The National Academies of

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

April 4, 2022 11AM – 5PM ET

WEBCAST LINK:

https://www.nationalacademies.org/event/04-04-2022/committee-on-use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research-meeting-2-and-public-workshop

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

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

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

Public Workshop: April 4, 2022

Table of Contents

1. WORKSHOP AGENDA 3

2. STUDY INFORMATION 9

Statement of Task 10

Committee Roster 12

Committee Biosketches 14

3. WORKSHOP INFORMATION 21

Speaker Biosketches 22

Speaker Guidance 27

4. BACKGROUND INFORMATION 30

Links to Additional Resources 31

Byeon et al., 2021 33

Huddart et al., 2019 42

Morales et al., 2018 57

Duster, 2015 67

Kahn, 2006 94

AGENDA

The National Academies of
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

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

PUBLIC WORKSHOP
April 4, 2022
All times listed in Eastern Standard Time

Register Here: https://www.nationalacademies.org/event/04-04-2022/committee-on-use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research-meeting-2-and-public-workshop

---

11:00 am ET Welcome and Goals for the Workshop

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

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

Session I: Historical and Current Use of Population Descriptors in Genomics Research

---

Moderator: Sandra Soo-Jin Lee, Columbia University

Session Objectives:

- To explore historical use of population descriptors to better understand current use
- To examine whom we study in genomic investigations
- To explore why we identify individuals and populations in genomic studies

- To examine and identify the criticisms and challenges in current use of population descriptors in genomics research

11:10 am **Brief introduction to the session by the moderator**

11:15 am **Speakers' Talks (15 minutes each)**

Pilar Ossorio
Professor of Law and Bioethics
University of Wisconsin-Madison, Law School
University of Wisconsin-Madison, Medical School

Joseph Graves
Professor of Biological Sciences
PI: IBIEM@ AT and BEACON@A&T
Associate Director, Triangle Center for Evolutionary Medicine
Department of Biology
North Carolina A&T State University

Andrew Clark
Jacob Gould Schurman Professor of Population Genetics
Nancy and Peter Meinin Family Investigator
Associate Director, Cornell Center for Comparative and Population Genomics
Interim Chair, Department of Computational Biology
Cornell University

Rina Bliss
Associate Professor of Sociology
Rutgers University

12:15 pm **Q&A with speakers**

1:00 pm **Break**

Session II: Future Use of Population Descriptors in Genomics Research

Moderator: Rick Kittles, City of Hope

Session Objectives:

- To consider the diverse types of population and individual descriptors (e.g. origins, definitions, and usage in the U.S., implications for non-U.S. participants)
- To discuss possible ideal descriptors of populations and individuals
- To consider standardized or ideal systems of population descriptors

1:30 pm **Brief introduction to the session by the moderator**

1:35 pm **Speakers' Talks (15 minutes each)**

Tesfaye Mersha
Associate Professor of Human Quantitative Genetics
Department of Pediatrics
University of Cincinnati
Cincinnati Children's Hospital Medical Center

Melinda Mills
Director, Leverhulme Centre for Demographic Sciences
Nuffield Professor of Sociology
University of Oxford

Joanna Mountain
Consultant
23andMe

Eimear Kenny
Founding Director, Institute for Genomic Health
Professor of Genetics and Medicine
Icahn School of Medicine at Mount Sinai

Stephanie Malia Fullerton
Professor of Bioethics and Humanities
University of Washington School of Medicine
Adjunct Professor
Departments of Epidemiology, Genome Sciences, and Medicine
University of Washington
Affiliate Investigator, Public Health Sciences Division
Fred Hutchinson Cancer Research Center

2:55 pm **Q&A with speakers**

3:50 pm **Break**

Session III: Community Input on Population Descriptors in Genomics Research

Moderator: Katrina Claw, University of Colorado Denver – Anschutz Medical Campus

Session Objectives:

- 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?

4:00 pm **Brief introduction to the session by the moderator**

4:05 pm **Speakers' Comments (5 minutes each)**

Catherine Potenski
Chief Editor
Nature Genetics

Donna Cryer
President & CEO
Global Liver Institute

Agustín Fuentes
Professor
Department of Anthropology
Princeton University

Charles Rotimi
President
American Society for Human Genetics

Judit Kumuthini
Bioinformatics Manager
Human Capacity Development Manager: Bioinformatics
University of Western Cape

Shishi Luo
Associate Director, Bioinformatics and Infectious Diseases
Helix Genomics

Julia Ortega
Vice President
iHope Genetic Health
Genetic Alliance

4:55 pm **Concluding Remarks**

5:00 pm **Adjourn**

STUDY INFORMATION

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Board on Health Sciences Policy

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

The National Academies of
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

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

Division of Behavioral and Social Science and Education
COMMITTEE ON POPULATION

The National Academies of
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'.

The National Academies of
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.

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.

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.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

HEALTH AND MEDICINE DIVISION
BOARD ON HEALTH SCIENCES POLICY

Division of Behavioral and Social Science and Education
COMMITTEE ON POPULATION

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

The National Academies of
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

April 4, 2022

Speaker Biosketches

Rina Bliss, Ph.D., is Associate Professor of Sociology at Rutgers University. She teaches courses in the sociology of health and illness, and science and technology. She is the author of *Race Decoded: The Genomic Fight for Social Justice* and *Social by Nature: The Promise and Peril of Sociogenomics*. Dr. Bliss's research examines the sociology of today's newest avenues in science and medicine – namely, genomics and postgenomics. In *Race Decoded: The Genomic Fight for Social Justice*, Bliss reveals how DNA science has emerged to become the newest authority on the meaning of race. She demonstrates the institutionalization of academic-industry-governmental ties as well as the crystallization of deterministic notions of race that perpetuate social inequality even as they aim to prevent it. In *Social by Nature: The Promise and Peril of Sociogenomics*, Bliss illuminates cutting-edge developments in gene-environment science as natural and social scientists partner to create DNA analyses of social behavior. Following a small group of innovators, the book exposes the evolution of a new field of genomics that evinces novel patterns in interdisciplinarity, such as transdisciplinarity marked by power imbalances in genetic versus social science. These imbalances shape how the field constructs concepts of race, gender, and sexuality. Dr. Bliss received her Ph.D. in Sociology from the New School for Social Research in 2009.

Andrew Clark, Ph.D., is the Jacob Gould Schurman Professor of Population Genetics in the Department of Molecular Biology and Genetics, a Nancy and Peter Meinig Family Investigator, the Associate Director of the Cornell Center for Comparative and Population Genomics (3CPG), and the chair of the Department of Computational Biology at Cornell University. Dr. Clark is a population geneticist focused on empirical and analytical problems associated with genetic variation in populations. Dr. Clark researches the genetic basis of adaptive variation in natural populations, with an emphasis on quantitative modeling of phenotypes as networks of interacting genes. He has several projects centered on the genetic basis for complex traits, especially in cases that have a well understood gene regulatory network underlying the trait. His work in humans concentrates on cardiovascular disease risk, population genetic applications of genome-wide SNP data, and the phenomenon of genomic imprinting. He has published more than 400 peer-reviewed papers in the field of population genetics and is co-author with Dan Hartl of Principles of Population Genetics. Dr. Clark was elected Fellow of the American Association for the Advancement of Science (AAAS) in 1994, and he serves on review panels for the

National Institutes of Health, National Science Foundation, and the Max Planck Society. In May of 2012, Dr. Clark was elected to the National Academy of Science. He received his Ph.D. in Population Genetics at Stanford University in 1980.

Stephanie Malia Fullerton, D.Phil, is Professor of Bioethics and Humanities at the University of Washington School of Medicine. She is also Adjunct Professor in the UW Departments of Epidemiology, Genome Sciences, and Medicine (Medical Genetics), as well as an affiliate investigator with the Public Health Sciences division of the Fred Hutchinson Cancer Research Center. Dr. Fullerton's work focuses on the ethical and social implications of genomic research and its equitable and safe translation for clinical and public health benefit. She contributes to a range of empirical projects focused on clinical genomics translation and precision medicine approaches to the treatment and prevention of common complex diseases in diverse patient populations. She received a Ph.D. in Human Population Genetics from the University of Oxford and later re-trained in Ethical, Legal, and Social Implications (ELSI) research with a fellowship from the NIH National Human Genome Research Institute.

Joseph Graves, Jr., Ph.D., is Professor of Biological Sciences in the Biology Department of North Carolina A&T State University. His research examines the evolutionary genomics of adaptation, biological aging, and bacterial responses to nanomaterials. Dr. Graves has authored multiple books on the biology of race, including *The Emperor's New Clothes: Biological Theories of Race* (2001), *The Race Myth: Why We Pretend Race Exists in America* (2005), and *Racism, Not Race: Answers to Frequently Asked Questions* (2021) with Alan Goodman. Dr. Graves was elected a Fellow of the Council of the American Association for the Advancement of Science (AAAS) in 1994. In 2012, he was chosen as one of the "Sensational Sixty" commemorating 60 years of the National Science Foundation Graduate Research Fellowship Award. In 2017, he was listed as an "Outstanding Graduates" in Biology at Oberlin College and was named an "Innovator of the Year" in US Black Engineer Magazine. Graves received his Ph.D. in Environmental, Evolutionary and Systematic Biology from Wayne State University in 1988.

Eimear Kenny, Ph.D., is a Professor of Medicine and Genetics, and the Founding Director of the Institute for Genomic Health, at the Icahn School of Medicine at Mount Sinai. She is a renowned expert in population genetics and translational genomics. She leads a multidisciplinary team of geneticists, computer scientists, clinician scientists, working on problems at the interface of genetic ancestry, genomics, and medicine. Her research is focused on uncovering the clinical impact of human genetic variation and accelerating the implementation of genomic information in routine clinical care in diverse populations. Her work seeks to build better tools, resources, and best practices to broaden diversity and representation in genomic research and ameliorate health disparities in the implementation of genomic medicine. She is Principal Investigator in 6

large national programs focused on genomic research, medicine and health. She is a scientific advisor to many genomic and genomic medicine initiatives in government, non-profit and industry arenas. She has published over 130 papers in leading journals like *Science*, *Nature*, *Nature Genetics*, *NEJM*, with over 18,000 citations, and her work has been featured by many media outlets including the *New York Times*. She has a B.A. in Biochemistry from Trinity College Dublin, a Ph.D. in computational genomics from Rockefeller University, and did her postdoctoral training in population genetics at Stanford University

Tesfaye B. Mersha, Ph.D., is Associate Professor in Human Genetics at the Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine. Dr. Mersha leads the Population Genetics, Ancestry and Bioinformatics (pGAB) Laboratory. Dr. Mersha's research combines quantitative, ancestry and statistical genomics to unravel genetic and non-genetic contributions to complex diseases and racial disparities in human populations, particularly asthma and asthma-related allergic disorders. Mersha is a recognized expert in the field of genetic ancestry, race, ethnicity, admixture mapping and mining functional genomic databases related to complex diseases. Much of his research is at the interface of genetic ancestry, statistics, bioinformatics, and functional genomics, and he is interested in crossline disciplines to unravel the interplay between genome and environment underlying racial disparities in asthma risk. His long-term research goal is to understand and dissect how biologic predisposition and environmental exposures interact to shape racial disparities in complex disorders. His long-term research goals are to understand and dissect the role of genetic and genetic-modifying causes of asthma and reduce health disparities in children. Current research interests and projects include ancestry analysis, association mapping, gene expression analysis, biological pathways/networks analysis, data mining and multi-omics integration, and socio-cultural and exposure studies. Dr. Mersha has been appointed to serve on the National Institutes of Health (NIH) Genetics of Health and Disease Study Section and invited to speak at the NIH on the topic of "My Road to Success in Science." He has also been invited to moderate panels on the use of ancestry, race, and ethnicity in biomedical research at the NIH and at the Missouri State University Public Affairs Conference. Dr. Mersha is the recipient of awards and honors that include a Faculty Research Achievement Award from Cincinnati Children's Hospital Medical Center, the African Professionals Network (APNET) Business and Professional Achievement Award, and a Keystone Symposia Early Career Investigator Award. His research is continuously funded by the National Institute of Health (NIH).

Melinda Mills, Ph.D., is Director of Leverhulme Centre for Demographic Science (LCDS) and the Nuffield Professor of Sociology at the University of Oxford. She is the co-founder of the GWAS Diversity Monitor that tracks ancestral and geographic diversity of research subjects in medical and genetic discoveries in an effort to promote accountability among researchers. Dr. Mills is Principal Investigator of the Leverhulme Trust Large Centre Grant for the Leverhulme

Centre for Demographic Science, the European Research Council (ERC) Advanced Grant CHRONO, and the ERC-funded social business enterprise called DNA4Science. She is on the Scientific and Ethics Advisory Board of Our Future Health, the new 5 million person UK data collection project. Her research spans multiple topics in demography, empirical sociology, statistics and genetics. Her recent work focuses on sociogenomics, combining a social science and molecular genetic approach to the study of behavioral outcomes, with a focus on reproduction (fertility), assortative mating, chronotypes, and nonstandard employment. Other interests include behavioral approaches to health interventions, such as behavioral and policy responses to face coverings and vaccine deployment. Her work is widely cited, and Dr. Mills has published seven books and over 100 articles across multiple scientific disciplines including *Nature Genetics*, *Science*, *Proceedings of the National Academy of Sciences*, *Annual Review of Sociology*, *JAMA Psychiatry*, *Journal of Marriage and Family*, and *Social Forces*. She has written two statistical textbooks, *Introducing Survival and Event History Analysis* (2011) and *An Introduction to Statistical Genetic Data Analysis* (2020). Mills received her Ph.D. in Demography from the University of Groningen in the Netherlands in 2000.

