structural errors[43]. Furthermore, all methods attempting to separate haplotype sequences performed much better in generating highly contiguous assemblies than those that merged the consensus between haplotypes into one assembly. The algorithm that gave the highest haplotype separation accuracy for contigs was HiFiasm[44], which incorporates separation of reads of each haplotype into the assembly graph[45]. Generation of contigs were more structurally accurate than scaffolds, where the HPRC identified areas of improvements that were necessary to prevent contig miss-joins, missed-joins, collapsed repeats and other structural assembly errors. On the basis of these findings, an initial set of 47 1KGP genomes from parent–offspring trios was assembled with HiFiasm[43,44], creating high-quality diploid contig-only genome assemblies. Going forward, we will further optimize sequencing, assembly and analysis methods with the goal of creating fully-phased T2T diploid genomes, including repetitive and structurally variable regions such as centromeres, telomeres and segmental duplications. We anticipate that the high-quality assemblies created in the project will drive tool creation and improvement for diploid genome assembly and quality control in which new and recently created existing tools (from the T2T assembly of CHM13 (ref. [22])) are applied to diploid genome assembly.

The first HPRC data release comprises the sequencing data from 47 participants, mostly from the 1KGP (listed and described in Supplementary Table 1). All sequencing data are publicly available and can be downloaded without egress fees from the Amazon Web Services (AWS) Public Datasets program and can be analysed with the AWS cloud. Data are also available for analysis within the AnVIL (Analysis, Visualization and Informatics Lab-space) cloud platform, organized as a public workspace (https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_HPRC). AnVIL is the genomic data science analysis, visualization, and informatics lab-space of the NHGRI that provides a cloud environment for analysis of large genomic datasets, and supports multiple globally used analysis tools including Terra, Bioconductor, Jupyter and Galaxy[46].

Pangenome reference

We are building a pangenome reference with three complementary parts: (1) the haplotypes, which are the sequences within the input assemblies; (2) the pangenome alignment, which is a sequence graph and an efficient embedding of each of the input haplotypes as paths within this graph; and (3) the coordinate system, which is a backward-compatible coordinate system and set of sequences that make it possible to refer to all variations encoded within the reference equally (Fig. 3). The haplotypes provide hundreds of individual representations of the genome, spanning global diversity. Each haplotype assembly will be useful individually as a reference for studying genomic sequences that are divergent from the current human reference assembly. The pangenome alignment represents the homology relationships among the individual assemblies. This canonical alignment will support coordinate translation (liftOver) between the haplotypes and defines the allelic relationships. It will form the substrate for many emerging pangenome tools and pipelines that will improve important genomic workflows, for example, by making genotyping accuracy less dependent on ancestry. The coordinate system provides a global, unambiguous means to refer to all the variations within the pangenome. It makes all the variations within the haplotypes first-class objects that can be referred to equally. Ultimately, it will provide a more complete means

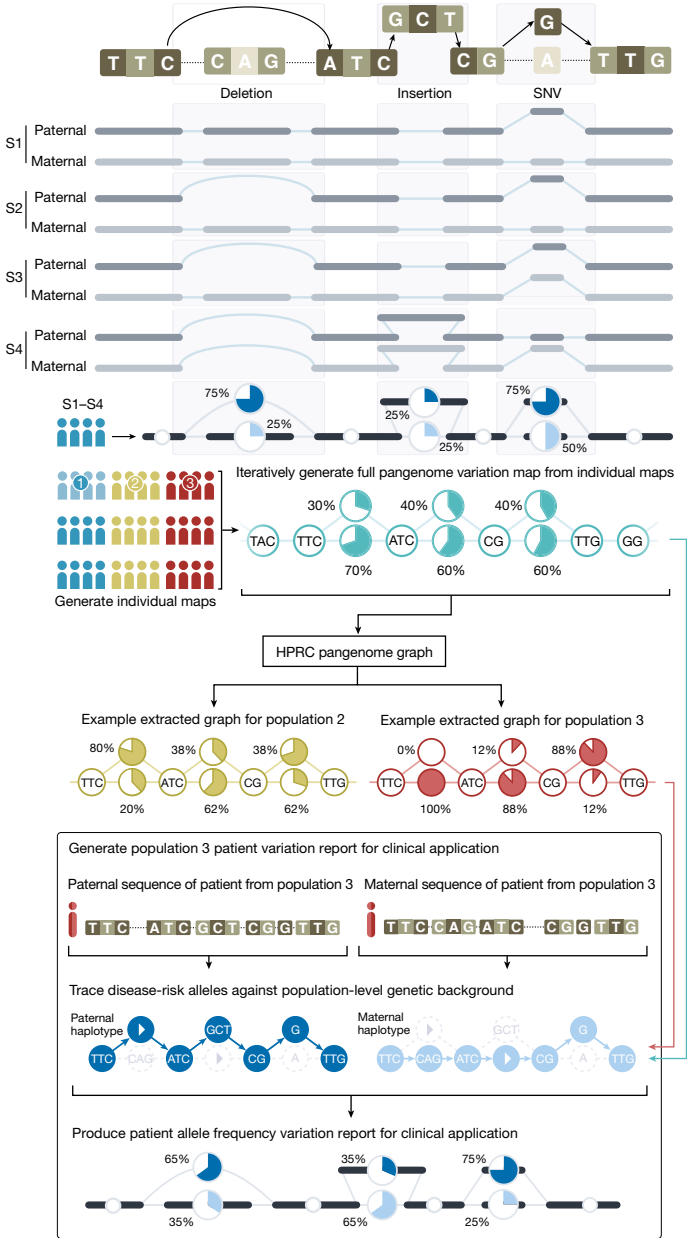


Fig. 3 | The human pangenome reference. Graph-aware mappers can be used to genotype samples by directly mapping against the graph. This simplified example shows how to create a pangenome graph for four people and calculate the allele frequency of three variants. Iterating through each individual produces the structure of the graph, which improves as new genomes are added. Genomic data are arranged into a sequence variation map based on edges. Alternative haplotypes are depicted as alternate pathways across the graph, with the edges being the primary data-bearing elements. The pangenome reference catalogues genomic variation and allows for population-scale analysis because of its graph structure. Tracing a path through the network and connecting sequences at access edges yield haplotypes for individuals. For clinical interpretation, allele frequencies are reported. SNV, single-nucleotide variant.

to refer to variations not contained within the existing linear reference, proving useful for databases and tooling that will build on the pangenome reference.

Supporting these parts is a new proposed set of file standards[47], notably, the rGFA format for representing a pangenome and the GAF format for representing read mappings to a pangenome. We hope that these will have an effect on the field similar to how SAM/BAM[48] and VCF[49] formats have generated a broad range of interoperable tools that have become widely used and accessible. To start this process, we have developed the vg toolkit[50] and minigraph[47], which incorporate downstream tools for graph construction and long-read and short-read mapping and genotyping.

We anticipate releasing an alpha pangenome reference based on existing variant calls and assembled contig genomes. Using the proposed incremental coordinate system, we will subsequently release updated graphs that incorporate the growing numbers of assemblies.

Variant detection

A central aim of this research is to document the genetic similarities and differences among the human genomes included in the pangenome reference. Comprehensive variant detection, however, is still a challenge even when high-quality genome assemblies are available. No single data type or bioinformatic approach yet achieves high performance across all variant classes and genomic regions[51,52]. Therefore, we are pursuing multiple complementary approaches to variant detection using a combination of whole-genome multiple assembly alignment, pairwise assembly–assembly alignment and traditional reference-based read alignment.

Ideally, we will accomplish variant detection in a single step that is designed to build pangenome graphs directly from whole-genome, multiple assembly alignments. Genetic variants will be represented naturally as features in the resulting graph because any variant would be captured by the assembly process. This offers a substantial advantage, enabling optimal breakpoint reconstruction via joint analysis of all input genomes. Accurate multiple alignment and graph construction of entire human genomes is extremely challenging, but recent improvements to tools such as minimap2 (ref. [53]), minigraph[47], cactus[54] and pggb[55] make this feasible. However, errors in variant calling can still arise from errors in assembly and sequence alignment, especially in repetitive regions of the genome. Given this, and the fact that pangenome graph construction tools have not been thoroughly evaluated at scale with real-world data, we are also pursuing the complementary approaches described below.

An alternative approach to multiple alignments is to map variants from pairwise assembly–assembly alignments. Towards this end, we are using minimap2 and Winnowmap to align each draft assembly to the GRCh38 and T2T-CHM13 references to perform variant detection of single-nucleotide variants, indels and other structural variants. This approach is more straightforward than whole-genome multiple alignment; however, complications can arise from reference genome effects and the need to merge results across many pairwise comparisons. The exact coordinates of complex and repetitive variants may differ due to alignment ambiguity. To alleviate reference effects, we are mapping variants via pairwise alignment of the two haploid genomes of each individual, enabling detection of natural heterozygous variants within sequences that are missing or poorly represented in GRCh38 and CHM13. Methods of pairwise alignment assembly help to control for potential errors from the multiple alignment and graph construction process outlined above; however, they still fail to detect variants that are not captured in the underlying assemblies. We are also running a host of traditional variant callers that rely on the alignment of raw reads to the GRCh38 and CHM13 references to control for potential assembly errors. Although limited by reference genome quality and alignment accuracy, these traditional tools are able to capture a subset of variants that are not accurately assembled, and they will serve as a cross-check on newer and less mature assembly-based tools.

In summary, we expect the above methods to capture most genetic variants in genomic regions that are accessible to current assembly and alignment methods. We will compare our variant calls to published call sets from the 1KGP (https://www.international-genome.org), HGSVC (Human Genome Structural Variation Consortium,

Box 4

Pangenome graph tools

Graph building
- minigraph[47]
- PGGB[55]

Graph aligners
- deBGA[82]
- BGREAT[83]
- BrownieAligner[84]
- GenomeMapper[85]
- HISAT2 (ref. [86])
- VG[87]
- GraphAligner[88]
- GRAF (https://www.internationalgenome.org/human-genome-structural-variation-consortium)
- PaSGAL[89]
- SPAligner[90]

Graph indexing
- CHOP[91]
- PSI[92]

Graph visualization
- Bandage[93]
- GfaViz[94]
- SGTK[95]
- AGB[96]
- Sequence Tube Map[97]
- MoMI-G[98]
- VG view[87]
- VG vis[87]
- ODGI viz[99]

Gene prediction
- Path Racer[100]

Variant detection
- PanGenie[101]
- Cortex with Bubbleparse[102]
- BayesTyper[103]
- Paragraph[104]
- GraphTyper2 (ref. [105])
- VG[87]

https://www.internationalgenome.org/human-genome-structural-variation-consortium) and GIAB[42] using samples from these projects that are also included in the HPRC references to assess quality. We will evaluate and validate variant calls using independent data types generated by the HPRC but that are not used for contig assembly—such as ONT, Hi-C, Strand-seq and BioNano data—and assess the read-level support for each variant call based on the alignment of raw data to assemblies and pangenome graphs.

Achieving comprehensive T2T variant detection across the entire genome will require improved methods for genome assembly, multiple alignments and graph construction. The development and application of these methods in subsequent years is a major goal of the HPRC, and will help to extend the impact of pangenomics to the full spectrum of variant classes.

Pangenome annotation

Annotation of the current GRCh38 reference includes genes and genomic features, such as repeats, CpG islands, regulatory regions and chromatin immunoprecipitation–seq peaks, among others. The pangenome reference will have these same utilities and more, including the following.