Joanna Mountain, Ph.D., is a consultant and former Senior Director of Research at 23andMe. Dr. Mountain joined 23andMe from Stanford University where she specialized in human evolutionary genetics as a faculty member within the Anthropological Sciences and Genetics Departments. Her areas of interest include the following: phenotype and the interactions among genotype, environment, and culture; the extent to which genetic data can reveal details of human history; ethical issues regarding human genetics; biology, genetics, and concepts of race and ethnicity; and the development of statistical tools for analyzing population genetic data. Dr. Mountain has been the recipient of a number of academic awards, including grants from the National Science Foundation and National Institutes of Health. Dr. Mountain received her Ph.D. in Genetics from Stanford University in 1994 and subsequently conducted postdoctoral research on human population genetics within the Integrative Biology Department at the University of California, Berkeley.

Pilar Ossorio, J.D., Ph.D., is Professor of Law and Bioethics at the University of Wisconsin-Madison (UW), where she is on the faculties of both the Law School and the Medical. In 2011, Dr. Ossorio became the inaugural Ethics Scholar-in-Residence at the Morgridge Institute for Research, a private, nonprofit research institute that is part of the Wisconsin Institutes of Discovery. In addition, at UW, she serves as the co-director of the Law and Neuroscience Program and is a faculty member in the Masters in Biotechnology Studies program and the Graduate Program in Population Health. Centering on research ethics and the protection of research participants, Dr. Ossorio's primary research interests include governance of large bioscience projects, data sharing in scientific research, the use of race in biomedical and social science research, ethical and regulatory issues in research with human subjects, the regulation

and ethics of online research, and policy issues raised through scientific discovery and translational research. Dr. Ossorio has participated in numerous advisory committees and boards that aid governments in establishing science policy. She has advised the U.S. National Institutes of Health, the Food and Drug Administration, Genome Canada, and Health Canada. She is an elected fellow of the American Association for the Advancement of Science (AAAS). She has also served as a member of, or liaison to, several boards and committees of the National Academies of Sciences, Engineering, and Medicine, including the National Cancer Policy Board, the Human Embryonic Stem Cell Advisory Committee, and the Committee on Intellectual Property Rights. Ossorio received her Ph.D. in Microbiology and Immunology from Stanford University in 1990 and her J.D. from the University of California at Berkeley in 1997.

The National Academies of
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 I: Historical and Current Use of Population Descriptors in Genomics Research

Objectives

- To explore historical use of population descriptors to better understand current use.
- To examine whom we study in genomic investigations.
- To explore why we identify individuals and populations in genomic studies.
- To examine and identify the criticisms and challenges in current use of population descriptors in genomics research.

Key Questions for Speakers:

1. What is the rationale for whom we study? Are the rationales different in different types of genomic studies?
2. How do we sample for these studies?
3. How and why have we addressed population diversity in these contexts?
4. How do genomic scientists define human populations? How is the notion of human populations contested?
5. What are the rationales for identifying people in genomic studies? Why do researchers sort people and populations in these studies?
6. What are the current descriptors used and why do we use them?

The National Academies of
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 II: Future Use of Population Descriptors in Genomics Research

Objectives

- To consider the diverse types of population and individual descriptors (e.g. origins, definitions, and usage in the U.S., implications for non-U.S. participants)
- To discuss possible ideal descriptors of populations and individuals
- To consider standardized or ideal systems of population descriptors

Key Questions for Speakers:

1. What might the features of a genetically or biologically based system of population descriptors be? What are the caveats of such a system? How can these systems be clear and understandable to the general public?
2. What should not be used as a population descriptor and why?
3. How could population descriptors capture true individual diversity?
4. How do population descriptors used in the U.S. translate, or not translate, in other countries?
5. What would it take to create globally standardized population descriptors? (e.g. stakeholders, collaborations, etc.)

The National Academies of
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. Byeon, Y. J. J., R. Islamaj, L. Yeganova, W. J. Wilbur, Z. Lu, L. C. Brody, and V. L. Bonham. 2021. Evolving use of ancestry, ethnicity, and race in genetics research – A survey spanning seven decades. *The American Journal of Human Genetics* 108: 2215-2223. doi:10.1016/j.ajhg.2021.10.008. https://www.sciencedirect.com/science/article/pii/S0002929721003852?via%3Dihub

2. Huddart, R., A. E. Fohner, M. Whirl-Carrillo, G. L. Wojcik, C. R. Gignoux, A. B. Popejoy, C. D. Bustamante, R. B. Altman, and T. E. Klein. 2019. Standardized biogeographic grouping system for annotating populations in pharmacogenetic research. *Clinical Pharmacology & Therapeutics* 105(5):1256-1262. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6465129/

3. Morales, J., D. Welter, E. H. Bowler, M. Cerezo, L. W. Harris, A. C. McMahon, P. Hall, H. A. Junkins, A. Milano, and E. Hastings. 2018. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS catalog. *Genome biology* 19(1):1-10. https://pubmed.ncbi.nlm.nih.gov/29448949/

4. Duster, T. 2015. A post-genomics surprise. The molecular reinscription of race in science, law and medicine. *The British Journal of Sociology* 66(1). https://pubmed.ncbi.nlm.nih.gov/25789799/

5. Kahn, J. 2006. Genes, race, and population: Avoiding a collision of categories. *American Journal of Public Health* 96(11):1965-1970. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1751810/

Links to Additional Resources

Session I: Historical and Current Use of Population Descriptors in Genomics Research

Abuabara, K., You, Y., Margolis, D. J., Hoffmann, T. J., Risch, N., & Jorgenson, E. (2020). Genetic ancestry does not explain increased atopic dermatitis susceptibility or worse disease control among African American subjects in 2 large US cohorts. *Journal of Allergy and Clinical Immunology*, *145*(1), 192-198. https://www.sciencedirect.com/science/article/abs/pii/S0091674919309674

Ali-Khan, S. E., T. Krakowski, R. Tahir, and A. S. Daar. 2011. The use of race, ethnicity and ancestry in human genetic research. *HUGO Journal* 5(1-4): 47-63. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3237839/

Duster, Troy. 2021. "I-CAST #14: Sociologist Troy Duster on U.S. Racial Politics." Temple University, Japan Campus. Suggested section to view: minute 42 – 49 [7 minutes total]. https://youtu.be/8RwQ5B-OgZY?t=2520

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. *Human Genetics* 126(3):355-362. https://link.springer.com/article/10.1007/s00439-009-0674-1

Race Ethnicity Genetics Working Group. 2005. The use of racial, ethnic, and ancestral categories in human genetics research. *American Journal of Human Genetics* 77(4): 519-532. https://pubmed.ncbi.nlm.nih.gov/16175499/

Session II: Future Use of Population Descriptors in Genomics Research

Flanagin, A., T. Frey, and S. L. Christiansen. 2021. Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* 326(7):621. https://jamanetwork.com/journals/jama/fullarticle/2783090

Lee, S. S. J., J. Mountain, B. Koenig, R. Altman, M. Brown, A. Camarillo, L. Cavalli-Sforza, M. Cho, J. Eberhardt, M. Feldman, R. Ford, H. Greely, R. King, H. Markus, D. Satz, M. Snipp, C. Steele, and P. Underhill. 2008. The ethics of characterizing difference: Guiding principles on using racial categories in human genetics. *Genome Biology* 9(404). https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-7-404

Minari, J., M. Yokono, K. Takashima, M. Kokado, R. Ida, and Y. Hishiyama. 2021. Looking back: Three key lessons from 20 years of shaping Japanese genome research regulations. *Journal of Human Genetics* 66: 1039-1041. https://www.nature.com/articles/s10038-021-00923-z

Session III: Community Input on Population Descriptors in Genomics Research

The *Committee on Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research* 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. The comments may be used, with attribution, in the committee's final report which is planned for release in 2023. The committee will also invite a few authors representing a sample of the submissions to present their comments at a public workshop in June. We will be accepting submissions to this form through June 1st, 2022.

Questions to Consider:

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

Submit your comment here.

Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades

Yen Ji Julia Byeon,[1,3] Rezarta Islamaj,[2] Lana Yeganova,[2] W. John Wilbur,[2] Zhiyong Lu,[2] Lawrence C. Brody,[4,*] and Vence L. Bonham[3,*]

Summary

To inform continuous and rigorous reflection about the description of human populations in genomics research, this study investigates the historical and contemporary use of the terms "ancestry," "ethnicity," "race," and other population labels in *The American Journal of Human Genetics* from 1949 to 2018. We characterize these terms' frequency of use and assess their odds of co-occurrence with a set of social and genetic topical terms. Throughout *The Journal*'s 70-year history, "ancestry" and "ethnicity" have increased in use, appearing in 33% and 26% of articles in 2009–2018, while the use of "race" has decreased, occurring in 4% of articles in 2009–2018. Although its overall use has declined, the odds of "race" appearing in the presence of "ethnicity" has increased relative to the odds of occurring in its absence. Forms of population descriptors "Caucasian" and "Negro" have largely disappeared from *The Journal* (<1% of articles in 2009–2018). Conversely, the continental labels "African," "Asian," and "European" have increased in use and appear in 18%, 14%, and 42% of articles from 2009–2018, respectively. Decreasing uses of the terms "race," "Caucasian," and "Negro" are indicative of a transition away from the field's history of explicitly biological race science; at the same time, the increasing use of "ancestry," "ethnicity," and continental labels should serve to motivate ongoing reflection as the terminology used to describe genetic variation continues to evolve.

Introduction

The field of human genetics has struggled since its inception with the task of conceptualizing and describing geographic and population-based genetic variation. First thought of as hierarchical and unequal taxonomic types, then reframed as isolates that differ in allele frequency,[1] and now in terms of genetic ancestry,[2] the idea of the "population" in human genetics has continuously evolved since the field's earliest decades. Today, advances in genomics continue to spur discussions about how the field can accurately describe human genetic diversity.[3] Central to these discussions is how it will reconcile its legacy of scientific racism.[4] We use this phrase to refer both to the historical practice of studying races as distinct biological groups and more broadly to the incorrect conceptualization of racial difference as biological in ways that contribute to social stratification and inequity.

Today, three concepts take center stage in these discussions, each of which brings its own challenges: ancestry, ethnicity, and race. Racial and ethnic group membership is used as a covariate in genomic studies to account for confounding related to genetic ancestry or social determinants of health. For example, geneticists may address confounding due to genetic ancestry by stratifying analyses by racial or ethnic categories or improve power to detect genetic associations by including a race or ethnicity variable that accounts for variation due to social stratification.[5] Although the field has made progress in rejecting the idea of racial and ethnic categories as discrete biological units, the continuing use of race and ethnicity as proxies for genetic ancestry remains scientifically and socially problematic.[6] Ancestry, more specifically, genetic ancestry, has been described as information about the ancestors or populations from whom one has inherited genetic material.[7] Although ancestry may lend itself to a quantitative description of human genetic variation, a unified definition of this concept has yet to be developed, and even a precise definition of the "populations" from whom one has inherited genetic material remains elusive.[7,8]

Given the complexity of these concepts and their underlying histories, there is a lack of consensus in the field on how ancestry, ethnicity, and race should be understood. This is reflected in the increasingly heterogeneous ways that the concepts are employed in clinical research and practice.[9,10] Members of the genetics community have called for consensus on how these data should and should not be used[6] as well as called on the National Institutes of Health to support the National Academy of the Sciences, Engineering, and Medicine in developing a consensus statement on best practices for characterizing human genetic diversity in research.[11,12] Others have proposed standardized systems for annotating populations[13] and expressed optimism that advances in genetic technologies may allow the field to move past the use of race and ethnicity.[3,14]

An important component of ongoing efforts to establish consensus in this area of human genetics is knowledge about the social and historical paths through which the

[1]Department of Sociology, Princeton University, Princeton, NJ 08544, USA; [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; [3]Social and Behavioral Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; [4]Division of Genomics and Society, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
*Correspondence: bonhamv@mail.nih.gov (V.L.B.), lbrody@mail.nih.gov (L.C.B.)
https://doi.org/10.1016/j.ajhg.2021.10.008.

Table 1. Percentage of 200-token segments and articles containing terms used for analysis, 1949–2018

Term	Articles, n (%)	Even segments, n (%)
Admixture, admixtures, admix, admixed	981 (8.5)	2,969 (2.5)
African, Africans (excluding African American)	1,116 (9.6)	3,398 (2.9)
Allele, alleles, allelic	7,984 (68.9)	37,219 (31.5)
Ancestry, ancestries, ancestral, ancestrally	2,351 (20.3)	6,799 (5.8)
Asian, Asians (excluding Asian American)	950 (8.2)	2,755 (2.3)
Behavior, behaviors, behavioral	1,608 (13.9)	2,928 (2.5)
Caucasian, Caucasians, Caucasoid, Caucasoids	1,391 (12.0)	3,381 (2.9)
Diversity, diverse	2,114 (18.2)	4,527 (3.8)
Environment, environments, environmental, environmentally	2,843 (24.5)	5,966 (5.1)
Ethnicity, ethnicities, ethnic, ethnically	2,208 (19.1)	4,344 (3.7)
European, Europeans (excluding European American)	2,637 (22.8)	6,545 (5.5)
Frequency, frequencies	7,769 (67.0)	28,741 (24.2)
Geography, geographies, geographic, geographically	1,131 (9.8)	2,438 (2.1)
Haplotype, haplotypes, haplotypic	3,720 (32.1)	15,489 (13.1)
Hispanic, Hispanics	397 (3.4)	883 (0.7)
Language, languages, linguistic, linguistically	870 (7.5)	1,992 (1.7)
Latino, Latinos, Latina, Latinas, Latinx	67 (0.6)	181 (0.2)
Linkage, linkages	5,605 (48.4)	21,950 (18.6)
Locus, loci	7,111 (61.4)	31,446 (26.7)
Negro, Negroes, Negroid, Negroids	373 (3.2)	1,121 (1.0)
Population, populations	7,572 (65.3)	31,899 (27.0)
Race, races, racial, racially	852 (7.4)	1,691 (1.6)
Religion, religions, religious, religiously	247 (2.1)	386 (0.3)
Social, socially	1,038 (9.0)	1,898 (1.6)
Socioeconomic, socioeconomically	215 (1.9)	353 (0.3)
Total	11,590 (100.0)	117,986 (100.0)

field has come to its current understanding of ancestry, ethnicity, and race. To this end, we investigated how the frequency of the terms "ancestry," "ethnicity," "race," and other population labels have changed over the 70-year publication history of *The American Journal of Human Genetics* (1949–2018). Additionally, in order to assess the evolving context in which the three concepts were used, we tested for non-random term co-occurrences between "ancestry," "ethnicity," and "race" and a predetermined set of social, genetic, and population terms from 1949 to 2018. In doing so, we aim to push for continuous and rigorous reflection surrounding the use of these population concepts in human genetics.