For genes, the two primarily used gene sets in genomic analysis are the National Center for Biotechnology Information's (NCBI's) RefSeq[56], which exists as independent mRNA definitions, and EMBL-EBI's Ensembl/GENCODE[57], which is built on the GRCh38 reference. The pangenome reference will support both the RefSeq and the Ensembl/GENCODE gene sets. We will map both annotations to each haplotype. Specifically, we will evaluate the mapping of the core reference set of human transcriptome data to each haplotype and incorporate putative new genes that are not represented in either RefSeq or Ensembl/GENCODE. Mapping these gene sets in conjunction with other transcriptomic datasets will annotate the pangenome graph. Other tools will support spliced alignments and transcript reconstruction on a mature graphical data structure. We will integrate the results of these approaches into the annotation released for each haplotype, accompanied by a description of whether transcripts are identical by both methods or whether changes were identified, including transcripts that are disabled, duplicated or missing on a given haplotype. We will also annotate all transcript haplotypes for their global frequency using Haplosaurus tools[58]. We will initially annotate haplotype-by-haplotype, and will explore methods for direct annotation of the pangenome, such as those currently being developed in the GENCODE Consortium. Direct annotation methods simultaneously cover all relevant haplotypes and result in both an annotated genome graph and haplotype-specific annotations. One of the critical use cases of direct annotation of the pangenome will be large transcriptomic datasets aligned directly to the graphical structure that natively annotate it.

For functional elements and other genome features, a central goal in biology is to understand how sequence variants affect genome function to influence phenotypes. Genome function includes regulatory regions that influence gene expression, enhancers that modulate expression levels, and the three-dimensional interactions that control chromosome structural organization within a cell. We will use the pangenome reference to annotate such functional information using existing RNA sequencing, methylC sequencing and assay for transposase-accessible chromatin with high-throughput sequencing datasets from Roadmap Epigenomics, ENCODE, 4D Nucleome (4DN), Genotype-Tissue Expression (GTEx) and the Center for Common Disease Genomics (CCDG), among others. This will enhance the functional human genetic variation catalogue.

Integrating functional data with the pangenome reference will facilitate the development of toolkits and analysis pipelines that evaluate the effect of genetic variants on complex traits and variation in phenotypes. The HPRC will work with developers to define rules and mechanisms to engage with multimodal 'big bio-data' for both data providers and consumers. We will co-create user-friendly informatics platforms to manage, integrate, visualize and compare highly heterogeneous datasets in the context of the genetic diversity represented in the pangenome. Box 4 lists available resources for working with pangenome graphs. We will also make all haplotype-by-haplotype annotation methods available in AnVIL so that others can run them to create custom annotation tracks on all or a selected subset of assemblies. These platforms will serve as a foundation for significant clinical datasets and global biobank initiatives that will ultimately improve precision medicine and medical breakthroughs. For example, the NHGRI is establishing the Impact of Genomic Variation on Function (IGVF) Consortium, which aims to develop a framework for systematically understanding the effects of genomic variation on genome function. Data generated by the IGVF will include high-resolution identification and annotation of functional elements and cell-type-specific perturbation studies to assess the effect of genomic variants on function. The pangenome will be an important foundation for predicting functional outcomes in these studies.

Data sharing

To enhance community access and sharing, we will submit sequence data (PacBio HiFi, ONT and Hi-C, among others), assemblies and pangenomes produced by the consortium to AnVIL[46] and the International Nucleotide Sequence Database Collaboration (INSDC)[59]. Data will also be stored and made publicly available on both S3 and Google Cloud Storage. This general model supports future efforts to use cloud-based strategies for biological data analysis that spans multiple centres. Users of various clouds worldwide will know that they are using the same datasets. Data coordination within the consortium will leverage the established methods in use and the constant development since the inception of the 1KGP more than a decade ago[60,61]. These processes will ensure that we rapidly release data in an organized manner, with proper accessioning of archival datasets, and future traceability of analysis objects and primary data items. Data stored in the INSDC will use BioProjects and BioProject umbrella structures similar to the 1KGP and the Vertebrate Genome Project[8] to ensure that data are appropriately organized and easily identifiable in the public archives. This approach ensures that sample identifiers are effectively managed via the BioSamples database[62], including metadata provisions, and makes any data generated from the same samples readily tractable. The INSDC will archive all reads and assembly data, and other relevant archives will be used, as appropriate for a specific data type. Each haplotype assembly will receive a genome collections accession number (GCA_*), which we will version as we make assembly updates. We will address additional data sharing considerations as they arise through our expanded recruitment and sampling efforts to broaden the diverse representation of global variation.

Adoption and outreach

Achieving widespread international adoption of a pangenome reference will be a challenge[11]. The HPRC will design a pragmatic model and transition plan that are simple and compelling enough to gain traction among researchers and clinical laboratories. Working across scientific and other stakeholder communities, we will foster a new ecosystem of analysis tools. We will maintain and improve the reference, establish scalable bioinformatics methods for resolving errors, improve resolution in difficult-to-resolve genomic regions and respond to user feedback. Importantly, we envision an integrated pangenome transition plan that involves broad community engagement via outreach and education, from tool developers to end-users. These efforts will create a software ecosystem and expert user base to support the next generation of human genetics. The pangenome reference will provide improved genomic research standards, data sharing and reproducible cloud-based workflows. Understanding the barriers to adoption will lead to effective outreach and training, ensuring that the pangenome reference resource is widely adopted.

Adoption will ultimately be driven by the creation of a data resource that sustains continued improvement in its accuracy and completeness, enables a range of uses and improves genomic analyses. We will actively publicize the benefits of using the pangenome. As a starting point for our outreach efforts, we have created a website (https://humanpangenome.org) to publicize the consortium. We have also created social media accounts for the human pangenome that directly connect our consortium with the end-user community (for example, @HumanPangenome on Twitter).

To facilitate adoption, we will explore who the user community will be, their needs and, most importantly, the technical and non-technical barriers that they may encounter. Addressing potential obstacles is essential, as we know that adopting an updated version of the linear reference can result in significant bottlenecks for many laboratories. The cost of switching can be significant, and the HPRC is aware that many clinical laboratories worldwide still use the GRCh37 build from February 2009 for this reason. The HPRC will examine how to reduce switching costs and expedite transition. User data will be collected in self-reporting surveys, including user characteristics, location, specific applications and barriers to adopting a pangenome reference framework.

Creating a coordinate system that builds on GRCh38 and includes both GRCh37 and GRCh38 assemblies is central to user adoption. The HPRC will develop training materials that explain the additional sequences included in the pangenome reference coordinates and how these sequences relate to GRCh37 or GRCh38. Existing linear reference tools will continue to work with the expanded pangenome reference coordinate system, and pangenome-based results will be translatable to these existing coordinate systems with improved genotype accuracy.

We will develop liftOver tools that make it easy to go backward from the pangenome reference to GRCh37 or GRCh38 when necessary. We already have algorithms for this purpose and demonstration of functionality to predict read mappings from a prototype pangenome to GRCh37 or GRCh38. We will precompute all mappings between the previous assemblies and the pangenome and provide these coordinate translation functions with the pangenome reference release. This information should ease the transition of other databases and resources that rely on these coordinates and provide an annotation directly onto the GRCh37 or GRCh38 assemblies in areas where mappings and interpretation on the pangenome are more reliable than current linear sequence representations.

We will augment the displays of the human genome browser to transition to any haplotype assembly in the pangenome reference and display the haplotype alignments. Visualizations will include relevant genetic backgrounds for specific tracks, for example, picking the right HLA haplotype for a read mapping track. To ensure that we use these tools effectively, we will add detailed information that explains these novel views to our existing training materials and make this information part of our respective workshops.

We have adopted the GA4GH principles and will develop exchange formats analogous to SAM/BAM and utility libraries analogous to htslib/samtools, facilitating the development of transition tools and workflows for the pangenome reference. We will deposit these tools and their guides in the HPRC resource repository. We have also developed a prototype transcript archive that facilitates annotation discovery in GRCh37, GRCh38, CHM13 and the pangenome, and visualizes the differences between two transcripts (for example, on two different genomes).

We aim to engage pilot users to obtain feedback about these resources. The HPRC programme and related tool developers connected with the community of users will develop new tools that gain additional value from using the pangenome reference rather than linear reference genome assemblies. We will report on our discoveries in publications and talks, through the blog, webinars and on the HPRC website, and provide educational tools and forums on using and switching to a pangenome reference.

Relevance to disease research

We expect that the resources and methods that we are developing will profoundly impact studies of the genetic basis of human disease and precision medicine. Although we recognize that adoption by the clinical research community will take time, there are three important benefits to using a pangenome reference. First, a more complete reference that incorporates and displays human genetic diversity will produce fewer ambiguous mappings and more accurate analyses of copy number variation throughout the genome when patient samples are sequenced and analysed[63,64]. This will improve genetic diagnosis and the functional annotation of variants. Second, the resource will enable the discovery of disease-risk alleles and previously unobserved rare variants, especially in regions that are inaccessible to standard, short-read sequencing technologies. Studies of unsolved Mendelian genetic disease, for example, have shown that approximately 25% of 'missing' disease variants

can be recovered when longer reads are applied and more complex repetitive regions are characterized[65]. Important genetic risk loci, such as *SMN1* and *SMN2* (spinal muscular atrophy), *LPA* (lipoprotein A and coronary heart disease), *CYP2D6* (pharmacogenomics), as well as numerous triplet repeat expansion loci are now being sequenced and assembled in large human cohort studies. These studies are revealing the standing pattern of natural genetic variation for loci that are typically excluded from previous analyses[51,63]. Resolution of these loci by long-read sequencing in even a limited number of human haplotypes improves our ability to genotype them in other patient-derived short-read datasets, allowing for the discovery of new genetic associations, through both genome-wide association study and expression quantitative trait locus methods[51]. Last, the pangenome approach represents a fundamental change in how human genetic variation is discovered. Instead of simply mapping sequence reads to a reference, we are constructing phased genome assemblies and aligning them to the graph, which in turn will pinpoint all genetic differences, both large and small, at the base-pair level[26,66]. As long-read sequencing costs fall and pangenome methods evolve[26], we predict that patient samples will probably be sequenced using long-read technology to increase sensitivity and accuracy.

Outlook

As we write this Perspective article, the world is reeling from the COVID-19 pandemic and the spread of new SARS-CoV-2 variants. Scientists can trace the epidemiology of the virus, determine why humans are susceptible[67,68] and determine why some individuals are more susceptible than others[69,70]. The current GRCh38 human reference is one of many resources that have made this possible, but we know that it can be improved. Through years of strategic investments in the public and private sectors, we find ourselves with the technologies and methods to build additional references that better represent global human genomic diversity.