Material and methods

Data

We obtained digital versions of the full text of every document published in *The Journal* from its founding in 1949 up to 2018. These files were held by the National Library of Medicine (NLM) at the National Institutes of Health and obtained for purposes of research with permission from the American Society of Human Genetics. We sought matches for all articles in this archive to a PMID in MedLine and/or a PMCID in PubMed Central. Articles without a PMCID or PMID, which comprised book reviews, abstract books, disciplinary announcements, indexes, and tables of contents, were not included in the dataset. The majority of the remaining articles were scientific research articles (11,360), and a small minority (275) were award speeches and other communications. Of the 11,635 articles included in analysis, 6,750 were in the form of extracted text from optical character recognition (OCR) versions of scanned journal pages. For the remaining 4,885 articles, the full text was readily available in XML format.

After removing the references sections of articles, all text was converted to the ASCII character set. For text obtained from OCR versions of PDF files, words broken over line breaks were repaired as described in the supplemental materials and methods. Punctuation, numerical tokens, and single-character terms were removed. Finally, we ensured that any occurrence of the term

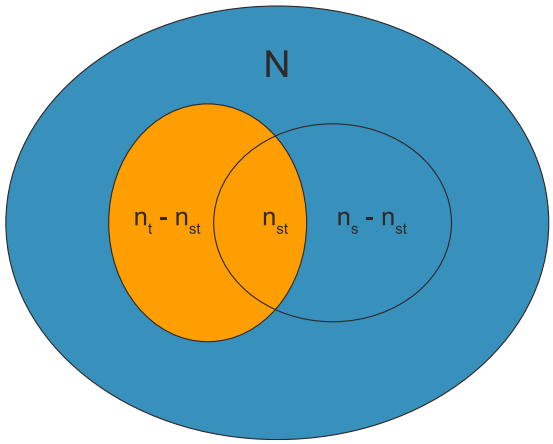


Figure 1. Visualization of parameters of random variable *y*, n_t, n_{st}, and n_s

Visualization of parameters of random variable *y*, n_t, n_{st}, and n_s, where *t* is the target term, *s* is the co-occurring word of interest, and *N* is the total number of 200-token segments in a given range of years. n_t represents number of segments containing *t*, n_s represents number of segments containing *s*, and n_{st} represents number of segments containing both.

"race" in the dataset referred solely to the population concept, as the term could also be a part of an author name (e.g., Robert Race), an abbreviation for a molecular biology method (rapid amplification of cDNA ends), or the word in the sense of a competition (e.g., "race to the finish line"). Informed by term associations, manual review, and orthographic characteristics of the term "race," we converted "race" to "xace" wherever it was not used in the population sense.

Selection of terms for analysis

We preselected 25 terms for which to calculate frequencies of use and odds of co-occurrence. In addition to "ancestry," "ethnicity," and "race," we examined 15 topical terms that have been related to these concepts ("admixture," "allele," "behavior," "diversity," "environment," "frequency," "geography," "haplotype," "language," "linkage," "locus," "population," "religion," "social," "socioeconomic") and seven population descriptors ("African," "Asian," "Caucasian," "European," "Hispanic," "Latina/o/x," "Negro"). The population descriptors "African," "Asian," and "European" refer to these specific forms of the descriptors and exclude uses of "African American," "Asian American," and "European American." Terms were selected for their relevance to ancestry, ethnicity, and race as well as specificity of meaning (e.g., "culture" was not chosen because it could also refer to cell cultures; "Black" and "White" could refer to the colors, as in "the black arrows indicate…" or "white blood cell"). We expanded each selected term to include alternate forms of the word with the same stem; for example, instances of "ancestral," "ancestries," and "ancestrally" were all counted as uses of "ancestry." Table 1 lists the 25 terms, their alternate forms, and their frequencies.

In order to investigate the ideas associated with and relationships between "ancestry," "ethnicity," and "race" in *The Journal*, we determined and compared co-occurrence patterns between (1) pairs of "ancestry," "ethnicity," and "race" (i.e., "ancestry" + "ethnicity," "ancestry" + "race," "ethnicity" + "race"), (2) 15 topical terms and each of "ancestry," "ethnicity," and "race" (i.e., each topical term + "ancestry," each topical term + "ethnicity," each topical term + "race" for 45 comparisons total), and (3) seven population descriptors and each of "ancestry," "ethnicity," and "race" (i.e., each population descriptor + "ancestry," each population descriptor + "ethnicity," each population descriptor + "race" for 21 comparisons total).

Measuring term co-occurrence

Using co-occurrence as a measure of relatedness between a pair of words is guided by a fundamental distributional hypothesis in linguistics stating that the meaning of words is determined by the contexts in which they occur or the words with which they occur.[15] This hypothesis informs statistical methods such as language modeling,[16] word embeddings,[17,18] and word similarity measures.[19,20] Given a pair of words, we analyze whether they co-occur more than expected by chance and interpret this as evidence that they have a semantic relationship.

Co-occurrence refers to how often a pair of words appear together in the same texts or documents. Ideally, these documents would all be of the same length, as longer documents are *a priori* more likely to contain a given word than shorter documents. To eliminate this bias, we split documents into disjoint text windows or segments of 200 words (space separated tokens) and defined the co-occurrence between two terms as the number of text windows in which both terms occur.[20] An advantage of this partitioning is that the relatively small size of the segments implies a closer relationship between words that co-occur. Using smaller pieces of text also increases the sensitivity of statistical testing to determine whether terms co-occur at a higher frequency than predicted by random mixing. Although paragraphs could also be used as text segments, a substantial proportion of the text we analyzed was obtained by OCR applied to scanned text, making it difficult to identify paragraph boundaries.

When partitioning documents into 200-token segments, it is possible that our two terms of interest become separated by the border between adjacent segments. To account for terms that are in close proximity but separated by the border of adjacent segments, we started a new segment after every 100 tokens, such that each had a 100-token overlap with adjacent segments. To eliminate double-counting of co-occurrences caused by the overlaps, we numbered the segments and computed results separately for even- and odd-numbered segments. We report results by using even segments. Results computed from odd segments were not substantially different and are reported in Table S3. Yearly counts of articles and segments are shown in Figure S1.

Figure 1 illustrates our method for assessing whether two terms co-occurred more often than expected by chance in a given set of segments. In our analyses, we considered either the even or odd segments from 10-year intervals at a time, incrementing the decades by one year in order to identify temporal trends (e.g., 1949–1958, 1950–1959, 1951–1960, …, 2008–2017, 2009–2018). The outside blue oval represents the set of all 200-token segments in a decade. Term *t* splits the space into two subsets of segments: those that contain the term (orange oval) and the rest. Further, consider term *s* and assume that it co-occurs with *t* in n_{st} segments. The p value is the probability that the observed or greater overlap between the two terms would happen by chance as determined by the size of the space of segments and the number of segments containing *s* and the number containing *t*. Mathematically, a random variable *y* representing the overlap between the target term *t* and

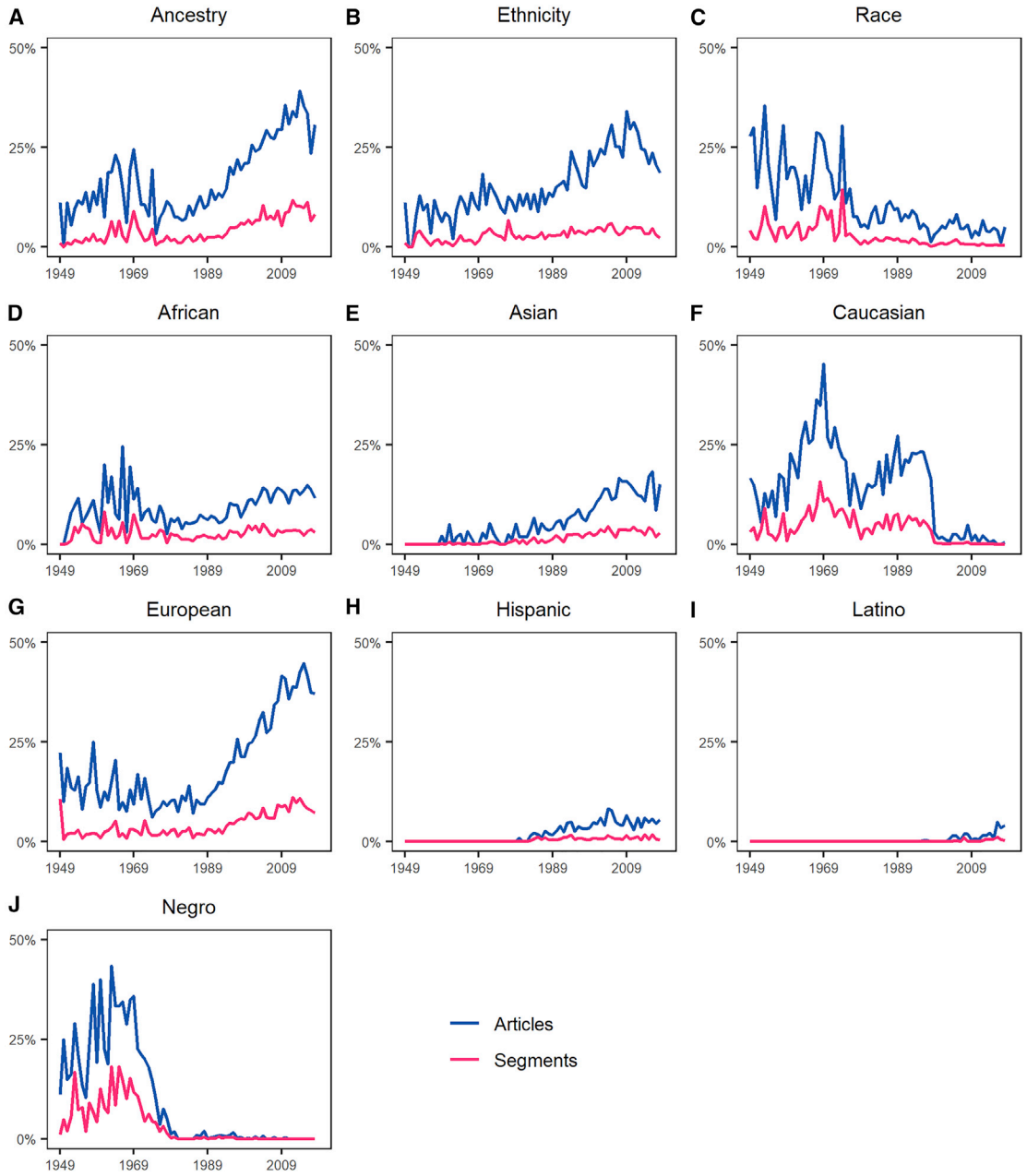


Figure 2. Percentage of 200-token segments containing "ancestry," "ethnicity," "race," and other population descriptors
(A–J) Yearly percentage of 200-token segments (pink) and articles (blue) containing "ancestry" (A), "ethnicity" (B), "race" (C), "African" (D), "Asian" (E), "Caucasian" (F), "European" (G), "Hispanic" (H), "Latina/o/x" (I), and "Negro" (J) from 1949–2018.

term s is a hypergeometric random variable with parameters n_s, n_t, n_{st} and the probability function:[21]

$$P(y) = \binom{n_t}{y}\binom{N-n_t}{n_s-y} \Big/ \binom{N}{n_s}$$

The mean of this distribution is the expected co-occurrence on a random basis and is given by $n_s n_t/N$. We compute the p value, i.e., the probability of the observed or a greater co-occurrence frequency arising by chance as the following:

$$p\ value = \sum_{y=n_{st}}^{\min(n_s,\ n_t)} P(y)$$

This calculation is equivalent to representing the presence or absence of two terms in a two-by-two contingency table and conducting a one-sided Fisher's exact test.[19]

We measured the "effect size" as the odds ratio (OR)[22]— the ratio of the odds of term t occurring in a segment where term s is present—to its odds of being present in the absence of s. The OR is given by $\frac{n_{st}(N-n_s-n_t+n_{st})}{(n_s-n_{st})(n_t-n_{st})}$, and N represents the total number of even or odd number of segments in a given decade and n the number of segments with s, t, or both (Figure 1). 95% confidence intervals (CIs) for the ORs are given by a well-known formula.[22] As p values and CIs were calculated with different distributions, ORs whose CIs include "1" can be statistically significant.

Table 2. Percentage of 200-token segments and articles containing population terms in 1949–58 and 2009–18

Term	Decade	Segments, % (n)	Articles, % (n)
Ancestry	1949–58	1% (48)	10% (31)
	2009–18	9% (2,585)	33% (721)
Ethnicity	1949–58	2% (56)	8% (25)
	2009–18	4% (1,150)	26% (571)
Race	1949–58	5% (149)	22% (69)
	2009–18	<1% (151)	4% (89)
African	1949–58	2% (74)	7% (23)
	2009–18	3% (905)	13% (288)
Asian	1949–58	0% (0)	0% (0)
	2009–18	3% (935)	14% (310)
Caucasian	1949–58	4% (131)	12% (38)
	2009–18	<1% (25)	<1% (22)
European	1949–58	2% (76)	15% (48)
	2009–18	9% (2,500)	40% (881)
Hispanic	1949–58	0% (0)	0% (0)
	2009–18	1% (287)	5% (112)
Latina/o/x	1949–58	0% (0)	0% (0)
	2009–18	<1% (128)	2% (45)
Negro	1949–58	7% (217)	21% (65)
	2009–18	<1% (2)	<1% (1)

Results

Frequency of use

The proportion of articles containing the population terms "ancestry," "ethnicity," "race," "African," "Asian," "Caucasian," "European," "Hispanic," "Latina/o/x," and "Negro" were calculated for each year from 1949–2018. The percent of articles that include "race" has declined since 1949 (Figure 2C), appearing in 22% of articles from 1949–58 and 5% in 2009–18 (Table 2). Conversely, the percent using "ancestry" and "ethnicity" have increased (Figures 2A and 2B). "Ancestry" increased in use from 10% of articles in 1949–58 to 33% in 2009–18. "Ethnicity" appeared in 8% of articles in 1949–58 and 26% in 2009–18 (Table 2). The continental terms "African," "Asian," and "European" have also increased in use, while the terms "Caucasian" and "Negro" have declined (Figures 2D–2J, Table 2). The proportion of articles containing the remaining 15 topical terms are shown in Figure S2. Yearly frequencies from 1949–2018, for both even and odd segments, are reported in Table S1.

Co-occurrence patterns

Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated for pairs of "ancestry," "ethnicity," and "race" for overlapping decades from 1949–2018. ORs have increased over time between "race" and "ethnicity," from 4.7 (CI 2.3, 9.5) in 1949–58 to 23.9 (CI 17.2, 33.1) in 2009–18. Ratios between "ancestry" and "race" and between "ancestry" and "ethnicity" have remained comparatively constant (Figure 3, Table 3). ORs were also generated between 15 topical terms and each of "ancestry," "ethnicity," and "race" (Figure 4, Figure S3) and between the seven population descriptors and each of "ancestry," "ethnicity," and "race" (Figure 5). All ORs, their 95% CIs, and p values for the observed co-occurrences between pairs of terms are given in Table S2 for even segments and Table S3 for odd segments.

Discussion

We have described the evolving usage of population terms in the 70-year publication history of *The American Journal of Human Genetics*. We find that from 1949–2018, the term "race" has declined in use, while increasing in co-occurrence with "ethnicity." At the same time, the use of "ancestry" and "ethnicity" has increased. We also describe changes in the use of specific population descriptors that may align with societal trends in their use outside of genetics.

The use of the term "race" in *The Journal* has consistently declined since 1949, while that of "ancestry" and

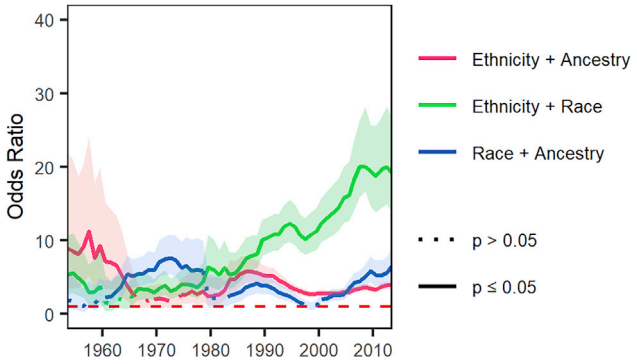


Figure 3. Odds ratios between "ancestry," "ethnicity," and "race"

ORs between "ethnicity" and "ancestry" (pink), "race" and "ancestry" (blue), and "race" and "ethnicity" (green) in overlapping decades from 1949–2018 (e.g., 1949–59, 1950–60, 1951–61 …). Each point on the line graph represents the value of the ratio for which the corresponding year is the midpoint (e.g., values at 1954 represent co-occurrence ratios for 1949–59). Solid line segments indicate decades where the number of co-occurrences was significantly greater than expected by chance, with $p \leq 0.05$. Dotted line segments indicate that the number of co-occurrences was not significantly greater than expected by chance. Shaded regions surrounding a curve and of the same color indicate 95% confidence intervals. For ease of viewing, upper confidence intervals are cut if they exceeded 40 (see Tables S2 and S3 for untruncated values). The horizontal red dashed line marks an OR of 1.0.

"ethnicity" have increased. These findings are consistent with those of Popejoy et al.'s survey of clinical geneticists in which participants reported ancestry, followed by ethnicity then race, as important to clinical variant interpretation and ordering genetic tests.[23] We hypothesize that as the field grows more cognizant of historical and ongoing debates about the use of race in genetics, ancestry and ethnicity may increasingly be perceived as more scientifically valid, historically neutral, or practically useful. This is not without its own criticisms, as we will discuss further below. We also found an increase in the odds ratio between "race" and "ethnicity" throughout the history of *The Journal*. This may be attributable to the increasing use of combined phrases such as "race/ethnicity" and "race and/or ethnicity," which have emerged as the distinction between the two concepts has become more ambiguous.

Furthermore, we report temporal changes in the use of specific population descriptors, adding support to the long-standing wisdom that population labels are not based on immutable biological order but shift in tandem with social context.[24] Along with the finding above that the use of "race" has declined, the labels "Caucasian"[25] and "Negro" have declined in *The Journal* over the past several decades. These terms, particularly in the form of "Caucasoid" and "Negroid," were used by 19th century race scientists and later by 20th century geneticists to refer to pseudoscientific biological race groups. "Hispanic" and "Latina/o/x" first appeared in *The Journal* in 1980 and 1996, respectively. Each of these changes in the use of population descriptors took place in a broader social context. For example, the decline of the term "Negro" can be connected not only to the discrediting of the idea of a "Negroid race" on scientific terms but also to African-descent Americans' efforts to reject or claim social identifiers in contexts outside of genetics.[26–28] Similarly, the adoption of "Hispanic" and "Latina/o/x" in genetics did not originate from within the field but from a convergence of commercial, activist, and government interests in creating a panethnic, institutionally recognized category from the diverse range of Latin American nationalities in the US.[29]

Some of the shifts described in this paper may signal constructive change. For example, the term "Caucasian," which has declined in use in *The Journal*, has been criticized for its historical connections to racist taxonomies and lack of scientific justification.[30] However, areas remain for continued investigation and critical reflection. For example, although the term "race" has declined, commentary in this area has pushed not necessarily for the complete removal of race from genetic and biomedical research but for a refocusing on racism and race as a social category with biological consequences.[4,11,12] Moreover, as numerous scholars have discussed, practices that racialize populations can persist in the sciences without explicit use of the term "race."[31–34] The continental population terms "African," "Asian," and "European," which we have shown are increasing in use in *The Journal*, have been critiqued for their resemblance to historical racial taxonomies and their inability to capture immense within-group heterogeneity.[8,35]

Table 3. Odds ratios between "ancestry," "ethnicity," and "race"

Term 1	Term 2	Decade	Odds ratio (95% CI)	p value
Ancestry	ethnicity	1949–58	8.2 (3.4, 20.1)	2.2×10^{-4}
		2009–18	3.9 (3.3, 4.4)	1.4×10^{-65}
Ancestry	race	1949–58	4.2 (2.0, 9.2)	1.5×10^{-3}
		2009–18	6.6 (4.8, 9.2)	1.3×10^{-23}
Ethnicity	race	1949–58	4.7 (2.3, 9.5)	2.0×10^{-4}
		2009–18	23.9 (17.2, 33.1)	1.3×10^{-60}

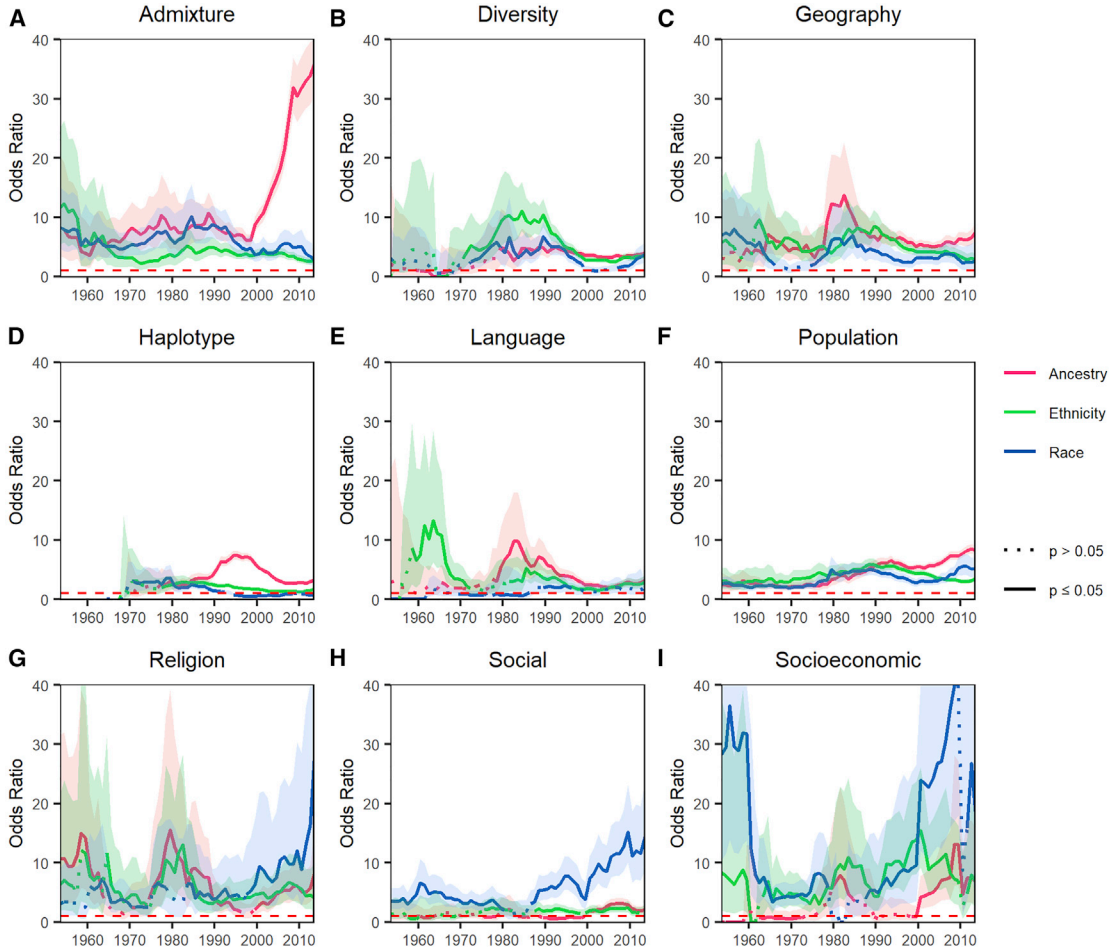


Figure 4. Odds ratios between "ancestry," "ethnicity," "race," and select topical terms
(A–I) ORs between "ancestry" (pink), "ethnicity" (green), and "race" (blue) and "admixture" (A), "diversity" (B), "geography" (C), "haplotype" (D), "language" (E), "population" (F), "religion" (G), "social" (H), and "socioeconomic" (I) in overlapping decades from 1949–2018 (e.g., 1949–59, 1950–60, 1951–61 …). Each point on the line graph represents the value of the ratio for which the corresponding year is the midpoint (e.g., values at 1954 represent co-occurrence ratios for 1949–59). Solid line segments indicate decades where the number of co-occurrences was significantly greater than expected by chance, with $p \leq 0.05$. Dotted line segments indicate that the number of co-occurrences was not significantly greater than expected by chance. Shaded regions surrounding a curve and of the same color indicate 95% confidence intervals. For ease of viewing, upper confidence intervals are cut if they exceeded 40 (see Tables S2 and S3 for untruncated values). The horizontal red dashed line marks an OR of 1.0.

This study has several limitations. First, we examined a single journal, and the trends we describe may not generalize to other contexts in the field. However, our analysis of the entire corpus of a single journal may be a strength relative to other studies of biomedical corpora, which tend to be limited to abstracts because of data availability. Second, we pre-selected a set of terms that we chose not to alter throughout the course of our analyses. As a result, we were limited in our ability to explore or discover new aspects of ancestry, ethnicity, and race that may deviate from our current biases about the concepts. We also could not examine many relevant descriptors such as "Black," "White," and "Native American," as these terms were either confounded by other meanings in the text or did not have high enough frequency in the dataset to conduct statistical analyses. Third, odds ratios were sensitive to the amount of data available, meaning that time periods with limited amounts of text or term uses were prone to large, not necessarily meaningful, fluctuations. Finally, although quantitative analyses of text are unique in their ability to detect patterns that are difficult through manual review, we recognize these methods' limited ability to provide insight into how our terms and concepts of interest were used qualitatively.

Nonetheless, our research has documented and quantitated historical changes in the use of population concepts in the entirety of *The Journal*'s text corpus. Our results can serve to motivate ongoing reflection as the concepts and population group labels used to study global genetic variation continues to evolve. Such reflection is critical to the field's ability to accurately describe human genetic variation and adopt new genomic methods in a way that is attentive to its troubled history with race.