The human pangenome reference will collect accurate haplotype-phased genome assemblies generated by efficient algorithmic innovations, which we anticipate will be widely used by the scientific community. The collection of individual genomes, comprising sequence information, genomic coordinates and annotations, will be a critical resource with more accurate representation of human genomic diversity. The original Human Genome Project enabled major advances in human health and genomic medicine[1–4]; it is time to build a more inclusive resource with better representation of human genomic diversity to better serve humanity.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Gibbs, R. A. The Human Genome Project changed everything. *Nat. Rev. Genet.* **21**, 575–576 (2020).
4. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
5. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
6. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
7. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
8. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
9. Schneider, V. A. 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).
10. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163–165 (2011).
Emphasizes the importance of reference data from ancestral and diverse genomes, as well as stating that researchers should invest time and money into education and outreach to explain why studying global (and local) health is so important.
11. Miga, K. H. & Wang, T. The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).
12. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
13. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
A model for presenting genomes that aims to improve read mapping by representing genetic variation in the reference.
14. Martiniano, R., Garrison, E., Jones, E. R., Manica, A. & Durbin, R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol.* **21**, 250 (2020).
15. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
16. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
17. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
18. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
19. Li, R. et al. Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
20. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
The sequence of the first complete human chromosome.
21. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
22. Nurk, S. et al. The complete sequence of a human genome. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.26.445798 (2021).
The first complete genome assembly issued from the T2T Consortium, which closed all remaining gaps in the GRCh38, including all acrocentric short arms, segmental duplications and human centromeric regions.
23. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
24. Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
25. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **23**, 148–154 (2015).
26. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief Bioinform.* **19**, 118–135 (2018).
27. Eizenga, J. M. et al. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
28. Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
29. Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
30. Popejoy, A. B. et al. The clinical imperative for inclusivity: race, ethnicity, and ancestry (REA) in genomics. *Hum. Mutat.* **39**, 1713–1720 (2018).
31. Popejoy, A. B. et al. Clinical genetics lacks standard definitions and protocols for the collection and use of diversity measures. *Am. J. Hum. Genet.* **107**, 72–82 (2020).
32. Bonham, V. L. et al. Physicians' attitudes toward race, genetics, and clinical medicine. *Genet. Med.* **11**, 279–286 (2009).
33. Race, Ethnicity & Genetics Working Group. The use of racial, ethnic, and ancestral categories in human genetics research. *Am. J. Hum. Genet.* **77**, 519–532 (2005).
34. Dodson, M. & Williamson, R. Indigenous peoples and the morality of the Human Genome Diversity Project. *J. Med. Ethics* **25**, 204–208 (1999).
35. Couzin-Frankel, J. Ethics. DNA returned to tribe, raising questions about consent. *Science* **328**, 558 (2010).
36. Dukepoo, F. C. The trouble with the Human Genome Diversity Project. *Mol. Med. Today* **4**, 242–243 (1998).
37. Fox, K. The illusion of inclusion—the "All of Us" research program and Indigenous peoples' DNA. *N. Engl. J. Med.* **383**, 411–413 (2020).
38. Devaney, S. A., Malerba, L. & Manson, S. M. The "All of Us" program and Indigenous peoples. *N. Engl. J. Med.* **383**, 1892 (2020).
39. Hudson, M. et al. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nat. Rev. Genet.* **21**, 377–384 (2020).
40. Carroll, S. R., Herczog, E., Hudson, M., Russell, K. & Stall, S. Operationalizing the CARE and FAIR principles for Indigenous data futures. *Sci. Data* **8**, 108 (2021).
41. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
42. Genome in a Bottle. *NIST* https://www.nist.gov/programs-projects/genome-bottle (updated 16 February 2022).
43. Jarvis, E. D. et al. Automated assembly of high-quality diploid human reference genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.06.483034 (2021).
44. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with HiFiasm. *Nat. Methods* **18**, 170–175 (2021).
HiFiasm is a haplotype-resolved assembler specifically designed for PacBio HiFi reads that aims to represent haplotype information in a phased assembly graph.
45. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
46. Schatz, M. C. et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom.* **2**, 100085 (2022).
The AnVIL platform provides scalable solutions for genomic data access, analysis and education.
47. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with Minigraph. *Genome Biol.* **21**, 265 (2020).
The Minigraph toolkit has been used to efficiently construct a pangenome graph, which is useful for mapping and constructing graphs that encode structural variation.
48. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

49. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
50. Rosen, Y., Eizenga, J. & Paten, B. Modelling haplotypes with respect to reference cohort variation graphs. *Bioinformatics* **33**, i118–i123 (2017).
51. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
The use of long-read data from 64 human genomes to predict structural variants and the patterns of variation across diverse populations.
52. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
54. Paten, B. et al. Cactus: algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
Cactus is a highly accurate, reference-free multiple genome alignment program that is useful for studying general rearrangement and copy number variation.
55. Pangenome Graph Builder. *GitHub* https://github.com/pangenome/pggb (2022).
56. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
57. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
58. Spooner, W. et al. Haplosaurus computes protein haplotypes for use in precision drug design. *Nat. Commun.* **9**, 4128 (2018).
59. Arita, M., Karsch-Mizrachi, I. & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
60. Clarke, L. et al. The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).
61. Clarke, L. et al. The International Genome Sample Resource (IGSR): a worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res.* **45**, D854–D859 (2017).
62. Courtot, M. et al. BioSamples database: an updated sample metadata hub. *Nucleic Acids Res.* **47**, D1172–D1178 (2019).
63. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.26.445678 (2021).
64. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.12.452063 (2021).
The importance of complete T2T genomes in novel variant discovery and of offering major improvements of variant calls within clinically relevant genes are highlighted.
65. Miller, D. E. et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).
66. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
67. Kim, D. et al. The architecture of SARS-CoV-2 transcriptome. *Cell* **181**, 914–921.e90 (2020).
68. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
69. Toh, C. & Brody, J. P. Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. *Hum. Genomics* **14**, 36 (2020).
70. Zeberg, H. & Paabo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).
71. Okubo, K., Sugawara, H., Gojobori, T. & Tateno, Y. DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res.* **34**, D6–D9 (2006).
72. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
73. Navarro Gonzalez, J. et al. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
74. Stalker, J. et al. The Ensembl web site: mechanics of a genome browser. *Genome Res.* **14**, 951–955 (2004).
75. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
76. Zhou, X. et al. The Human Epigenome Browser at Washington University. *Nat. Methods* **8**, 989–990 (2011).
77. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU Epigenome Browser update 2019. *Nucleic Acids Res.* **47**, W158–W165 (2019).
78. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
Analysis of sample descriptions included in the genome-wide association study catalogue indicates that some populations are still under-represented and left behind in studies of genomic medicine.
79. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, 9 (2019).
80. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
81. Ulahannan, N. et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. Preprint at *bioRxiv* https://doi.org/10.1101/833590 (2019).
82. Liu, B., Guo, H., Brudno, M. & Wang, Y. deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinformatics* **32**, 3224–3232 (2016).
83. Limasset, A., Cazaux, B., Rivals, E. & Peterlongo, P. Read mapping on de Bruijn graphs. *BMC Bioinformatics*. **17**, 237 (2016).
84. Heydari, M., Miclotte, G., Van de Peer, Y. & Fostier, J. BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs. *BMC Bioinformatics* **19**, 311 (2018).
85. 1001 Genomes. GenomeMapper. *1001 Genomes* https://www.1001genomes.org/software/genomemapper_graph.html (accessed 2021).
86. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
87. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
88. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
89. Jain, C., Misra, S., Zhang, H., Dilthey, A. & Aluru, S. Accelerating sequence alignment to graphs. *IEEE Int. Parallel and Distributed Processing Symp. (IPDPS)* 451–461 (2019).
90. Dvorkina, T., Antipov, D., Korobeynikov, A. & Nurk, S. SPAligner: alignment of long diverged molecular sequences to assembly graphs. *BMC Bioinformatics* **21**, 306 (2020).
91. Mokveld, T., Linthorst, J., Al-Ars, Z., Holstege, H. & Reinders, M. CHOP: haplotype-aware path indexing in population graphs. *Genome Biol.* **21**, 65 (2020).
92. Ghaffaari, A. & Marschall, T. Fully-sensitive seed finding in sequence graphs using a hybrid index. *Bioinformatics* **35**, i81–i89 (2019).
93. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
94. Gonnella, G., Niehus, N. & Kurtz, S. GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics* **35**, 2853–2855 (2019).
95. Kunyavskaya, O. & Prjibelski, A. D. SGTK: a toolkit for visualization and assessment of scaffold graphs. *Bioinformatics* **35**, 2303–2305 (2019).
96. Mikheenko, A. & Kolmogorov, M. Assembly Graph Browser: interactive visualization of assembly graphs. *Bioinformatics* **35**, 3476–3478 (2019).
97. Beyer, W. et al. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics* **35**, 5318–5320 (2019).
98. Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y. & Kasahara, M. MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformatics* **20**, 548 (2019).
99. ODGI. *GitHub* https://github.com/pangenome/odgi (2021).
100. Shlemov, A. & Korobeynikov, A. in *Algorithms for Computational Biology* (eds Holmes, I., Martín-Vide, C. & Vega-Rodríguez, M. A.) 80–94 (Springer, 2019).
101. Ebler, J. et al. Pangenome-based genome inference. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.11.378133 (2020).
102. Leggett, R. M. et al. Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de Bruijn graphs. *PLoS ONE* **8**, e60058 (2013).
103. Sibbesen, J. A. et al. Accurate genotyping across variant classes and lengths using variant graphs. *Nat. Genet.* **50**, 1054–1059 (2018).
104. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
105. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).

Acknowledgements We thank the NHGRI for funding multiple components to improve and update the Human Genome Reference Program, which has supported the work represented by this report (1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973 and 1U01HG010963). This work was also supported, in part, by the Intramural Research Program of the NHGRI, NIH (A.M.P.).

Author contributions All authors contributed to writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04601-8.

Correspondence and requests for materials should be addressed to Ting Wang, Paul Flicek, Heng Li, Karen H. Miga, Benedict Paten, Erich D. Jarvis, Ira M. Hall, Evan E. Eichler or David Haussler.

Peer review information *Nature* thanks Kazuto Kato and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

POLICY FORUM

GENETICS AND SOCIETY

Getting genetic ancestry right for science and society

We must embrace a multidimensional, continuous view of ancestry and move away from continental ancestry categories

By **Anna C. F. Lewis, Santiago J. Molina, Paul S. Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M. Fullerton, Nanibaa' A. Garrison, Nayanika Ghosh, Evelynn M. Hammonds, David S. Jones, Eimear E. Kenny, Peter Kraft, Sandra S.-J. Lee, Madelyn Mauro, John Novembre, Aaron Panofsky, Mashaal Sohail, Benjamin M. Neale, Danielle S. Allen**

Glaring health disparities have reinvigorated debate about the relevance of race to health, including how race should and should not be used as a variable in research and biomedicine (*1*). After a long history of race being treated as a biological variable, there is now broad agreement that racial classifications are a product of historically contingent social, economic, and political processes. Many institutions have thus been reexamining their use of race and racism and stating intentions about how race should be used going forward. One common proposal is to use genetic concepts—in particular, genetic ancestry and population categories—as a replacement for race (*2*). However, the use of ancestry categories has technical limitations, fails to adequately capture human genetic diversity and demographic history, and risks retaining one of the most problematic aspects of race—an essentialist link to biology—by allowing genetic ancestry categories to stand in its place.

The process of racialization entails a dynamic cognitive process of identification based on phenotype that is often highly context dependent. Although research has found genetic variation correlated with phenotypes that have been historically used to assign race categories, such as skin pigmentation or hair texture, it is the case that such genetic correlates are not distributed in a manner that correspond to racially defined groups. Race is a sociopolitical construct rather than a biological one. For example, in the United States, immigrants from southern and eastern Europe only began to be classified as "white" on the census in the 20th century (*3*); the American Indian/Alaska Native census category reflects colonizing histories and federal policies (*4*). As such, social scientists and others have argued that the strongest case for using race is limited to tracking the impact of racism on health outcomes, rather than as a proxy for anything biological (*5*).

Genetic ancestry, one of the main proposed alternatives to using race, is of relevance to statistical and population geneticists, epidemiologists, public health practitioners, physicians, and patients. In particular, genetic ancestry has renewed relevance for the clinical application of genetic technology because the accuracy of genetic risk scores varies across ancestries (*6*). Genetic ancestry and population categories are also relevant to the general public, as demonstrated by the tens of millions of individuals who have paid for ancestry reports from consumer companies. Across these different domains, a dominant description of genetic ancestry is associated with continents as meaningful groupings. Within genetics research, continental ancestry categories have become the most common type of group label (*7*). Similarly, consumer genetics products give customers a report with data based on a percentage of these continental groups from which an individual can trace their "ancestry."

Systems of racial classification have historically regarded continents as meaningful group boundaries; thus, it is not surprising that racial categories and continental ancestry categories are often confounded. Whenever continental ancestry categories are used, the risk is high that a misconception of race as a biological attribute will reenter through the back door (*8*). Insufficiently nuanced thinking about continental categories, genetic ancestry, and racial groups can lead to the conflation of the three.

A FLATTENED NOTION OF ANCESTRY

Our genetic ancestry is defined by the stretches of the genome that we inherit from our ancestors (*9*). Geneticists have a concept for this known as the ancestral recombination graph (ARG). Put simply, an individual's genetic ancestry is the subset of paths through the human family tree by which they have inherited DNA from specific ancestors. Most often, geneticists study the ARG of multiple individuals at the same time.