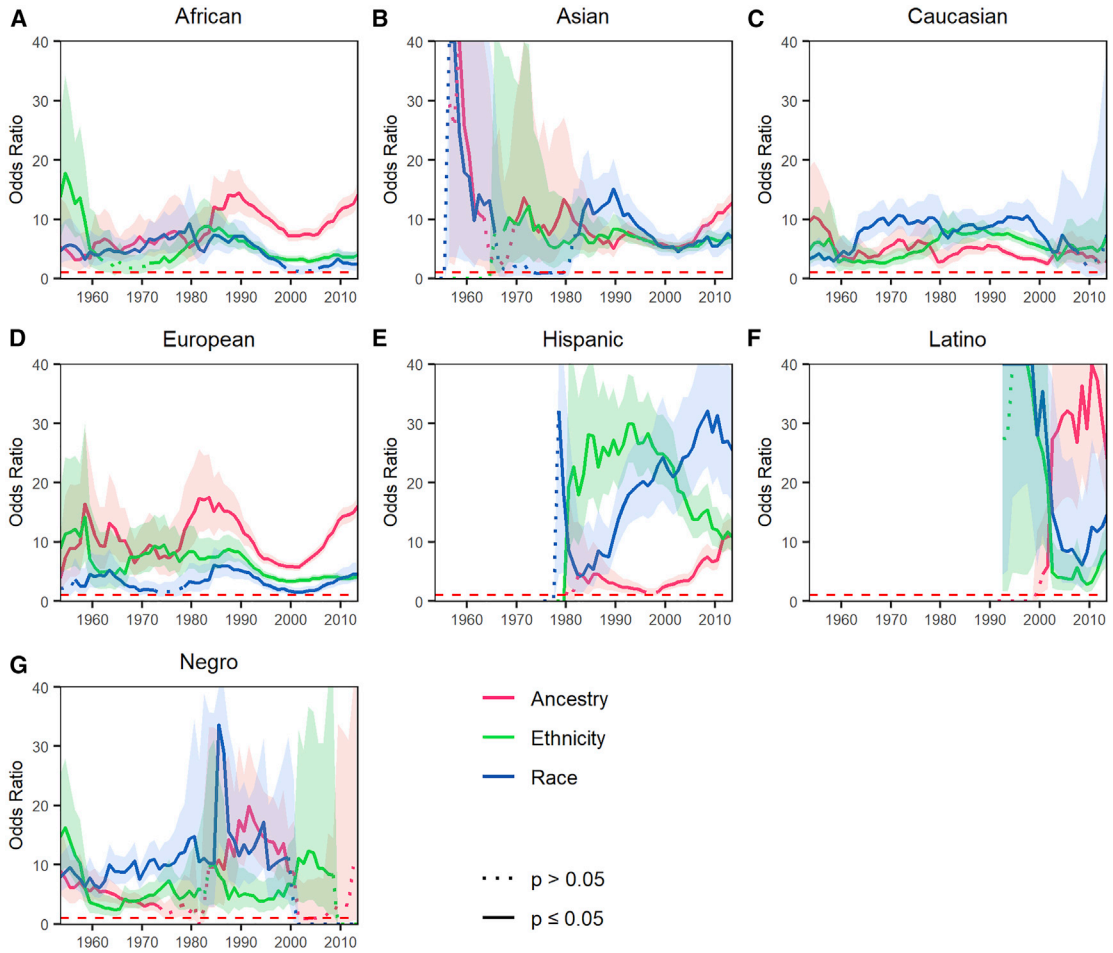


Figure 5. Odds ratios between "ancestry," "ethnicity," "race," and population descriptors

(A–G) ORs between "ancestry" (pink), "ethnicity" (green), and "race" (blue) and "African" (A), "Asian" (B), "Caucasian" (C), "European" (D), "Hispanic" (E), "Latina/o/x" (F), and "Negro" (G) in overlapping decades from 1949–2018 (e.g., 1949–59, 1950–60, 1951–61 …). Each point on the line graph represents the value of the ratio for which the corresponding year is the midpoint (i.e., values at 1954 represent co-occurrence ratios for 1949–59). Solid line segments indicate decades where the number of co-occurrences was significantly greater than expected by chance, with $p \leq 0.05$. Dotted line segments indicate that the number of co-occurrences was not significantly greater than expected by chance. Shaded regions surrounding a curve and of the same color indicate 95% confidence intervals. For ease of viewing, ORs and upper confidence intervals are cut if they exceeded 40 (see Tables S2 and S3 for untruncated values). The horizontal red dashed line marks an OR of 1.0.

Data and code availability

The code generated during this study and the data are available upon request. Please contact the corresponding authors for the data.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.10.008.

Acknowledgments

We would like to acknowledge the American Society of Human Genetics for sharing the *AJHG* archive for this research, Natalie Xie for contributions to data cleaning and the development of a graphical interface for data visualization, and Won Kim for assistance in data processing. This work was supported by the Intramural Research Programs of the National Center for Biotechnology Information and the National Human Genome Research Institute, National Institutes of Health.

Declaration of interests

The authors declare no competing interests.

Received: January 30, 2021
Accepted: October 20, 2021
Published: December 2, 2021

References

1. Veronika Lipphardt (2013). From "Races" to "Isolates" and "Endogamous Communities": Human Genetics and the Notion of Human Diversity in the 1950s. In Human Heredity

in the Twentieth Century, B. Gausemeier, S. Müller-Wille, and E. Ramsden, eds. (Pickering and Chatto), pp. 55–68.
2. Fujimura, J.H., and Rajagopalan, R. (2011). Different differences: the use of 'genetic ancestry' versus race in biomedical human genetic research. Soc. Stud. Sci. *41*, 5–30.
3. Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J., et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. Nature *586*, 683–692.
4. Brothers, K.B., Bennett, R.L., and Cho, M.K. (2021). Taking an antiracist posture in scientific publications in human genetics and genomics. Genet. Med. *23*, 1004–1007.
5. Khan, A., Mchugh, C., Conomos, M.P., Gogarten, S.M., and Nelson, S.C. (2020). Guidelines on the use and reporting of race, ethnicity, and ancestry in the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. arxiv, 2108.07858. https://arxiv.org/ftp/arxiv/papers/2108/2108.07858.pdf.
6. 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.
7. Mathieson, I., and Scally, A. (2020). What is ancestry? PLoS Genet. *16*, e1008624.
8. Weiss, K.M., and Long, J.C. (2009). Non-Darwinian estimation: my ancestors, my genes' ancestors. Genome Res. *19*, 703–710.
9. 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.
10. Panofsky, A., and Bliss, C. (2017). Ambiguity and Scientific Authority: Population Classification in Genomic Science. Am. Sociol. Rev. *82*, 59–87.
11. Yudell, M., Roberts, D., DeSalle, R., Tishkoff, S.; and 70 signatories (2020). NIH must confront the use of race in science. Science *369*, 1313–1314.
12. Yudell, M., Roberts, D., DeSalle, R., and Tishkoff, S. (2016). Taking race out of human genetics. Science *351*, 564–565.
13. Huddart, R., Fohner, A.E., Whirl-Carrillo, M., Wojcik, G.L., Gignoux, C.R., Popejoy, A.B., Bustamante, C.D., Altman, R.B., and Klein, T.E. (2019). Standardized Biogeographic Grouping System for Annotating Populations in Pharmacogenetic Research. Clin. Pharmacol. Ther. *105*, 1256–1262.
14. Bonham, V.L., Callier, S.L., and Royal, C.D. (2016). Will Precision Medicine Move Us beyond Race? N. Engl. J. Med. *374*, 2003–2005.
15. Harris, Z.S. (1954). Distributional Structure. Distrib. Struct. WORD *10*, 146–162.
16. Schütze, H., Manning, C.D., and Raghavan, P. (2008). Introduction to information retrieval (Cambridge University Press Cambridge).
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (Red Hook, NY, USA: Curran Associates Inc.), pp. 3111–3119.
18. Li, Y., and Yang, T. (2018). Word Embedding for Understanding Natural Language: A Survey (Cham: Springer), pp. 83–104.
19. Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics *19*, 61–74.
20. Terra, E., and Clarke, C.L.A. (2003). Frequency estimates for statistical word similarity measures (Association for Computational Linguistics), pp. 165–172.
21. Larson, H. (1982). Introduction to Probability Theory and Statistical Inference (New York: Wiley).
22. Tenny, S., and Hoffman, M.R. (2020). Odds Ratio (Treasure Island, FL: StatPearls Publishing).
23. 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.
24. Morning, A. (2014). Race and its Categories in Historical Perspective (Brooklyn Hist. Soc. Crossing Borders, Bridg. Gener).
25. Mukhopadhyay, C.C. (2018). Getting rid of the word "caucasian.". In Privilege (Routledge), pp. 231–236.
26. Grant, R., and Orr, M. (1996). Language, Race and Politics: From "Black" to "African-American.". Polit. Soc. *24*, 137–152.
27. Martin, B.L. (1991). From Negro to Black to African American: The Power of Names and Naming. Polit. Sci. Q. *106*, 83–107.
28. Smith, T.W. (1992). Changing Racial Labels: From "Colored" to "Negro" to "Black" to "African American.". Public Opin. Q. *56*, 496–514.
29. Mora, G.C. (2014). Making Hispanics: How Activists, Bureaucrats, and Media Constructed a New American (University of Chicago Press).
30. Popejoy, A.B. (2021). Too many scientists still say Caucasian. Nature *596*, 463.
31. Roberts, D. (2011). Redefining Race in Genetic Terms. In Fatal Invention: How Science, Politics, and Big Business Re-Create Race in the Twenty-First Century (The New Press), pp. 58–80.
32. Fullwiley, D. (2008). The biologistical construction of race: 'admixture' technology and the new genetic medicine. Soc. Stud. Sci. *38*, 695–735.
33. Fujimura, J.H., and Rajagopalan, R.M. (2020). Race, ethnicity, ancestry, and genomics in Hawai'i: Discourses and practices. Hist. Stud. Nat. Sci. *50*, 596–623.
34. Duello, T.M., Rivedal, S., Wickland, C., and Weller, A. (2021). Race and Genetics vs. 'Race' in Genetics: A Systematic Review of the Use of African Ancestry in Genetic Studies (Evol. Med. Public Heal).
35. Rajagopalan, R., and Fujimura, J. (2012). Making history via DNA, making DNA from history: Deconstructing the race-disease connection in admixture mapping. In Genetics and the Unsettled Past: The Collision of DNA, Race, and History, K. Wailoo, A. Nelson, and C. Lee, eds. (Rutgers University Press), pp. 143–163.

HHS Public Access

Author manuscript

Clin Pharmacol Ther. Author manuscript; available in PMC 2020 May 01.

Published in final edited form as:

Clin Pharmacol Ther. 2019 May ; 105(5): 1256–1262. doi:10.1002/cpt.1322.

Standardized biogeographic grouping system for annotating populations in pharmacogenetic research

Rachel Huddart[1,8], **Alison E. Fohner**[1,2,8], **Michelle Whirl-Carrillo**[1,8], **Genevieve L. Wojcik**[1], **Christopher R. Gignoux**[3], **Alice B. Popejoy**[1,4], **Carlos D. Bustamante**[1,5], **Russ B. Altman**[1,5,6,7], and **Teri E. Klein**[1,7,*]

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA

[2]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA

[3]Division of Bioinformatics and Personalized Medicine, and Department of Biostatistics, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA

[4]Stanford Center for Integration of Research on Genetics and Ethics, Stanford, CA 94305, USA

[5]Department of Genetics, Stanford University, Stanford, CA 94305, USA

[6]Department of Biomedical Engineering, Stanford University, Stanford, CA 94305, USA

[7]Department of Medicine, Stanford University, Stanford, CA 94305, USA

[8]These authors contributed equally to this work

Abstract

The varying frequencies of pharmacogenetic alleles between populations have important implications for the impact of these alleles in different populations. Current population grouping methods to communicate these patterns are insufficient as they are inconsistent and fail to reflect the global distribution of genetic variability. To facilitate and standardize the reporting of variability in pharmacogenetic allele frequencies, we present seven geographically-defined groups: American, Central/South Asian, East Asian, European, Near Eastern, Oceanian, and Sub-Saharan African, and two admixed groups: African American/Afro-Caribbean and Latino. These nine groups are defined by global autosomal genetic structure and based on data from large-scale sequencing initiatives. We recognize that broadly grouping global populations is an oversimplification of human diversity and does not capture complex social and cultural identity. However, these groups meet a key need in pharmacogenetics research by enabling consistent

[*]**Correspondence** Dr. Teri E. Klein, Shriram Center for BioE & ChemE, 443 Via Ortega, Stanford, CA 94305-4125, Phone: (650) 736-0156, feedback@pharmgkb.org.

Author Contributions

R.H., A.E.F., M.W-C., G.L.W., C.R.G., A.B.P., C.D.B., R.B.A., and T.E.K. wrote the manuscript; A.E.F., M.W-C., C.R.G., and T.E.K. designed the research; M.W-C., G.L.W., and C.R.G. analyzed the data.

Conflicts of Interest

CRG owns stock in 23andMe, Inc and is a founder of and advisor to Encompass Bioscience, Inc. CDB is a member of the scientific advisory boards for Liberty Biosecurity, Personalis, 23andMe Roots into the Future, Ancestry.com, IdentifyGenomics, and Etalon and is a founder of CDB Consulting. RBA is a stockholder in Personalis Inc. and 23andMe, and a paid advisor for Youscript. Remaining authors have no conflicts of interest.

communication of the scale of variability in global allele frequencies and are now used by PharmGKB.

Keywords

Pharmacogenetics; pharmacogenomics; PharmGKB; population groups

Introduction

Interindividual variability in pharmacogenes has important consequences for drug efficacy and toxicity.(1, 2) Unlike the low frequencies of alleles that are considered actionable with respect to disease risk, pharmacogenetic variants with clinical relevance are common and, in fact, both presence and absence of variants provide valuable dosing information.(3, 4) The frequencies of many pharmacogenetic alleles vary greatly by global population, meaning that people with different ancestries can have considerably different likelihoods of carrying an allele that is associated with a particular drug response. For example, the CYP3A5*3 allele has been found at a frequency of 98% in an Iranian population but at 11% in a Ngoni population from Malawi. (5, 6) A single value for global allele frequency would fail to reflect this pattern. Presenting the differences in frequencies of pharmacogenetic alleles is important for communicating the scale of their expected impact on drug response and the degree of variation between populations. This information is invaluable for furthering pharmacogenetic research and implementation.

Many pharmacogenetic studies present allelic data for very specific populations, such as from a single country or ethnic group, which are difficult to incorporate into broader research or implementation. Literature curation and gene summaries, such as those from the Pharmacogenomics Knowledgebase (PharmGKB: www.pharmgkb.org), must group these specific populations when annotating pharmacogenetic studies to allow users to easily compare information from multiple studies. As such, tagging studies with population group identifiers is an important component of knowledge extraction from curated literature. These population group labels then are used in aggregating and evaluating overall evidence for gene-drug associations, which eventually inform clinical implementation guidelines, such as those of the Clinical Pharmacogenetics Implementation Consortium (CPIC: www.cpicpgx.org).

Similar to other areas of biomedical research, (7) current methods for grouping global populations in pharmacogenetics are based on subjective, vague, and inconsistent geographical boundaries, or on populations that are geographically straightforward to cluster and reflect little admixture.(8–12) As an example of the issues with current grouping methods, some studies cluster participants of Egyptian descent with African populations, while others cluster them with Middle Eastern populations.(13, 14) While this discrepancy illustrates inconsistencies of geographic borders, the clustering of African-descent populations of the Americas with populations from Africa, as seen in the 1000 Genomes African (AFR) superpopulation, provides another example of challenges posed by employing a small number of categories to describe a broad spectrum of genomically diverse

groups. The genetic patterns seen in American populations with African ancestry differs dramatically from populations in Africa due to admixture primarily with European and American Indian populations. (15–17) While sharing common ancestry, the recent admixture typically observed in the Americas can complicate average allele frequency estimation or, at a minimum, make these combined groupings less homogeneous.(16) These insufficient grouping systems, often ad-hoc and not fully representative evidence from population genomic studies, create a barrier to understanding and interpreting pharmacogenetic allele frequencies in a globally representative fashion.

Until July 2018, PharmGKB annotated studies using the five race categories defined by the US Office of Management and Budget (OMB): White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Pacific Islander, with an additional ethnicity OMB category of Hispanic/Latino. While PharmGKB serves as a global resource, these OMB groups are US-centric and, as socio-cultural measures of identity, lack the capacity to capture the scale of global human diversity. We also investigated the utility of the biogeographic categories employed by the Human Genome Diversity Panel - Centre d'Etude du Polymophisme Humain (HGDP - CEPH), which groups its 52 populations into Africa, Europe, Middle East, South and Central Asia, East Asia, Oceania and the Americas. (8, 18, 19) These population labels work well for the populations included in the HGDP data set, which are not located in ambiguous. However, papers curated at PharmGKB can include populations located all over the world, including in the transitional zones between HGDP geographical regions and admixed populations. This leads to ambiguity in how such populations would be grouped using HGDP categories. In conclusion, existing systems are insufficient for capturing the diversity of study populations in a replicable manner that is consistent with patterns of human genetic variation.

Therefore, we sought to define a grouping system of global populations that could be used consistently to annotate pharmacogenetic studies and relevant alleles, and could capture global human population genetic patterns. Using population genetics data sources, including the 1000 Genomes Phase 3 data release and the HGDP, we propose a simple and robust grouping pattern based on nine broad biogeographic regions that represent major geographic regions of the world (Figure 1). It is important to note that classifying individuals and communities into a few distinct groups with defined boundaries conflicts with our understanding of human variation, history, and social/cultural identities. *As a result, we respectfully present these groups as a tool to represent broad differences in frequencies of pharmacogenetic variation rather than as a classification of human diversity*.

Results

We chose this geographic clustering pattern because geography has historically been the greatest predictor of genetic variation between human populations, with genetic distance increasing as geographic distance increases.(20) This geographic pattern aids consistency in population groupings by setting boundaries along national borders. To simplify utility, geographic boundaries between groupings are drawn predominantly along country borders, with only Russia divided into east and west along the Ural Mountains boundary due to the large size and genetic heterogeneity of the country. We intend these groups to represent

peoples with a predominance of ancestors who were in the region pre-Diaspora and pre-colonization.

We have also included two admixed groups representing populations with recent gene flow between geographically-based populations and therefore, have distinct genetic patterns which are not adequately reflected by any single geographically-based group. (7) While many populations reflect a degree of admixture, we selected these two populations because they are frequently reported in pharmacogenetic studies.

We consider these nine groups sufficient to better illustrate the broad diversity in global allele frequencies, yet small enough to apply easily and to be tractable in grouping specific populations.(21–24) The groups are given below with their abbreviations.

Geographical populations

American (AME): The American genetic ancestry group includes populations from both North and South America with ancestors predating European colonization, including American Indian, Alaska Native, First Nations, Inuit, and Métis in Canada, and Indigenous peoples of Central and South America.

Central/South Asian (SAS): The Central and South Asian genetic ancestry group includes populations from Pakistan, Sri Lanka, Bangladesh, India, and ranges from Afghanistan to the western border of China.

East Asian (EAS): The East Asian genetic ancestry group includes populations from Japan, Korea, and China, and stretches from mainland Southeast Asia through the islands of Southeast Asia. In addition, it includes portions of central Asia and Russia east of the Ural Mountains.

European (EUR): The European genetic ancestry group includes populations of primarily European descent, including European Americans. We define the European region as extending west from the Ural Mountains and south to the Turkish and Bulgarian border.

Near Eastern (NEA): The Near Eastern genetic ancestry group encompasses populations from northern Africa, the Middle East, and the Caucasus. It includes Turkey and African nations north of the Saharan Desert.

Oceanian (OCE): The Oceanian genetic ancestry group includes pre-colonial populations of the Pacific Islands, including Hawaii, Australia, and Papua New Guinea.

Sub-Saharan African (SSA): The Sub-Saharan African genetic ancestry group includes individuals from all regions in Sub-Saharan Africa, including Madagascar.(25)

Admixed populations

African American/Afro-Caribbean (AAC): Individuals in the African American/Afro-Caribbean genetic ancestry group reflect the extensive admixture between African, European, and Indigenous ancestries(26) and, as such, display a unique genetic profile

compared to individuals from each of those lineages alone. Examples within this cluster include the Coriell Institute's African Caribbean in Barbados (ACB) population and the African Americans from the Southwest US (ASW) population, (27) and individuals from Jamaica and the US Virgin Islands.

Latino (LAT): The Latino genetic ancestry group is not defined by an exclusive geographic region, but includes individuals of Mestizo descent, individuals from Latin America, and self-identified Latino individuals in the United States. Like the African American/Afro-Caribbean group, the admixture in this population creates a unique genetic pattern compared to any of the discrete geographic regions, with individuals reflecting mixed Native and Indigenous American, European, and African ancestry.

The Central/South Asian, East Asian and European groups presented here are equivalent to the 1000 Genomes South Asian (SAS), East Asian (EAS) and European (EUR) super populations, respectively. As such, we have adopted the relevant 1000 Genomes super population codes as abbreviations for each of these groups to maintain consistency. While the 1000 Genomes Ad Mixed American (AMR) super population shows complete overlap with the Latino group, we have opted to use the abbreviation LAT for this group. This removes the potential for confusion between the Latino group and the other admixed group of African American/Afro-Caribbean.

Figure 1 illustrates the countries included in each of the seven geographical groups and removes any ambiguity of the group boundaries. As this map shows the boundaries of each group pre-colonization and pre-Diaspora, the two admixed groups, African American/Afro-Caribbean and Latino are not shown. We intend this map to be used as a guide for grouping genetic ancestral populations. Study subjects of an ancestry that is not within the geographic cluster in which they currently live will be included in the geographic cluster reflecting their ancestry. For example, South Africans of Dutch descent would be included in the European cluster rather than the Sub-Saharan African cluster. However, when lacking a clear description otherwise, the population will be included in the group that includes its home country.

This approach highlights the importance of understanding and recording detailed self-identified and self-reported race and ethnicity in the context of genetic studies. While self-reported race and ethnicity can be influenced by an individual's social and cultural background and thus may not perfectly correlate with genetic ancestry (28), it is more reliable than assignment of race or ethnicity by another person (e.g. a healthcare professional) (29). However, it should be noted that self-reported measures can be complicated by collection processes, (30) including an incomplete selection of possible identity categories, or allowing only one selection and thus failing to capture whether an individual may identify with multiple categories or none at all (29). These classification limitations can be particularly prevalent among populations with a high degree of admixture.

To validate the genetic variability distinguished by these population groups, we conducted Principal Components Analysis (PCA) using autosomal genotype data of unrelated individuals from 1000 Genomes and HGDP. As seen in Figure 2A, the first two principal

components (PCs) separate populations by geographic region, especially along continental boundaries, and illustrate the increasing genetic distance between populations of increasing geographic distance. As can be seen in the overlapping PC distribution of individuals of different population groups, human genetic diversity is a spectrum,(19) and therefore the geographic boundaries of these groups should be understood as an obligatory divide to create relevant groupings, with the acknowledgement that these borders are constrained by modern country borders and therefore are inherently arbitrary in geographic space.(19) However, as shown in Figure 2B, only a few PCs are needed to accurately predict these population clusters. Even with only 4 PCs, the minimum area under the curve (AUC) for correct cluster prediction is 97.9% for most populations using multiple logistic regression. The only outlier is the African American/Afro-Caribbean cluster, consistent with ancestral similarity to the African cluster.(15, 31) Here still, with a larger number of PCs, the AUC is above 93%, even with the observed ancestry outliers present in the 1000 Genomes African Americans in the Southwest US (ASW) population.(32) While no categorization will result in perfect prediction, given the spectrum of human diversity, the statistical validation of this clustering from broad autosomal data makes these clusters both relevant and useful for PharmGKB.

In Figure 3, we demonstrate that the groups we have selected are effective for representing the diversity of global allele frequencies in pharmacogenes. We present here the frequency of four single nucleotide polymorphisms (SNPs) with important pharmacogenetic implications. The ‘A’ allele of rs1065852 is the defining SNP of the *cytochrome P450 2D6 (CYP2D6) *10* haplotype and is also found in combination with other variants in multiple CYP2D6 haplotypes. Haplotypes containing this SNP are associated with decreased CYP2D6 activity, which has important implications for drugs that are CYP2D6 substrates, including codeine, selective serotonin reuptake inhibitors, ondansetron, and tricyclic antidepressants.(33–36) The *CYP2C9* alleles **2* (defined by rs1799853), **3* (defined by rs1057910), and **8* (defined by rs7900194) are associated with reduced enzyme function and therefore are associated with recommended changes to the dosing of warfarin and phenytoin, which are substrates of CYP2C9.(37, 38) Using data from the 1000 Genomes, we show the frequency of the four SNPs in these biogeographic groups. The range of frequencies between populations illustrates the importance of showing allele frequency by group in order to convey its impact on drug response globally.

The SNP rs1065852 shows stark continental patterns (Figure 3A). The ‘A’ allele is found at high frequencies within East Asian populations, ranging from 66.2% in Vietnam (KHV) to 36.1% in Japan (JPT). This allele is less frequent in other continental populations, such as Sub-Saharan African (3.5–16.5%), European (14.6–24.7%), and Central/South Asian (10.4–25.6%). As can be seen from the range of frequencies of the three *CYP2C9* alleles, the most common reduced function allele varies globally, with the **8* allele much more common in Sub-Saharan African populations (1.8–7.6%) than the **2* (<1%) or **3* (monomorphic in Africa) (Figure 3B-D). Conversely, the **8* allele is rare in European populations (<1%), while **2* (8.1–15.2%) and **3* (5.6–8.4%) are more common. Patterns such as this one can result in bias in the utility of dosing algorithms, such as the International Warfarin Pharmacogenetics Consortium (IWPC) dosing algorithm for warfarin, which adjusts dose based on the presence of the **2* and **3* alleles but does not include the **8* allele.(39)

Discussion

While individual pharmacogenetic testing (either pre-emptive or at point-of-care) remains the most effective and appropriate way to implement pharmacogenetic knowledge for the care of an individual,(40, 41) we recognize the need in clinical and genetic research for a standardized method to broadly group populations based on biogeographic region. For example, identifying populations with high frequencies of certain pharmacogenetic alleles can help to direct targeted screening when resources are constrained and inform priorities for future pharmacogenetic research.(20) However, the groups we present are large and the summary information presented should be understood as an approximation dependent on existing studies in that region, which may be limited to a few locations. As such, these groups are not suitable for use in guiding specific implementation programs; rather, they should be seen as a tool for research purposes.

It should be noted that this grouping system does have limitations. Classifying individuals into these population groups can be complicated by social and cultural identities(8, 10, 42–44) and membership of an individual within one of these population groups is inherently an imperfect surrogate for predicting the likelihood that the individual carries a particular genetic variant.(41, 45) As can be seen in the analysis of rs1065852 above, the frequency of the ‘A’ allele can vary by up to 30% between populations which are all included in the East Asian group. Furthermore, while the grouping system is based on overall genome-wide average patterns, which typically follow a clinal variation pattern correlated with geographic proximity,(8, 23, 24, 46, 47) variation in individual genes or individual populations do not always follow these gradual patterns.(9–12, 41) In an attempt to mitigate some of these limitations, we encourage researchers using this grouping system to also provide specific details regarding the geographical and racial or ethnic origins of their subjects.

Because aggregate annotations of pharmacogenetic research and summary allele frequencies are based only on available studies, additional studies are needed that include a greater diversity of populations to make pharmacogenetic research and allele frequency summaries more representative.(48) For example, the Sub-Saharan African (SSA) grouping represents a large swath of human genomic diversity, which is not adequately represented in the available data from HGDP and 1000 Genomes. Increased representation of these populations in pharmacogenetics studies may lead to the discovery of clinical differences within the larger grouping. Furthermore, large, reference genetic studies with targeted allele information, like that emerging from the Population Architecture using Genomics and Epidemiology (PAGE) study (www.pagestudy.org), may provide compelling evidence to adjust these group boundaries based on frequency patterns specific to pharmacogenetic alleles. Continued evolution of this grouping system will be key to ensuring that misclassification of individuals is kept to a minimum. However, it should be understood that some misclassification is inevitable and will only be truly avoided when every patient can access comprehensive pharmacogenetic testing.

Despite these limitations, broad population groups are needed for illustrating global diversity with respect to pharmacogenetic variation and the average predicted phenotypes in populations. These nine proposed biogeographic groups provide a consistent way to present

these data based on a system that is grounded in robust data on population genetic patterns, and their introduction is particularly timely given the recent commentaries by Bonham *et al.* and Cooper *et al.* (7, 49) PharmGKB is now using these population groups in curation activities, and we recommend that these groups and accompanying map be considered the standard grouping mechanism for population pharmacogenetics. Ultimately, individual pharmacogenetic testing of all patients, regardless of ancestry, is needed to deliver truly personalized medicine. However, the population groups we present are useful for the standardized presentation of pharmacogenetic studies, global allele frequency summaries in pharmacogenetic research and broad clinical screening.