Crucially, this definition makes clear that there are two things that are not necessary to the definition of genetic ancestry. The first is any categorization by populations or groups. And the second is any contextualization of the individuals apart from their genealogical connections—for example, by labeling these individuals with geographical or cultural information. Yet current practices around ancestry estimation and reporting almost always impose categories and, when they do so, very often default to just one way to contextualize individuals: by continent of origin. Both practices limit the accuracy and reliability of claims being made by researchers about human genetic difference.

There are many statistical methodologies across subfields of genetics and genomics whose outputs are framed as "genetic ancestry," most of which do not attempt to approximate the ARG and several of which only capture genetic similarity (*9*). The majority of these methods involve placing individuals into categories or modeling them as mixtures of discrete categories. For some methods, the categories are predefined and prelabeled. For others, the categories emerge from the analysis. In these cases, not only are the resulting categories very sensitive to which individuals are included in the analysis, they may not even represent shared ancestries (*10*). In other cases, categories and their labels are imposed in downstream analysis.

The concern about use of categories goes beyond these technical limitations. Imposing categories on genetic ancestry fails to adequately capture human genetic diversity and what we know of human demographic history. A standard way to visualize patterns of genetic similarity is by plotting results of principal components analysis of genetic variation data, a technique that reduces the dimensionality of that data. Most genetic analyses use data from reference populations to contextualize a study's data. The most commonly used reference data were created by sampling individuals from a few dozen places spread across the globe. If individuals from these populations are graphed in this manner, distinct clusters roughly representing continental categories are visible (see the figure). A prominent early result was that genetic ancestry was strongly concordant with continental origins when ascertaining for individuals

Author affiliations are available in the supplementary materials. Email: annalewis@g.harvard.edu

whose four grandparents were from the recruitment sites (*11*).

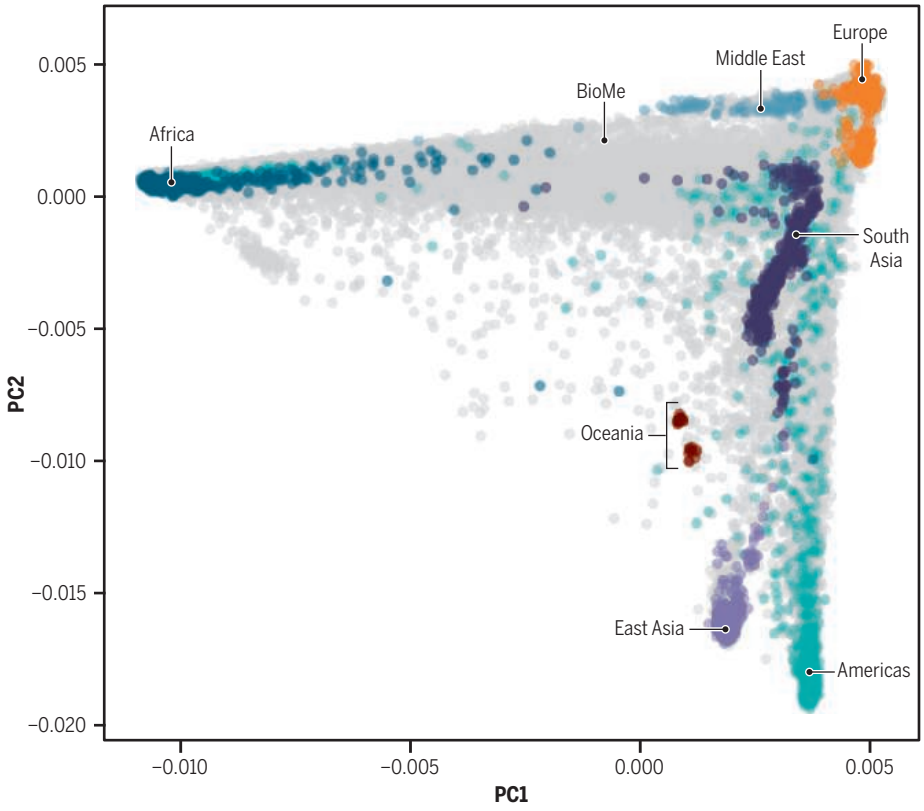
But newly assembled datasets show that if people are sampled differently, such as individuals living in New York City, it becomes clear how impoverished this view of a structure of distinct clusters is (see the figure) (*12*). The clearly separated clusters of reference population individuals, corresponding to different continental groups, merge into a background of continuous genetic variation. This is consistent with what we know of human demographic history, in which mass migration and constant mixing across groups have been the norm. The impact of these histories leads to different structures of genetic variation in different parts of the world. Such studies illustrate just how inappropriate use of discrete continental categories can be, particularly when information framed as genetic ancestry can potentially influence medical care.

The use of the terms admixture and "admixed individuals"—defined as those who have recent ancestry from more than one population, and typically continental ancestry populations—reinforces notions of discrete categories within humanity. This use does not escape the notion of continental ancestry categories but rather compounds the errors of using such categories because these individuals are typically conceptualized as a mixture of otherwise "pure" continental ancestry populations.

Our conceptualization of ancestry must be general enough to describe every human; the only way to do this is to use concepts and tools that acknowledge that ancestry is continuous. Categories have their legitimate uses—for example, in reporting the differences in predictive power of genetic risk scores (even in this case, differences in performance are due to many factors, and focusing on only one factor such as ancestry can lead to essentializing differences between groups) (*6*). But the default appeal to any one set of categories risks essentializing those groups, making it more likely that differences between these abstract groups are treated as though they were concrete.

The continuous, category-free, nature of genetic variation

Colored dots (n = 4149) are reference panel individuals from 87 populations representing ancestry from seven continental or subcontinental regions projected onto the first two principal components (PC1 and PC2) of genetic similarity. Gray dots (n = 31,705) are participants from BioMe, a diverse biobank based in New York City. Clearly delineated continental ancestry categories (the islands of color) are shown to be a by-product of sampling strategy. They are not reflective of the diversity in this real-world dataset, which is made evident by the continuous sea of gray.



GRAPHIC: K. FRANKLIN/*SCIENCE* BASED ON (*12*)

In addition to not requiring the use of categories, the definition of genetic ancestry is silent on any aspect of the context of an individual's ancestors. Although the ancestral recombination graph does have structure, it does not by itself indicate anything about an individual's geographical location or their culture. Researchers face choices in whether and how to provide this context. Crucially, we can give multiple contexts depending on the time horizon considered because we each have ancestors from every generation in our species' past. Advances in ancient DNA and in population genetics are providing us with more and more information about population structure at different points in our histories. A contemporary human genome can hence increasingly give us visibility into the chronologically layered ancestral record for that person.

Yet this historical notion of genetic ancestry is flattened when just one set of categories is used. In the case of continental ancestry categories, their use reflects the assumption that at some specific point in time, humans were mostly divided into homogeneous groups by the natural geographical barriers between continents. This is a gross oversimplification of human history. It also obscures other time slices when different categories would be relevant—for example, ~50,000 years ago, *Homo Sapiens* and Neanderthal categories; or ~5000 years ago, "Steppe-related," "European" hunter-gatherer, and "Near Eastern" farmer categories in Europe (*13*); or ~500 years ago, when waves of migration and the slave trade were forging new patterns of human genetic diversity in the Americas.

A MORE COMPLEX NOTION OF ANCESTRY

What are the implications for researchers who want to invoke genetic ancestry? They should first ask whether they need to impose categories at all to answer their research question. There are many situations in which categorization has been thought essential but has subsequently been shown to be avoidable, such as in correcting for population stratification in genome-wide association studies (*14*). In cases in which genetic ancestry categories can be avoided, they should be avoided. If researchers are able to justify a scientific need to impose categories, they should next think about whether they have to provide labels (be it geographic, ethnic, linguistic, or other) to the groupings they impose. If they do need to provide labels, they should give the scientific justification for that choice and show that they have considered potential disadvantages of imposing these labels.

Additionally, researchers should use multiple types of categories, reflecting that

genetic ancestry is a historical concept: We all have multiple ancestries depending on the time horizon considered. No individual has a single "ancestry"; the plural should always be used. Different geographical resolutions—for example, "Yoruban" versus "West African"—can serve as proxies for different time slices. Ancestry categories from different time points may be of medical relevance. The incorporation of ancient DNA information can also allow for probing different time slices, although the promise of this approach will depend on how much ancient DNA can actually be recovered and analyzed. The use of continental ancestry categories as a proxy for one of the time slices considered must be particularly carefully justified because of the conflation of continental ancestry categories with racial groupings. Additionally, future work should find better ways to conceptualize the genetic ancestry of individuals whose recent ancestors come from distant parts of the ARG.

For some diseases that have a different prevalence in different populations, genetic risk factors may indeed be at play, a result of differences in the chance arrival of new mutations, demographic history, and historical environmental exposures. But although it is possible that genetics is playing a causal role in such cases, genetic ancestry may also be serving as a proxy for differences in environmental effects, including the effects of discrimination. Whenever researchers invoke any categories in understanding health outcomes, they need to make careful efforts to jointly model genetic and environmental effects and acknowledge that a failure to explain differences could be due to unmodeled factors.

Science is reductive, and a model that uses simple continental categories has been useful in starting the process of understanding human genetic diversity. But all models have their legitimate domains of application and limits, and a much more complex set of models should now be the norm across a wide variety of use cases. This is particularly important because although human genetics falls under the biological sciences, it is in fact a science at the intersection of several disciplines, including anthropology, demography, epidemiology, history, and sociology. Even if the limitations of models used are well understood by statistical and population geneticists, others may take the models to be descriptive of realities rather than recognizing that they merely formalize approximations and estimates, using reductive categories to do so. Hence, one of the risks of using these categories is that others may interpret them as true natural kinds, which is inaccurate. Instead, they are heuristics permitting the approximation or answering of very narrow sorts of questions. Because of the association of continental ancestry categories with racial groupings, this is particularly important for continental categories.

An individual researcher's use of continental ancestry categories is not in and of itself racist, but the cumulative impact of this practice has led to and sustains racism. Typological thinking about human difference has had damaging social consequences. Continued reliance on continental ancestry categories contributes to failures of inference, miscommunication between fields, and reported findings that are rooted in reductive and limited ways of understanding human difference. These are likely to exacerbate medical stereotypes about individuals and groups, contribute to health disparities rather than addressing them, and reify (mis)understandings of race as biological. Moreover, this problem is not limited to continental ancestry categories; national categories can and have been reified as biological for political goals (*15*).

The solution will require addressing the issues with how ancestry is conceptualized and used across the entire biomedical research ecosystem. This will involve the development, operationalization, and widespread use of a more complex notion of ancestry—one that disambiguates what is meant by genetic ancestry from related concepts, wherever possible does not treat ancestry as a categorical variable, and treats ancestry as reflecting a historical process, meaning that any study should use many different types of categories.

To aid this transition, a solid empirical understanding of how and why different fields use and operationalize the concept of ancestry is needed. To ensure that this more complex notion of ancestry is then used in practice will require systems-level change. New computational tools and data structures will be required—for example, a wider variety of proxies for genetic ancestry that do not impose categories, as well as easily accessible software tools to enable use of ancestry categories representing multiple time horizons. Further development and adoption of methodologies that directly estimate the ARG should be encouraged. Educational materials will need to be developed for scientists and physicians. Scientists of all stripes who engage in research that uses biological categories for humans should not work in isolation but as part of interdisciplinary teams, ideally including engagement with affected communities. In support of these efforts, journal editors should set standards, professional societies should publish best practices, and funders should carefully consider which research agendas they will support. It is paramount, as these organizations rightly critique the use of race as a biological variable, that use of continental ancestry categories does not become the new default. The US National Academies of Sciences, Engineering, and Medicine recently formed an ad hoc committee, "Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research"; we are hopeful that this represents an opportunity for consideration and consolidation of the points raised here.

Adoption of a more complex notion of ancestry should in turn continue to inform the research agenda in population and statistical genetics and in ancient DNA research. It is in these fields, the home turf of the concept of genetic ancestry, that change in practice may have the largest overall impact. These changes are a prerequisite to any research that looks for connections between genetics and health disparities. More generally, with a more complex notion of ancestry that reflects continuous variation and historical depth, we can start to pave the way for a science that reflects the complex histories of human groups, including the power dynamics among them. ■

REFERENCES AND NOTES

1. D. A. Vyas, L. G. Eisenstein, D. S. Jones, *N. Engl. J. Med.* **383**, 874 (2020).
2. A. Oni-Orisan, Y. Mavura, Y. Banda, T. A. Thornton, R. Sebro, *N. Engl. J. Med.* **384**, 1163 (2021).
3. D. R. Roediger, *Working Toward Whiteness: How America's Immigrants Became White: The Strange Journey from Ellis Island to the Suburbs* (Basic Books, Text is Free of Markings ed., 2005).
4. E. A. Haozous, C. J. Strickland, J. F. Palacios, T. G. A. Solomon, *J. Environ. Public Health* **2014**, e321604 (2014).
5. J. H. Fujimura, T. Duster, R. Rajagopalan, *Soc. Stud. Sci.* **38**, 643 (2008).
6. A. R. Martin *et al.*, *Nat. Genet.* **51**, 584 (2019).
7. A. Panofsky, C. Bliss, *Am. Sociol. Rev.* **82**, 59 (2017).
8. T. Duster, *Backdoor to Eugenics* (Routledge, ed. 2, 2003).
9. I. Mathieson, A. Scally, *PLOS Genet.* **16**, e1008624 (2020).
10. D. J. Lawson, L. van Dorp, D. Falush, *Nat. Commun.* **9**, 3258 (2018).
11. N. A. Rosenberg *et al.*, *Science* **298**, 2381 (2002).
12. G. M. Belbin *et al.*, *Cell* **184**, 2068 (2021).
13. I. Olalde *et al.*, *Nature* **555**, 190 (2018).
14. G. L. Wojcik *et al.*, *Nature* **570**, 514 (2019).
15. W.-C. Sung, in *Asian Biotech: Ethics and Communities of Fate*, A. Ong, N. Chen, Eds. (Duke Univ. Press, 2010), pp. 263–288.

ACKNOWLEDGMENTS

This work was supported by National Institute of Mental Health administrative supplements 5000747-5500001474 to 3R37MH107649-06S1. B.M.N. and D.S.A. contributed equally to this work. A.C.F.L. owns stock in Fabric Genomics. E.E.K. has received personal fees from Regeneron Pharmaceuticals, 23&Me, and Illumina and serves on the advisory boards for Encompass Biosciences and Galateo Bio. B.M.N. is a member of the scientific advisory board at Deep Genomics and RBNC Therapeutics, a member of the scientific advisory committee at Milken, and a consultant for Camp4 Therapeutics and Merck.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm7530

10.1126/science.abm7530

Getting genetic ancestry right for science and society

Anna C. F. LewisSantiago J. MolinaPaul S. AppelbaumBege DaudaAnna Di RienzoAgustin FuentesStephanie M. FullertonNanibaa' A. GarrisonNayanika GhoshEvelynn M. HammondsDavid S. JonesEimear E. KennyPeter KraftSandra S.-J. LeeMadelyn MauroJohn NovembreAaron PanofskyMashaal SohailBenjamin M. NealeDanielle S. Allen

Science, 376 (6590), • DOI: 10.1126/science.abm7530

Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries

Michael E. Weale, PhD[a,1,*], Fernando Riveros-Mckay, PhD[a,1], Saskia Selzam, PhD[a], Priyanka Seth, PhD[a], Rachel Moore, PhD[a], William A. Tarran, PhD[a], Eva Gradovich, MSc[a], Carla Giner-Delgado, PhD[a], Duncan Palmer, PhD[a], Daniel Wells, DPhil[a], Ayden Saffari, PhD[a], R. Michael Sivley, PhD[a], Alexander S. Lachapelle, MD[a], Hannah Wand, MS[b], Shoa L. Clarke, MD, PhD[b], Joshua W. Knowles, MD, PhD[b], Jack W. O'Sullivan, MBBS, DPhil[b], Euan A. Ashley, MBChB, DPhil[b], Gil McVean, PhD[a], Vincent Plagnol, PhD[a,2], and Peter Donnelly, DPhil[a,2]

The American College of Cardiology / American Heart Association pooled cohort equations tool (ASCVD-PCE) is currently recommended to assess 10-year risk for atherosclerotic cardiovascular disease (ASCVD). ASCVD-PCE does not currently include genetic risk factors. Polygenic risk scores (PRSs) have been shown to offer a powerful new approach to measuring genetic risk for common diseases, including ASCVD, and to enhance risk prediction when combined with ASCVD-PCE. Most work to date, including the assessment of tools, has focused on performance in individuals of European ancestries. Here we present evidence for the clinical validation of a new integrated risk tool (IRT), ASCVD-IRT, which combines ASCVD-PCE with PRS to predict 10-year risk of ASCVD across diverse ethnicity and ancestry groups. We demonstrate improved predictive performance of ASCVD-IRT over ASCVD-PCE, not only in individuals of self-reported White ethnicities (net reclassification improvement [NRI]; with 95% confidence interval = 2.7% [1.1 to 4.2]) but also Black / African American / Black Caribbean / Black African (NRI = 2.5% [0.6−4.3]) and South Asian (Indian, Bangladeshi or Pakistani) ethnicities (NRI = 8.7% [3.1 to 14.4]). NRI confidence intervals were wider and included zero for ethnicities with smaller sample sizes, including Hispanic (NRI = 7.5% [−1.4 to 16.5]), but PRS effect sizes in these ethnicities were significant and of comparable size to those seen in individuals of White ethnicities. Comparable results were obtained when individuals were analyzed by genetically inferred ancestry. Together, these results validate the performance of ASCVD-IRT in multiple ethnicities and ancestries, and favor their generalization to all ethnicities and ancestries. **(Am J Cardiol 2021;148:157−164)**

Introduction

Current US guidelines for the primary prevention of cardiovascular disease are based on the quantification of an individual's predicted risk of atherosclerotic cardiovascular disease (ASCVD) over the following 10 years using the ASCVD pooled cohort equations tool (ASCVD-PCE).[1−3] The tool combines information from multiple clinical risk factors including age, sex, blood lipid levels, blood pressure, history of diabetes, smoking or anti-hypertensive treatment, and racial identity (entered as "White," "African American" or "Other," where "Other" is treated algorithmically as "White"). Although genetics is a known risk factor,[4] it is not directly included in ASCVD-PCE (family history of ASCVD is assessed separately, outside of the tool). Polygenic risk scores (PRSs), which combine information across thousands of common genetic variants in the human genome, can be added to the ASCVD-PCE tool, but previous studies have reported variable predictive performance.[5−8] Additionally, these studies focused primarily on individuals with European ancestries, but it is known that the predictive accuracy of a PRS tends to attenuate in individuals with non-European ancestries.[9−11] We therefore undertook a clinical validation study of a new 10-year ASCVD risk prediction tool (ASCVD-IRT) that integrates a PRS for ASCVD with the

[a]Genomics plc, Oxford, UK; and [b]Division of Cardiology, Department of Medicine, Stanford University School of Medicine, Stanford, California. Manuscript received December 28, 2020; revised manuscript received and accepted February 23, 2021.

Conflicts of interests: Peter Donnelly and Gil McVean are partners in Peptide Groove LLP. All other authors declare no competing interests.

[1]These authors contributed equally to the work.

[2]These authors jointly supervised the work.

See page 163 for disclosure information.

*Corresponding author: Tel: +44 (0) 1865 981 600.

E-mail address: mike.weale@genomicsplc.com (M.E. Weale).

https://doi.org/10.1016/j.amjcard.2021.02.032

established ASCVD-PCE tool, paying particular attention to its predictive performance in non-European ancestries and non-White ethnicities.

Methods

Following recent guidelines on the use and reporting of race, ethnicity and ancestry,[12] we clarify that in this study we use the term "ethnicity" to refer to social categories including both race and ethnicity, and we do not distinguish "race" from "ethnicity." In the testing cohorts described below, data on ethnicity were collected from questionnaire, census and other self-identification data, allowing us to infer that ethnicity was self-reported. We use the term "ancestry" or "genetic ancestry" to refer to inferences from genetic data. We note that the concepts of ethnicity and ancestry are correlated but not synonymous.[13] To infer ancestry, we used a method based on principal components derived from genetic data to infer membership to one of 5 high-level ancestry groups conforming to those used by the 1000 Genomes Project[14] (Sub-Saharan African [AFR], Native/Indigenous American [AMR], East Asian [EAS], European [EUR], and South Asian [SAS]; see Supplementary Materials for details, OTHER indicates mixed inferred ancestry). We note that this method is more accurately described as producing "ancestry-like" genetic similarity relationships.[15] We present performance evaluations for both ethnicity and ancestry groups, as the former relate to important social categories while the latter allow us to examine the ancestry-specific PRS attenuation issue[9–11] and its effect on IRT performance.

We tested the performance of the ASCVD-IRT using data from the Atherosclerosis Risk in Communities (ARIC) cohort, the Multi-Ethnic Study of Atherosclerosis (MESA), and UK Biobank (UKB). All individuals in these studies gave informed consent. Legal and ethical approval for our use of ARIC and MESA data is provided by the Western Institutional Review Board (Study Number 1264897, IRB Tracking Number 20192201). See Supplementary Materials for UKB approval information.

All 3 cohorts are prospective, contain participants that were extensively examined at baseline, and have continuing follow-up (via annual phone calls for ARIC and MESA, via linkage to national electronic healthcare and death records for UKB). ARIC comprises over 15,000 adults from predominantly 2 study-defined racial/ethnic groups ("Black" and "White"), from defined populations in 4 sites in the USA, aged 45 to 64 years when recruited between 1987 and 1989.[16] MESA comprises over 6,000 adults from 4 study-defined racial/ethnic groups ("African American," "Chinese American," "Hispanic," and "White/Caucasian"), recruited primarily via phone call invitation to 6 sites in the USA, aged 45 to 84 years and free of cardiovascular disease when recruited between 2000 and 2002.[17,18] UKB comprises over 500,000 adults, recruited via postal invitation to 22 sites in the UK, aged 40 to 69 when recruited between 2006 and 2010.[19,20] We carefully selected 88,666 UKB individuals from multiple ethnicities (labels defined from questionnaire data) for the "IRT testing" subgroup, in order to ensure that the IRT testing subgroup was maximally enriched for incident ASCVD cases from non-European ethnicities and ancestries, and to ensure the independence of these individuals from other UKB subgroups that were used to construct and train the IRT (see Supplementary Materials for further details).

We excluded individuals with cardiovascular disease or on cholesterol lowering medication at baseline, and also those related to others in the cohort at greater than third degree relative level according to the genetic inference method described in Bycroft et al.[20] A separate bridging analysis was performed on individuals free of cardiovascular disease who were on cholesterol lowering medication at baseline, which indicated similar performance for this subgroup (Supplementary Figure 1). White participants in MESA and ARIC were excluded from testing due to sample overlap with GWASs used to construct the ASCVD PRS (MEGASTROKE_EUR[21] and CARDIoGRAMplusC4D[22] respectively - see Supplementary Table 1). Black ARIC individuals were included, but it should be noted that they also formed part of the cohort data used to train the ASCVD-PCE model (outside of this study).[1] It may therefore be expected that the absolute prediction performance of both ASCVD-PCE and ASCVD-IRT is somewhat elevated in this group. However, our validation focused on the comparative performance of these 2 tools, which is not expected to be biased. All individuals in our testing cohorts were selected so as to be independent and unrelated to any individuals used in the training of the PRS or the IRT model.