Methods

The MVN joint callset for 1000 Genomes data Phase 3 (21) was downloaded directly form the website for downstream interpretation. For principal component analysis (PCA), we filtered sites with a MAF < 0.5% and thinned sites given windows of 100 kilobases or 10 variants and r2>0.2, resulting in 156,211 sites. PCA was performed in PLINK 1.9 (50). Forward stepwise logistic regression was subsequently performed, adding 1 PC at a time, to predict population labels in a bivariate fashion. Prediction accuracy was assessed using the AUC-ROC estimator, as included in the R package ‘epicalc.’ To make assessments transparent, we included all individuals with specific population labels, although it has been demonstrated in multiple venues that there are several known ancestry outliers within 1000 Genomes populations of the Americas (17, 32). Plots were performed in R and ggplot2.

Acknowledgments

Funding Information

This work was funded by NIH/NIGMS R24 GM61374 and NIH/NHGRI U01 HG007419–04.

References

(1). Roden DM PHarmacogenomics: Challenges and Opportunities. Annals of Internal Medicine 145, (2006).

(2). Dunnenberger HM et al. Preemptive clinical pharmacogenetics implementation: current programs in five US medical centers. Annu Rev Pharmacol Toxicol 55, 89–106 (2015). [PubMed: 25292429]

(3). Tabor HK et al. Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. Am J Hum Genet 95, 183–93 (2014). [PubMed: 25087612]

(4). Wright GEB, Carleton B, Hayden MR & Ross CJD The global spectrum of protein-coding pharmacogenomic diversity. The pharmacogenomics journal 18, 187–95 (2018). [PubMed: 27779249]

(5). Rahsaz M et al. Association between tacrolimus concentration and genetic polymorphisms of CYP3A5 and ABCB1 during the early stage after liver transplant in an Iranian population. Experimental and clinical transplantation : official journal of the Middle East Society for Organ Transplantation 10, 24–9 (2012). [PubMed: 22309416]

(6). Bains RK et al. Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa. BMC genetics 14, 34 (2013). [PubMed: 23641907]

(7). Bonham VL, Green ED & Pérez-Stable EJ Examining How Race, Ethnicity and Ancestry Data Are Used in Biomedical Research. JAMA **Epub September 24, 2018**, (2018).**Epub September 24, 2018**

(8). Rosenberg NA et al. Genetic structure of human populations. Science (New York, NY) 298, 2381–5 (2002).

(9). Rajagopalan R & Fujimura JH Will personalized medicine challenge or reify categories of race and ethnicity? The virtual mentor : VM 14, 657–63 (2012). [PubMed: 23351323]

(10). Gannett L Group Categories in Pharmacogenetics Research. Philosophy of Science 72, 1232–47 (2005).

(11). Wilson JF et al. Population genetic structure of variable drug response. Nat Genet 29, 265–9 (2001). [PubMed: 11685208]

(12). Race E, and Genetics Working Group. The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research. Am J Hum Genet 77, 519–32 (2005). [PubMed: 16175499]

(13). Relling MV et al. Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clin Pharmacol Ther 89, 387–91 (2011). [PubMed: 21270794]

(14). Scott SA et al. Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. Clin Pharmacol Ther 94, 317–23 (2013). [PubMed: 23698643]

(15). Bryc K et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America 107, 786–91 (2010). [PubMed: 20080753]

(16). Mathias RA et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nature communications 7, 12522 (2016).

(17). Martin AR et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 100, 635–49 (2017). [PubMed: 28366442]

(18). Cann HM et al. A human genome diversity cell line panel. Science (New York, NY) 296, 261–2 (2002).

(19). Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK & Feldman MW Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1, e70 (2005). [PubMed: 16355252]

(20). Burchard EG et al. The importance of race and ethnic background in biomedical research and clinical practice. The New England journal of medicine 348, 1170–5 (2003). [PubMed: 12646676]

(21). Auton A et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). [PubMed: 26432245]

(22). Elhaik E et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nature communications 5, 3513 (2014).

(23). Jakobsson M et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451, 998–1003 (2008). [PubMed: 18288195]

(24). Li JZ et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, NY) 319, 1100–4 (2008).

(25). Hurles ME, Sykes BC, Jobling MA & Forster P The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages. Am J Hum Genet 76, 894–901 (2005). [PubMed: 15793703]

(26). Maples BK, Gravel S, Kenny EE & Bustamante CD RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet 93, 278–88 (2013). [PubMed: 23910464]

(27). Genomes Project C et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012). [PubMed: 23128226]

(28). Mersha TB & Abebe T Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Human genomics 9, 1 (2015). [PubMed: 25563503]

(29). Magana Lopez M, Bevans M, Wehrlen L, Yang L & Wallen GR Discrepancies in Race and Ethnicity Documentation: a Potential Barrier in Identifying Racial and Ethnic Disparities. Journal of racial and ethnic health disparities, (2016).

(30). Shraga R et al. Evaluating genetic ancestry and self-reported ethnicity in the context of carrier screening. BMC genetics 18, 99 (2017). [PubMed: 29179688]

(31). Baharian S et al. The Great Migration and African-American Genomic Diversity. PLoS Genet 12, e1006059 (2016). [PubMed: 27232753]

(32). Mimno D, Blei DM & Engelhardt BE Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. Proceedings of the National Academy of Sciences of the United States of America 112, E3441–50 (2015). [PubMed: 26071445]

(33). Bell GC et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 genotype and use of ondansetron and tropisetron. Clin Pharmacol Ther, (2016).

(34). Hicks JK et al. Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update. Clin Pharmacol Ther, (2016).

(35). Hicks JK et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and CYP2C19 Genotypes and Dosing of Selective Serotonin Reuptake Inhibitors. Clin Pharmacol Ther 98, 127–34 (2015). [PubMed: 25974703]

(36). Crews KR et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. Clin Pharmacol Ther 95, 376–82 (2014). [PubMed: 24458010]

(37). Caudle KE et al. Clinical pharmacogenetics implementation consortium guidelines for CYP2C9 and HLA-B genotypes and phenytoin dosing. Clin Pharmacol Ther 96, 542–8 (2014). [PubMed: 25099164]

(38). Johnson JA et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. Clin Pharmacol Ther, (2017).

(39). International Warfarin Pharmacogenetics C et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. The New England journal of medicine 360, 753–64 (2009). [PubMed: 19228618]

(40). Foster MW, Sharp RR & Mulvihill JJ Pharmacogenetics, Race, and Ethnicity: Social Identities and Individualized Medical Care. Therapeutic Drug Monitoring 23, 232–8 (2001). [PubMed: 11360031]

(41). Yen-Revollo JL, Auman JT & McLeod HL Race does not explain genetic heterogeneity in pharmacogenomic pathways. Pharmacogenomics 9, 1639–45 (2008). [PubMed: 19018720]

(42). Braun L et al. Racial categories in medical practice: how useful are they? PLoS medicine 4, e271 (2007). [PubMed: 17896853]

(43). Ortega VE & Meyers DA Pharmacogenetics: implications of race and ethnicity on defining genetic profiles for personalized medicine. J Allergy Clin Immunol 133, 16–26 (2014). [PubMed: 24369795]

(44). Bamshad M, Wooding S, Salisbury BA & Stephens JC Deconstructing the relationship between genetics and race. Nat Rev Genet 5, 598–609 (2004). [PubMed: 15266342]

(45). Urban TJ Race, ethnicity, ancestry, and pharmacogenetics. Mt Sinai J Med 77, 133–9 (2010). [PubMed: 20309922]

(46). Risch N, Burchard E, Ziv E & Tang H Categorization of humans in biomedical research: genes, race and disease. Genome Biology 3, (2002).

(47). Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA & Jorde LB Human population genetic structure and inference of group membership. Am J Hum Genet 72, 578–89 (2003). [PubMed: 12557124]

(48). Bustamante CD, Burchard EG & De la Vega FM Genomics for the world. Nature 475, 163–5 (2011). [PubMed: 21753830]

(49). Cooper RS, Nadkarni GN & Ogedegbe G Race, Ancestry, and Reporting in Medical Journals. JAMA **Epub September 24, 2018**, (2018).**Epub September 24, 2018**

(50). Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM & Lee JJ Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 7 (2015). [PubMed: 25722852]

Study Highlights

What is the current knowledge on the topic?

The frequency of pharmacogenetic alleles can very significantly between different populations around the world. Grouping populations can simplify reporting of pharmacogenetic alleles but current methods used to group populations are inadequate and are applied inconsistently.

What question did this study address?

Can we improve how populations are grouped for the reporting of pharmacogenetic alleles?

What does this study add to our knowledge?

We present nine new biogeographical groups based on geographical location or recent genetic admixture for use in pharmacogenetic research. These groups have been validated using autosomal genetic data from large-scale sequencing initiatives.

How might this change clinical pharmacology or translational science?

These groups have already been adopted for use in curation activities at PharmGKB. It is hoped that use of these groups will become standard in pharmacogenetics research.

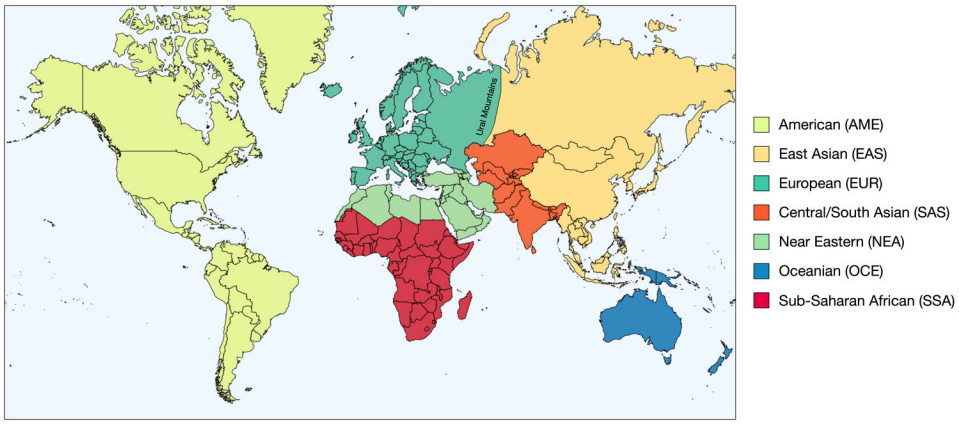


Figure 1: Map of geographical boundaries included in each geographical population group.

Group boundaries for the seven geographical groups fall predominantly along national boundaries to aid the assignment of group membership. The two admixed groups of African American/Afro-Caribbean and Latino are not shown on this figure as the map indicates the borders of each geographical group based on the location of genetic ancestors pre-Diaspora and pre-colonization, which cannot be applied to the two admixed groups. It should also be recognized that, due to the large geographical areas covered by each group, a single group does not accurately represent the large amount of genetic diversity found in that one region.

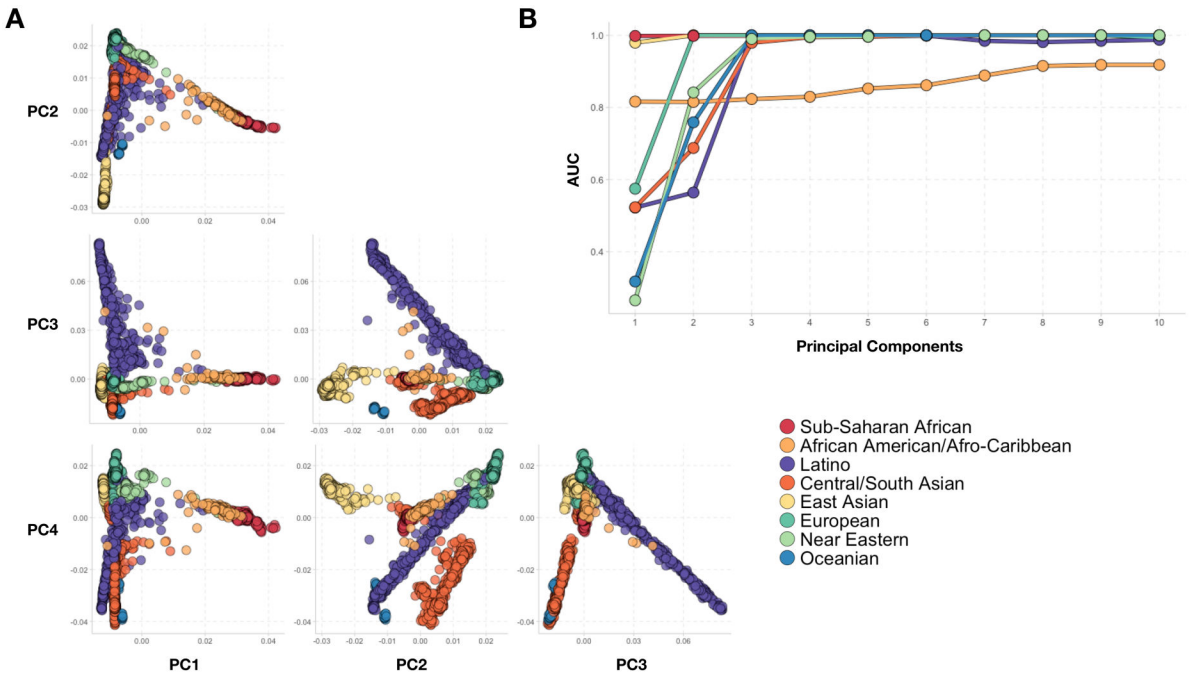


Figure 2: Principal component analysis comparing genetic distances of populations with close geographic proximity using 1000 Genomes and HGDP participants.

(A) The genetic gradient between populations is illustrated along PCs 1 vs 2 and PCs 3 vs 4, showing that, while completely discrete population boundaries are challenging, the groupings proposed here provide a statistically robust grouping. (B) AUCs of logistic regression to predict cluster membership, showing high degree of population structure. Note that, because none of the 1000 Genomes populations fall into the American (AME) group, no reference data were available to include this group in the analysis.