To define ASCVD cases, we used outcomes that closely matched the ASCVD-PCE tool. In MESA we used "CARDIOVASCULAR DISEASE (CVD), HARD," and in ARIC we used a union of "MI (myocardial infarction), heart attack, or fatal CHD (coronary heart disease) by December 31, 2004" and "Definite or probable ischemic incident stroke by December 31, 2004." The ASCVD definition for UK Biobank is described in Supplementary Materials.

The ASCVD-IRT tool is a function of the score obtained from the currently established ASCVD-PCE tool[1–3] and a PRS for ASCVD, trained from multiple datasets that each represent individuals from multiple ancestry groups and from different geographies. Ten GWAS datasets for different ASCVD subtraits were meta-analyzed to derive the PRS, and an additional 4 cohorts were used to train the PRS effect size. Further details regarding the construction of ASCVD-IRT are provided in Supplementary Materials.

We assessed predictive performance of ASCVD-IRT via relative performance comparisons to ASCVD-PCE, focusing in particular on differences in sensitivity and specificity and on the sum of these differences, also known as the Net Reclassification Improvement.[23] We used ASCVD events in the following 10 years to define cases and we used scores above and below a certain risk threshold to define positive and negative results. This relative performance approach is justified because ASCVD-PCE provides a strong basis for comparison, being the currently established and recommended tool for 10-year ASCVD risk prediction in the US.[1–3] We note that absolute values of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are difficult to interpret on their own. ASCVD-IRT and ASCVD-PCE are not diagnostic tests but risk predictors of an uncertain future event, where even under the best of conditions a "positive"

result has a measured probability of being a noncase and a "negative" result has a measured probability of being a case. For example, we label an individual with a risk of 8% a "positive" result, but they only have an 8% chance of becoming an ASCVD case, assuming the model is correct. But while absolute values are difficult to interpret, relative improvements remain good indicators of performance improvements. In particular, simultaneous improvements in both sensitivity and specificity are strongly indicative of superior test performance.

There is also a logical choice for which threshold to use for binary compartmentalization into "positive" (high-risk) and "negative" (low-risk) results, as required for metrics such as sensitivity and specificity. The same guidelines recommending ASCVD-PCE for risk prediction also lay out recommendations for actionable risk thresholds above which intervention is advised. This threshold is 7.5% risk over 10 years for most individuals, but is reduced to 5% for individuals presenting with additional risk factors, one of which is South Asian ethnicity.[2,3] We therefore applied the same thresholds for our clinical performance assessment, applying the 5% threshold for individuals of South Asian ethnicities or ancestries and 7.5% for all others.

Further details on the statistical methods used to calculate performance metrics and their confidence intervals are provided in Supplementary Materials.

Results

Tables 1 and 2 show the subgroup sample sizes, after exclusions, in each IRT testing cohort. Figure 1 and Supplementary Table 2 summarize the relative performance improvements of ASCVD-IRT over the currently established tool (ASCVD-PCE). When meta-analyzed across all testing cohorts, ASCVD-IRT shows significantly improved performance (as measured by 95% CI) across NRI, sensitivity, and specificity (combined results [with 95% CI]: NRI = 3.0% [1.7 to 4.3]; delta-sensitivity [equivalent to NRI in cases] = 1.5% [0.2 to 2.8]; delta-specificity [equivalent to NRI in noncases] = 1.5% [1.3 to 1.7]). Positive predictive value (PPV) and negative predictive value (NPV) were also significantly improved (Supplementary Table 2). The NRI is also significantly positive in all 3 cohorts, with the largest point estimates seen in the 2 US cohorts (Figure 1a, Supplementary Table 2). The within-cohort point estimates for changes in sensitivity, specificity, log(PPV), log(NPV) and Harrell's C[24] are also all positive, albeit in some cases with 95% confidence intervals that are large and overlap zero (Figure 1b-c, Supplementary Table 2).

We proceeded to investigate relative performance patterns within ethnicity (Figure 2, Supplementary Table 3) and ancestry (Supplementary Figure 2, Supplementary Table 4) groups (groups with >50 ASCVD cases across men and women are shown, metrics for groups with fewer cases can be found in Supplementary Tables 3 to 4). Both figures show similar patterns. The overall NRI remains significantly positive when individuals are meta-analyzed for the 2 largest groups we have data for − those corresponding to White or Black / African American / Black Caribbean / Black African self-reported ethnicities and EUR or AFR genetic ancestries. The significantly positive NRI results for Black / African American / Black Caribbean / Black African ethnicities (combined NRI = 2.5% [0.6 to 4.3]) and AFR ancestry (combined NRI = 2.2% [0.4 to 4.1]) are especially noteworthy, given the reported attenuation of PRS performance in individuals of African genetic ancestries.[9–11] The sample sizes and case numbers for other ethnicities and ancestries are lower, meaning that in some contexts it is not possible to reject the null hypothesis that there is no change in NRI. We note that there are no instances of significantly negative performance, whereas there are several instances of significantly positive NRI performance (for MESA-AMR, UKB "Indian, Bangladeshi, or Pakistani," and UKB-SAS), and point estimates are also generally positive. The changes in sensitivity (equivalent to NRI in cases), specificity (equivalent to NRI in noncases), log(PPV), log (NPV) and Harrell's C are either significantly positive or not significantly different from zero (Figure 2b, c, Supplementary Figure 2b, c, Supplementary Tables 3, 4). We also find that ASCVD-IRT has, across the same large ethnicity and ancestry groups, a larger NRI than that of a tool constructed in the same way as ours, but using an alternative PRS previously shown to have good cross-ancestry performance (the coronary artery disease PRS of Inouye et al,[25,26] Supplementary Figure 3).

Performance metrics for smaller groups (with fewer than 50 cases across men and women) are reported in Supplementary Tables 3 and 4. As the sample sizes are small, it is to be expected that most of the reported CIs in these smaller groups overlap zero. For example, UKB contains a small group of Chinese ethnicities (n = 979, with 6 cases), and MESA contains an even smaller group of Chinese American ethnicities (n = 5, with 1 case). The case numbers in these 2 groups are too low to provide an accurate estimate of performance.

We next proceeded to investigate relative performance patterns in 4 sex-by-age subgroups (Figure 3, Supplementary Table 5). This analysis reiterates patterns previously reported for individuals of European ancestries in UKB and using a different IRT built from a coronary artery disease PRS.[5] The overall NRI performance in 3 of the 4 subgroups is significantly positive, with the strongest performance seen in younger middle-aged men (40 to 54 year old) (NRI = 10.3% [5.7 to 15.0]). The NRI estimates within cohorts are either significantly positive (for ARIC 40 to 54 year old men and women) or not significantly different from zero. It is noteworthy that the largest NRI point estimate in ARIC, comprising individuals self-reporting as Black, is also for younger middle-aged men.

UKB sex-by-age subgroups vary in their sensitivity and specificity patterns (Figure 3b, c, Supplementary Table 5), with younger middle-aged men and women (40 to 54 year old) showing significant increases in sensitivity and smaller but significant decreases in specificity, while older middle-aged men (55 to 69 year old) show the opposite pattern. The 2 US cohorts are more balanced, with no significantly negative performance estimates in any subgroup, while some effects remain significantly positive (delta-specificities for ARIC 40 to 54 year old men and women, ARIC 55 to 69 year old women, and MESA 55 to 69 year old men and women).

Table 1
IRT testing cohort sample numbers (and percentage of cohort) by sex, age at recruitment, case status and self-reported ethnicity

Cohort	Self-reported ethnicity	Age at recruitment	Women		Men	
			Cases	Noncases	Cases	Noncases
ARIC	Black	40-54	47 (2.4%)	762 (39%)	38 (1.9%)	422 (21%)
		55-69	40 (2.0%)	383 (19%)	40 (2.0%)	233 (12%)
MESA	Chinese American	40-54	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
		55-69	0 (0.0%)	1 (0.1%)	1 (0.1%)	3 (0.2%)
		70-79	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	African American	40-54	4 (0.2%)	191 (9.9%)	8 (0.4%)	154 (8.0%)
		55-69	10 (0.5%)	252 (13%)	20 (1.0%)	216 (11%)
		70-79	15 (0.8%)	102 (5.3%)	16 (0.8%)	109 (5.7%)
	Hispanic	40-54	2 (0.1%)	134 (7.0%)	1 (0.1%)	146 (7.6%)
		55-69	10 (0.5%)	159 (8.3%)	25 (1.3%)	186 (9.7%)
		70-79	8 (0.4%)	59 (3.1%)	9 (0.5%)	84 (4.4%)
UKB (IRT testing)	White	40-54	135 (0.2%)	18191 (21%)	308 (0.4%)	14167 (16%)
		55-69	588 (0.7%)	25162 (28%)	1048 (1.2%)	18098 (20%)
	Indian, Bangladeshi or Pakistani	40-54	12 (0.01%)	1211 (1.4%)	53 (0.1%)	1136 (1.3%)
		55-69	17 (0.02%)	652 (0.7%)	54 (0.1%)	553 (0.6%)
	Black Caribbean or Black African	40-54	14 (0.02%)	1628 (1.8%)	17 (0.02%)	1194 (1.4%)
		55-69	16 (0.02%)	633 (0.7%)	18 (0.02%)	415 (0.5%)
	Chinese	40-54	1 (0.001%)	390 (0.4%)	0 (0.0%)	233 (0.3%)
		55-69	4 (0.01%)	238 (0.3%)	1 (0.001%)	112 (0.1%)
	Other	40-54	10 (0.01%)	948 (1.1%)	8 (0.01%)	624 (0.7%)
		55-69	13 (0.01%)	501 (0.6%)	14 (0.02%)	249 (0.3%)

IRT = Integrated Risk Tool.

Detailed tables of sensitivity, specificity, PPV, NPV, Harrell's C, and their comparative differences to ASCVD-PCE (deltas and NRI) are provided in Supplementary Tables 2 to 5). We also provide an equivalent Supplementary Table 6 for an analysis carried out using a different version of the IRT that used the same PRS and PRS coefficients but was integrated with the QRISK2 score that is recommended for use in the UK.[27,28]

Discussion

Our results indicate that ASCVD-IRT, a new tool for estimating 10-year ASCVD risk that incorporates a PRS for ASCVD, outperforms the existing standard-of-care tool ASCVD-PCE, and that this improvement extends across ethnicities and genetic ancestries. The 2 US-based cohorts used in our validation (ARIC and MESA) are drawn, in part, from minority US ethnic groups, allowing us to demonstrate that, in addition to individuals of White ethnicities or European ancestries, the significant improvement of ASCVD-IRT is also seen in individuals with Black or African American self-reported ethnicities and African genetic ancestries.

Data in other ethnicities and ancestries are more limited. However, the UKB contains a reasonably large group (n = 3,688, with 136 cases) of individuals of South Asian ("Indian, Bangladeshi, or Pakistani") ethnicities, and a significantly improved performance of ASCVD-IRT is seen in this group as well. MESA contains a smaller group of Hispanic ethnicities (n = 823, with 55 cases), and although the NRI point estimate (7.5%) is positive, there was insufficient power to reject the null hypothesis of no change. Groups of East Asian ethnicities are even smaller in the IRT testing data (n = 979, with 6 cases, of Chinese ethnicities in UKB; n = 5, with 1 case, of Chinese American ethnicities in MESA), and this resulted in poor estimation of NRI and wide 95% CIs that extend to either side of zero.

The extent to which the results from larger groups can be generalized to support the clinical use of ASCVD-IRT in individuals from ethnicities and ancestries that are poorly represented in the IRT testing data is an important question. Three lines of evidence support such a generalization. First, we have data from additional cohorts (described in Supplementary Materials) that indicate the ASCVD PRS has an effect size in individuals of Hispanic and East Asian ethnicities that is positive, significantly different from zero and comparable in size to that seen in individuals of White ethnicities, with an equivalent pattern also seen across genetically inferred ancestries (Supplementary Figure 4). Although these cohorts lack the necessary longitudinal and covariate information to calculate ASCVD-PCE at baseline, and therefore could not be used for IRT testing, these results permit the inference that the strong predictive performance seen at the PRS level should transfer to the IRT level. Second, in line with previous work,[5,8] we observe a low (and statistically nonsignificant) correlation between PRS values and ASCVD-PCE scores in the IRT testing cohorts (ARIC: r = 0.026, 95% CI −0.018 to 0.070 [Fisher's z-method]; MESA: r = 0.002, 95% CI −0.043 to 0.047; UKB: r = 0.002, 95% CI −0.005 to 0.008). This increases our confidence that the ASCVD PRS acts largely independently of ASCVD-PCE, and strengthens the inference that PRS results should therefore transfer to the IRT level. Third, both population genetic theory and prior data indicate that individuals of African ancestries should be most affected by attenuation in PRS effect size.[9–11] Thus, other non-European ancestries should be intermediate in attenuation.

Table 2.
IRT testing cohort sample numbers (and percentage of cohort) by sex, age at recruitment, case status and genetically inferred ancestry

Cohort	Genetic ancestry	Age at recruitment	Women		Men	
			Cases	Noncases	Cases	Noncases
ARIC	AFR	40-54	46 (2.3%)	748 (38%)	37 (1.9%)	410 (21%)
		55-69	39 (2.0%)	370 (19%)	40 (2.0%)	223 (11%)
	EUR	40-54	0 (0.0%)	4 (0.2%)	0 (0.0%)	2 (0.1%)
		55-69	0 (0.0%)	1 (0.1%)	0 (0.0%)	3 (0.2%)
	OTHER	40-54	1 (0.1%)	10 (0.5%)	1 (0.1%)	10 (0.5%)
		55-69	1 (0.1%)	12 (0.6%)	0 (0.0%)	7 (0.4%)
MESA	AFR	40-54	4 (0.2%)	202 (10%)	8 (0.4%)	156 (8.1%)
		55-69	10 (0.5%)	246 (13%)	21 (1.1%)	217 (11%)
		70-79	14 (0.7%)	106 (5.5%)	14 (0.7%)	105 (5.5%)
	AMR	40-54	0 (0.0%)	41 (2.1%)	0 (0.0%)	62 (3.2%)
		55-69	4 (0.2%)	55 (2.9%)	8 (0.4%)	91 (4.7%)
		70-79	2 (0.1%)	21 (1.1%)	4 (0.2%)	26 (1.4%)
	EAS	40-54	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
		55-69	0 (0.0%)	2 (0.1%)	1 (0.1%)	3 (0.2%)
		70-79	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	EUR	40-54	2 (0.1%)	39 (2.0%)	0 (0.0%)	39 (2.0%)
		55-69	5 (0.3%)	38 (2.0%)	6 (0.3%)	45 (2.3%)
		70-79	3 (0.2%)	14 (0.7%)	2 (0.1%)	28 (1.5%)
	OTHER	40-54	0 (0.0%)	43 (2.2%)	1 (0.1%)	43 (2.2%)
		55-69	1 (0.1%)	71 (3.7%)	10 (0.5%)	49 (2.5%)
		70-79	4 (0.2%)	20 (1.0%)	5 (0.3%)	34 (1.8%)
UKB (IRT testing)	AFR	40-54	20 (0.02%)	1876 (2.1%)	19 (0.02%)	1359 (1.5%)
		55-69	18 (0.02%)	722 (0.8%)	19 (0.02%)	462 (0.5%)
	EAS	40-54	3 (0.003%)	705 (0.8%)	0 (0.0%)	363 (0.4%)
		55-69	10 (0.01%)	412 (0.5%)	3 (0.003%)	152 (0.2%)
	EUR	40-54	135 (0.2%)	18182 (21%)	307 (0.3%)	14148 (16%)
		55-69	588 (0.7%)	25146 (28%)	1048 (1.2%)	18090 (20%)
	SAS	40-54	11 (0.01%)	802 (0.9%)	26 (0.03%)	714 (0.8%)
		55-69	12 (0.01%)	492 (0.6%)	33 (0.04%)	353 (0.4%)
	OTHER	40-54	3 (0.003%)	803 (0.9%)	34 (0.04%)	770 (0.9%)
		55-69	10 (0.01%)	414 (0.5%)	32 (0.04%)	370 (0.4%)

IRT = Integrated Risk Tool; AFR = Sub-Saharan African; AMR = Native/Indigenous American; EAS = East Asian; EUR = European; SAS = South Asian.

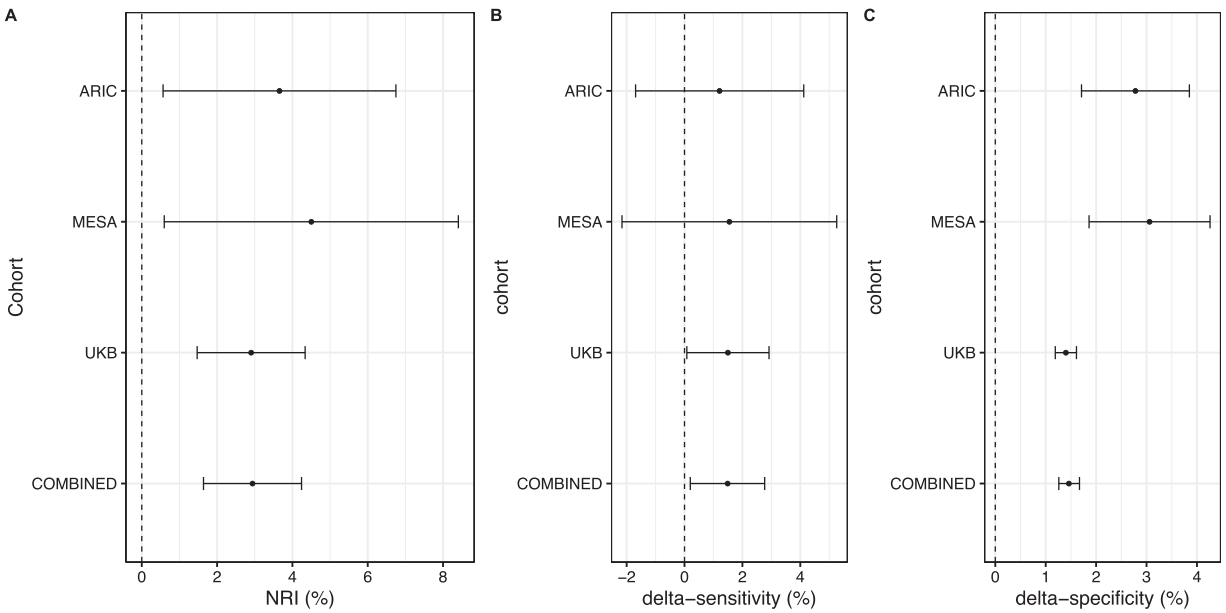


Figure 1. Relative performance of ASCVD-IRT over the currently established ASCVD-PCE tool, split by cohort and combined via meta-analysis across cohorts. (a) Net Reclassification Improvement (NRI), with 95% confidence intervals (CI). (b) Delta-sensitivity (equivalent to NRI in cases) with 95% CI. (c) Delta-specificity (equivalent to NRI in noncases) with 95% CI. Vertical dotted lines at zero indicate the null hypothesis of no change.

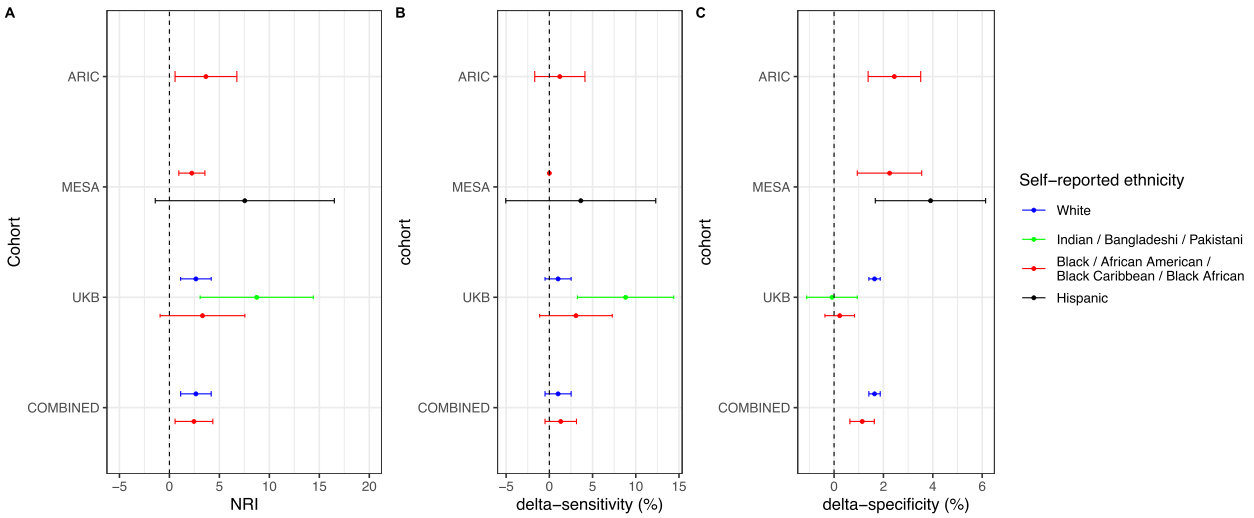


Figure 2. Relative performance improvements of ASCVD-IRT over the currently established ASCVD-PCE tool, split by self-reported ethnicities. (a) Net Reclassification Improvement (NRI), with 95% confidence intervals (CI). (b) Delta-sensitivity (equivalent to NRI in cases) with 95% CI. (c) Delta-specificity (equivalent to NRI in noncases) with 95% CI. Vertical dotted lines at zero indicate the null hypothesis of no change.

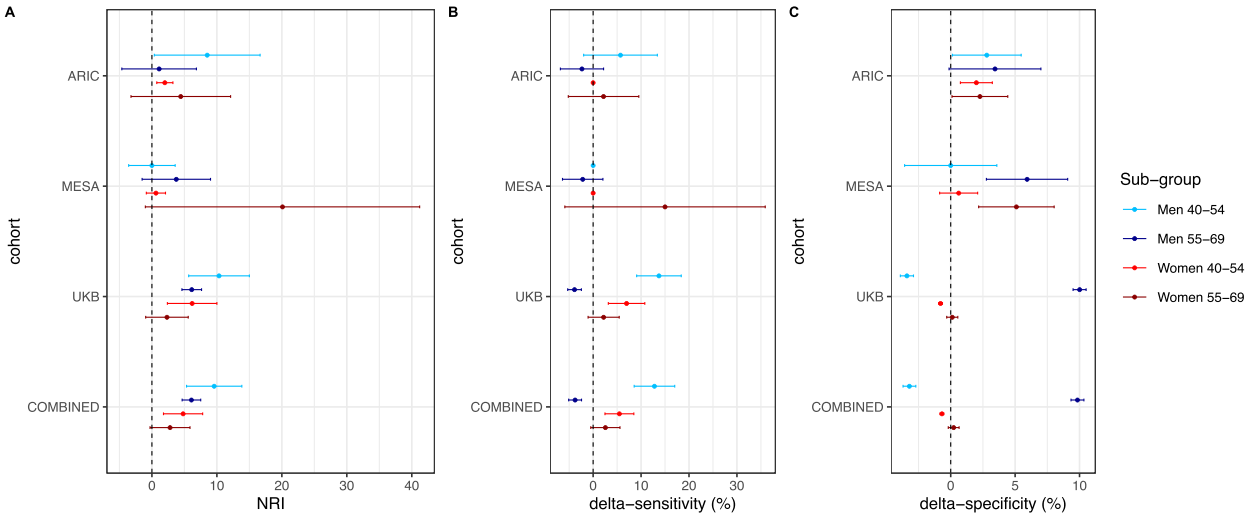


Figure 3. Relative performance improvements of ASCVD-IRT over the currently established ASCVD-PCE tool, split by 4 sex-by-age subgroups. (a) Net Reclassification Improvement (NRI), with 95% confidence intervals (CI). (b) Delta-sensitivity (equivalent to NRI in cases) with 95% CI. (c) Delta-specificity (equivalent to NRI in noncases) with 95% CI. Vertical dotted lines at zero indicate the null hypothesis of no change.

Just as ASCVD-PCE is itself an improvement over previous risk estimators,[1,29] so we conclude the addition of a PRS to ASCVD-PCE should lead to further predictive enhancement. But for clinical utility, the ASCVD-IRT tool requires gains in performance that are not only statistically significant but also clinically meaningful.[30] On this latter point, we note that large gains are seen especially in younger middle-aged men (40 to 54 year old), not only in this study (overall NRI=10.3% [95% CI 5.7 to 15.0]), but also in a previous study on individuals of European ancestry in UK Biobank, where an NRI of 15.4% (95% CI 11.6 to 19.3) was observed for coronary artery disease outcomes.[5] Furthermore, a large effect is also seen in younger middle-aged ARIC men of Black ethnicity in the current study (NRI = 8.5% [95% CI 0.4 to 16.7]), suggesting that this effect generalizes to other ethnicities and ancestries. We note that ASCVD is a more heterogenous condition than coronary artery disease, which may explain the observed drop in NRI.

There are limitations to this study. Tools that predict future risk, such as ASCVD-PCE and ASCVD-IRT, require larger datasets for validation than diagnostic clinical tools that typically have high sensitivity, specificity, PPV, and NPV. Thus, our conclusions remain data limited. We demonstrate significant gains in performance in some ethnicities and ancestries, but generalization to other groups requires additional inferential steps. More data should be collected to further validate and optimize ASCVD-IRT, improve risk prediction, further incorporate variable genetic ancestry among individuals, and assess performance gains in different subgroups. We also note this study does not address the question of analytical validation, and additional evidence is

required to demonstrate the value of specific sampling and genotyping or sequencing protocols for the accurate computation of risk scores.

In conclusion, using data from multiple ethnicities and ancestries, we have shown improved predictive performance of the ASCVD-IRT tool over the currently established ASCVD-PCE tool in multiple cohorts, multiple ethnicities, and multiple ancestries. To our knowledge, this is the first time, for any disease, that an integrated risk tool combining a current clinical risk tool and a PRS has been successfully validated across multiple ethnicities and ancestries.

Author Contributions

Michael E. Weale: Writing - Original draft preparation, Methodology, Investigation, Supervision. Fernando Riveros-Mckay: Writing - Review & Editing, Methodology, Formal analysis, Software, Validation, Investigation, Visualization. Saskia Selzam, Eva Gradovich: Data Curation, Software, Resources. Priyanka Seth, Rachel Moore, Carla Giner-Delgado, Duncan Palmer, Daniel Wells, Ayden Saffari: Methodology, Software, Resources. William A. Tarran, R. Michael Sivley: Software, Resources, Validation. Alexander S. Lachapelle: Conceptualization, Project administration. Hannah Wand, Lee Shoa Long Clarke, Joshua W. Knowles, Jack W O'Sullivan, Euan A Ashley: Writing - Review & Editing. Gil McVean: Conceptualization, Writing - Review & Editing, Project administration. Vincent Plagnol, Peter Donnelly: Conceptualization, Writing - Review & Editing, Supervision, Project administration.

Acknowledgments

We acknowledge and thank the participants and investigators of all the datasets used in this study. Please see Supplementary Materials for detailed acknowledgements.

Disclosures

Peter Donnelly and Gil McVean are partners in Peptide Groove LLP. All other authors declare no competing interests.

Supplementary materials

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.amjcard.2021.02.032.

1. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, 2013 Wilson PWF. ACC/AHA Guideline on the assessment of cardiovascular risk. *Circulation* 2014;129:S49–S73.
2. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, Braun LT, Ferranti S de, Faiella-Tommasino J, Forman DE, Goldberg R, Heidenreich PA, Hlatky MA, Jones DW, Lloyd-Jones D, Lopez-Pajares N, Ndumele CE, Orringer CE, Peralta CA, Saseen JJ, Smith SC Jr, Sperling L, Virani SS, Yeboah J. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the management of blood cholesterol. *Circulation* 2018;139:1082–1143.
3. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, Himmelfarb CD, Khera A, Lloyd-Jones D, McEvoy JW, Michos ED, Miedema MD, Muñoz D, Smith SC, Virani SS, Williams KA, Yeboah J, Ziaeian B. 2019 ACC/AHA Guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol* 2019;74:e177–e232.
4. Khera AV, Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet* 2017;18:331–344.
5. Riveros-Mckay F, Weale M, Moore R, Selzam S, Krapohl E, Sivley RM, Tarran W, Sørensen P, Lachapelle A, Griffiths J, Saffari A, Deanfield J, Spencer C, Hippisley-Cox J, Hunter D, O'Sullivan J, Ashley E, Plagnol V, Donnelly P. An integrated polygenic tool substantially enhances coronary artery disease prediction. *Circulation: Genomic and Precision Medicine* 2021. (in press). https://scholar.google.com/scholar_lookup?title=An%20integrated%20polygenic%20tool%20substantially%20enhances%20coronary%20artery%20disease%20prediction&author=F%20Riveros-Mckay&publication_year=2021.
6. Elliott J, Bodinier B, Bond T, Chadeau-Hyam M, Evangelou E, Moons K, Dehghan A, Muller DC, Elliott P, Tzoulaki I. Predictive accuracy of a polygenic risk score−enhanced prediction model vs a clinical risk score for coronary artery disease. *J Am Med Assoc* 2020;323:636–645.
7. Mosley J, Gupta D, Tan J, Yao J, Wells QS, Shaffer CM, Kundu S, Robinson-Cohen C, Psaty BM, Rich SS, Post WS, Guo X, Rotter JI, Roden DM, Gerszten RE, Wang TJ. Predictive accuracy of a polygenic risk score compared with a clinical risk score for incident coronary heart disease. *J Am Med Assoc* 2020;323:627–635.
8. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, Ahola-Olli A, Kurki M, Karjalainen J, Palta P, Neale BM, Daly M, Salomaa V, Palotie A, Widén E, Ripatti S. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* 2020;26:549–557.
9. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;51:584–591.
10. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* 2019;10:3328.
11. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun* 2020;11:3865.
12. Khan A, McHugh C, Conomos MP, Gogarten SM, Nelson SC, Race and Genetics Discussion Group. Guidelines on the use and reporting of race, ethnicity, and ancestry in the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. Available at: https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed. Accessed December 10, 2020.
13. Race Ethnicity and Genetics Working Group. The use of racial, ethnic, and ancestral categories in human genetics research. *Am J Hum Genet* 2005;77:519–532.
14. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;526:68–74.
15. Mathieson I, Scally A. What is ancestry? *PLoS Genet* 2020;16:e1008624.
16. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol* 1989;129:687–702.
17. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux A V, Folsom AR, Greenland P, Jacobs DR, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol* 2002;156:871–881.
18. Olson JL, Bild DE, Kronmal RA, Burke GL. Legacy of MESA. *Glob Heart* 2016;11:269–274.
19. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779.
20. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P,

Marchini J. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562:203–209.
21. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese AK, Van Der Laan SW, Gretarsdottir S, Anderson CD, Chong M, Adams HHH, Ago T, Almgren P, Amouyel P, Ay H, Bartz TM, Benavente OR, Bevan S, Boncoraglio GB, Brown RD, Butterworth AS, Carrera C, Carty CL, Chasman DI, Chen WM, Cole JW, Correa A, Cotlarciuc I, Cruchaga C, Danesh J, Bakker PIW De, Destefano AL, Den Hoed M, Duan Q, Engelter ST, Falcone GJ, Gottesman RF, Grewal RP, Gudnason V, Gustafsson S, Haessler J, Harris TB, Hassan A, Havulinna AS, Heckbert SR, Holliday EG, Howard G, Hsu FC, Hyacinth HI, Ikram MA, Ingelsson E, Irvin MR, Jian X, Jiménez-Conde J, Johnson JA, Jukema JW, Kanai M, Keene KL, Kissela BM, Kleindorfer DO, Kooperberg C, Kubo M, Lange LA, Langefeld CD, Langenberg C, Launer LJ, Lee JM, Lemmens R, Leys D, Lewis CM, Lin WY, Lindgren AG, Lorentzen E, Magnusson PK, Maguire J, Manichaikul A, McArdle PF, Meschia JF, Mitchell BD, Mosley TH, Nalls MA, Ninomiya T, O'Donnell MJ, Psaty BM, Pulit SL, Rannikmäe K, Reiner AP, Rexrode KM, Rice K, Rich SS, Ridker PM, Rost NS, Rothwell PM, Rotter JI, Rundek T, Sacco RL, Sakaue S, Sale MM, Salomaa V, Sapkota BR, Schmidt R, Schmidt CO, Schminke U, Sharma P, Slowik A, Sudlow CLM, Tanislav C, Tatlisumak T, Taylor KD, Thijs VNS, Thorleifsson G, Thorsteinsdottir U, Tiedt S, Trompet S, Tzourio C, Van Duijn CM, Walters M, Wareham NJ, Wassertheil-Smoller S, Wilson JG, Wiggins KL, Yang Q, Yusuf S, AFGen Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, International Genomics of Blood Pressure (iGEN-BP) Consortium, INVENT Consortium, STARNET, Bis JC, Pastinen T, Ruusalepp A, Schadt EE, Koplev S, Björkegren JLM, Codoni V, Civelek M, Smith NL, Trégouët DA, Christophersen IE, Roselli C, Lubitz SA, Ellinor PT, Tai ES, Kooner JS, Kato N, He J, Harst P van der, Elliott P, Chambers JC, Takeuchi F, Johnson AD, BioBank Japan Cooperative Hospital Group, COMPASS Consortium, EPIC-CVD Consortium, EPIC-InterAct Consortium, International Stroke Genetics Consortium (ISGC), METASTROKE Consortium, Neurology Working Group of the CHARGE Consortium, NINDS Stroke Genetics Network (SiGN), UK Young Lacunar DNA Study, MEGASTROKE Consortium, Sanghera DK, Melander O, Jern C, Strbian D, Fernandez-Cadenas I, Jr WTL, Rolfs A, Hata J, Woo D, Rosand J, Pare G, Hopewell JC, Saleheen D, Stefansson K, Worrall BB, Kittner SJ, Seshadri S, Fornage M, Markus HS, Howson JMM, Kamatani Y, Debette S, Dichgans M. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* 2018;50:524–537.
22. The CARDIoGRAMplusC4D Consortium. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;47:1121–1130.
23. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–172.
24. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *J Am Med Assoc* 1982;247:2543–2546.
25. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, Lai FY, Kaptoge S, Brozynska M, Wang T, Ye S, Webb TR, Rutter MK, Tzoulaki I, Patel RS, Loos RJF, Keavney B, Hemingway H, Thompson J, Watkins H, Deloukas P, Di Angelantonio E, Butterworth AS, Danesh J, Samani NJ. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J Am Coll Cardiol* 2018;72:1883–1893.
26. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, Fedotov A, Feng Q, Hakonarson H, Jarvik GP, Lee MTM, Pacheco JA, Rowley R, Sleiman PM, Stein CM, Sturm AC, Wei WQ, Wiesner GL, Williams MS, Zhang Y, Manolio TA, Kullo IJ. Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. *Am J Hum Genet* 2020;106:707–716.
27. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P. Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Br Med J* 2008;336:1475–1482.
28. National Institute for Health and Care Excellence. *Cardiovascular disease: risk assessment and reduction, including lipid modification (Clinical guideline [CG181])*. Available at: https://www.nice.org.uk/guidance/cg181.
29. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837–1847.
30. Khan SS, Cooper R, Greenland P. Do polygenic risk scores improve patient selection for prevention of coronary artery disease? *J Am Med Assoc* 2020;323:614–615.