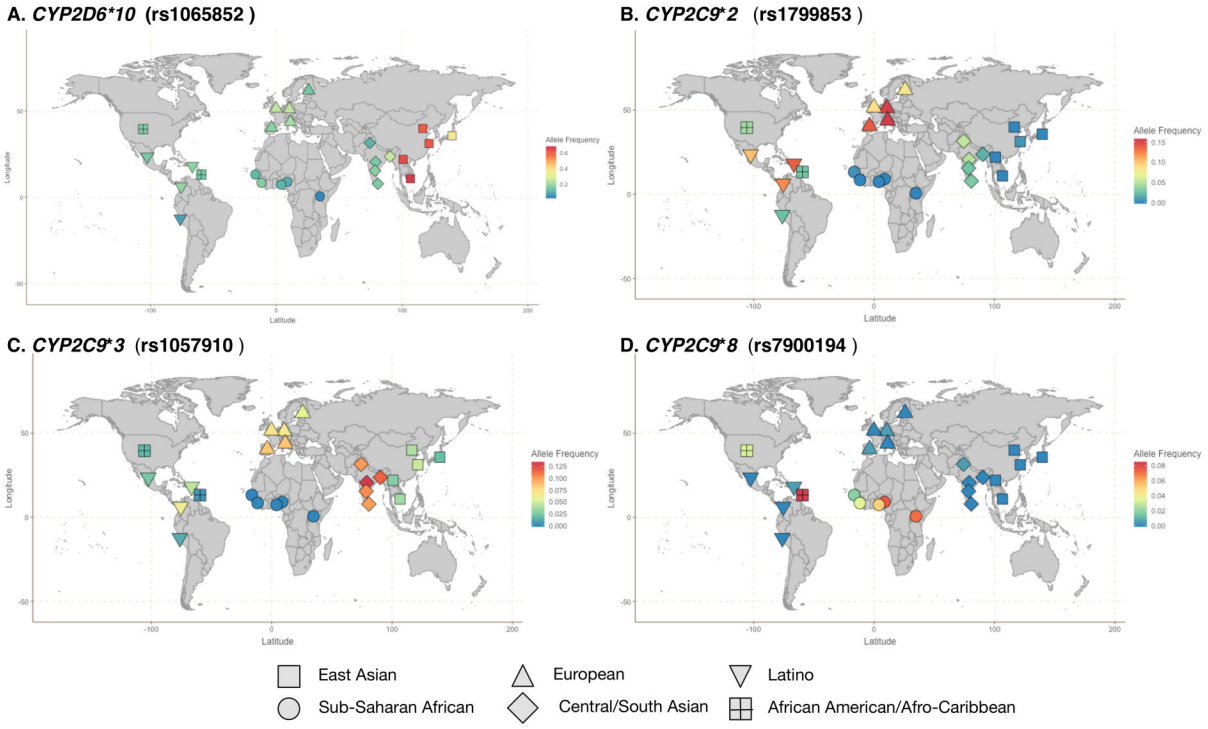


Figure 3: Maps illustrating how the proposed biogeographical grouping system can be used to illustrate the variability in global frequencies of key pharmacogenetic alleles.

Allele frequencies from 1000 Genomes are shown across global populations for (A) CYP2D6*10, (B) CYP2C9*2, (C) CYP2C9*3 and (D) CYP2C9*8.

Morales et al. Genome Biology (2018) 19:21
https://doi.org/10.1186/s13059-018-1396-2

Genome Biology

OPEN LETTER Open Access

A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog

Joannella Morales[1*], Danielle Welter[1], Emily H. Bowler[1], Maria Cerezo[1], Laura W. Harris[1], Aoife C. McMahon[1], Peggy Hall[2], Heather A. Junkins[2], Annalisa Milano[1], Emma Hastings[1], Cinzia Malangone[1], Annalisa Buniello[1], Tony Burdett[1], Paul Flicek[1], Helen Parkinson[1], Fiona Cunningham[1], Lucia A. Hindorff[2†] and Jacqueline A. L. MacArthur[1*†]

Abstract

The accurate description of ancestry is essential to interpret, access, and integrate human genomics data, and to ensure that these benefit individuals from all ancestral backgrounds. However, there are no established guidelines for the representation of ancestry information. Here we describe a framework for the accurate and standardized description of sample ancestry, and validate it by application to the NHGRI-EBI GWAS Catalog. We confirm known biases and gaps in diversity, and find that African and Hispanic or Latin American ancestry populations contribute a disproportionately high number of associations. It is our hope that widespread adoption of this framework will lead to improved analysis, interpretation, and integration of human genomics data.

Keywords: Genomics, Genome-wide association studies, GWAS Catalog, Ancestry, Diversity, Population genetics

Background

The past 15 years have seen a dramatic growth in the field of genomics, with numerous efforts focused on understanding the etiology of common human disease and translating this to advances in the clinic. Essential to the interpretation of this vast amount of data is the accurate and unambiguous description of the ancestry of samples. Degrees of genetic diversity and patterns of linkage disequilibrium (LD) vary by ancestry, with implications for the generalizability of results and the identification of disease-causing variants. The standardized representation of ancestry data is also indispensable to facilitate data access in bioinformatics resources and to support the integration of information from different sources, ultimately enabling more robust analyses of "big data" sets. The need for genetic studies in more ancestrally diverse populations has been repeatedly articulated [1], most recently by Popejoy and Fullerton [2]. Although inclusion efforts are improving over time, it is challenging to assess the status of such efforts without a standardized way of representing ancestry data.

There are currently no established guidelines for the description of ancestral information. We here provide a framework to represent, in an accurate and standardized manner, the ancestry of samples included in human genomics studies. We utilize our method to describe samples analyzed in over 3200 publications included in the NHGRI-EBI GWAS Catalog [3–5], validating its applicability to large and complex data sets. We also present a new and expanded analysis of Catalog ancestry content using, for the first time, our standardized framework. We thus demonstrate the efficacy of categories to facilitate data analysis, including tracking trends in the area of diversity. Finally, to ensure broader applicability beyond the Catalog to other studies or resources involving human subjects, we offer recommendations to authors

* Correspondence: jmorales@ebi.ac.uk; jalm@ebi.ac.uk
†Equal contributors
[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
Full list of author information is available at the end of the article

and provide an ancestry-specific ontology for application to bioinformatics resources. We also apply our method to the 1000 Genomes [6] and HapMap [7] project populations to enable integration with any samples described utilizing these well-established reference populations and of any variation data generated from these projects.

Results

Ancestry framework

Our framework involves representing the ancestry of samples in two forms: (1) a detailed description and (2) an ancestry category from a controlled list (Table 1). Detailed descriptions aim to capture accurate, informative, and comprehensive information regarding the ancestry or genealogy of each distinct sample. Category assignment reduces complexity within data sets and enables the establishment of hierarchical relationships, placing samples in context with other samples, groups, and populations. This is extremely useful, empowering more precise search functionalities and improved access to data in bioinformatics resources. This process also facilitates integration of results from multiple sources, ultimately enabling the community to better interpret findings and perform further analyses.

Validation in the NHGRI-EBI GWAS Catalog

To validate the framework, we applied our method to all publications included in the GWAS Catalog—3200 publications, representing 4600 separate GWA studies, 60000 associations, and 110 million individuals, as of November 2017. The Catalog is widely used, and invaluable for researching existing findings on common diseases and supporting investigations to identify causal variants, understand disease mechanisms, and establish targets for treatment [8–11]. As one of the largest repositories and visual summaries of genomic association data, the Catalog provided an ideal substrate on which to test our method and its applicability to large and complex data sets.

Each Catalog study entry comprises one or more samples, designated as "Initial" or "Replication" samples, depending on the stage of the GWA study in which they were analyzed (Fig. 1; Additional file 1: Figure S1 and Additional file 1: Figure S2a). For each sample, we created the detailed description by extracting the ancestry descriptor utilized by the author in the relevant publication. To generate the controlled description, we selected, from a limited list of terms (Table 1), the category noted by the author or, if not stated, the category that best correlates with the detailed description for the same sample. For example, we selected the category "East Asian" for detailed descriptions containing the descriptor "Han Chinese".

We relied heavily on data stated by authors in the GWAS publication, giving precedence to information inferred using genomic methods, such as principal component analysis (PCA; see Additional file 1: Box S1 for a list of methods commonly used to ascertain ancestry). In some cases, we considered other sources, but only when the information provided by authors was limited or ambiguous. We consulted peer-reviewed population genetics publications to obtain additional information on lesser-known groups that were not adequately characterized by authors or when samples were described using ethno-cultural terms (for example, "Punjabi Sikh"). When the only information provided in the publication was the location of recruitment, we consulted The United Nations M49 Standard of Geographic Regions [12] and The World Factbook [13]. The latter is a regularly updated compendium of worldwide demographic data, covering all countries and territories of the world. Additional file 2: Table S1 provides a list of countries of recruitment in the Catalog, together with the sources that were consulted and the inferred categories.

In rare instances, the ancestry information provided by authors was not detailed enough to allow the resolution of samples into ancestrally distinct sets. For these samples, we created complex, multi-ancestry detailed descriptions and selected multiple categories (for example, Catalog entry for Jiang R et al. [14, 15]). For admixed samples, we selected either one of the categories that includes individuals with well-defined admixture ("African American or Afro-Caribbean" and "Hispanic or Latin American") or the category "other admixed ancestry". We also captured additional information to describe the ancestral backgrounds that contribute to the admixture. No ancestry-informative detailed descriptions were generated in the absence of ancestry or recruitment data; for those samples, the category "Not Reported" was selected.

Where possible we also curated country of recruitment (Fig. 1; Additional file 1: Figure S2b) and country of origin as this provides additional and complementary demographic information. Country of origin was extracted when the country of origin of the study participant's grandparents was stated or when the genealogy of the sample could be traced to a particular country.

The detailed extraction guidelines utilized by Catalog curators are included in Additional file 1: Supplementary Methods. A full list of Catalog detailed descriptions and categories is provided in Additional file 3: Table S2. Examples that illustrate application to specific samples can be found in Additional file 4: Table S3. All curated ancestry data are available from the GWAS Catalog website [4] (Fig. 1) and via download [16].

Table 1 Ancestry categories: distinct regional population groupings used in this framework

Ancestry category	Definition	Examples of detailed descriptions for samples included in the category
Aboriginal Australian	Includes individuals who either self-report or have been described by authors as Australian Aboriginal. These are expected to be descendants of early human migration into Australia from Eastern Asia and can be distinguished from other Asian populations by mtDNA and Y chromosome variation [29, 30]	Martu Australian Aboriginal
African American or Afro-Caribbean	Includes individuals who either self-report or have been described by authors as African American or Afro-Caribbean. This category also includes individuals who genetically cluster with reference populations from this region, for example, 1000 Genomes and/or HapMap ACB or ASW populations. We note that there is likely to be significant admixture with European ancestry populations	African American, African Caribbean
African unspecified	Includes individuals that either self-report or have been described as African, but there was not sufficient information to allow classification as African American, Afro-Caribbean or Sub-Saharan African	African, non-Hispanic black
Asian unspecified	Includes individuals that either self-report or have been described as Asian but there was not sufficient information to allow classification as East Asian, Central Asian, South Asian, or South-East Asian	Asian, Asian American
Central Asian	Includes individuals who either self-report or have been described by authors as Central Asian [31]. We note that there does not appear to be a suitable reference population for this population and efforts are required to fill this gap	Silk Road (founder/genetic isolate)
East Asian	Includes individuals who either self-report or have been described by authors as East Asian or one of the sub-populations from this region (e.g., Chinese). This category also includes individuals who genetically cluster with reference populations from this region, for example, 1000 Genomes and/or HapMap CDX, CHB, CHS, and JPT populations	Chinese, Japanese, Korean
European	Includes individuals who either self-report or have been described by authors as European, Caucasian, white, or one of the sub-populations from this region (e.g., Dutch). This category also includes individuals who genetically cluster with reference populations from this region, for example, 1000 Genomes and/or HapMap CEU, FIN, GBR, IBS, and TSI populations	Spanish, Swedish
Greater Middle Eastern (Middle Eastern, North African, or Persian)	Includes individuals who self-report or were described by authors as Middle Eastern, North African, Persian, or one of the sub-populations from this region (e.g., Saudi Arabian) [32]. We note there is heterogeneity in this category with different degrees of admixture as well as levels of genetic isolation. We note that there does not appear to be a suitable reference population for this category and efforts are required to fill this gap	Tunisian, Arab, Iranian
Hispanic or Latin American	Includes individuals who either self-report or are described by authors as Hispanic, Latino, Latin American, or one of the sub-populations from this region. This category includes individuals with known admixture of primarily European, African, and Native American ancestries, though some may have also a degree of Asian (e.g., Peru). We also note that the levels of admixture vary depending on the country, with Caribbean countries carrying higher levels of African admixture when compared to South American countries, for example. This category also includes individuals who genetically cluster with reference populations from this region, for example, 1000 Genomes and/or HapMap CLM, MXL, PEL, and PUR populations [17, 33]	Brazilian, Mexican
Native American	Includes indigenous individuals of North, Central, and South America, descended from the original human migration into the Americas from Siberia [34]. We note that there does not appear to be a suitable reference population for this category and efforts are required to fill this gap	Pima Indian, Plains American Indian
Not reported	Includes individuals for which no ancestry or country of recruitment information is available	
Oceanian	Includes individuals that either self-report or have been described by authors as Oceanian or one of the sub-populations from this region (e.g., Native Hawaiian) [35]. We note that there does not appear to be a suitable	Solomon Islander, Micronesian

Table 1 Ancestry categories: distinct regional population groupings used in this framework *(Continued)*

Ancestry category	Definition	Examples of detailed descriptions for samples included in the category
	reference population for this category and efforts are required to fill this gap	
Other	Includes individuals where an ancestry descriptor is known but insufficient information is available to allow assignment to one of the other categories	Surinamese, Russian
Other admixed ancestry	Includes individuals who either self-report or have been described by authors as admixed and do not fit the definition of the other admixed categories already defined ("African American or Afro-Caribbean" or "Hispanic or Latin American")	
South Asian	Includes individuals who either self-report or have been described by authors as South Asian or one of the sub-populations from this region (e.g., Asian Indian). This category also includes individuals who genetically cluster with reference populations from this region, for example, 1000 Genomes and/or HapMap BEB, GIH, ITU, PJL, and STU populations	Bangladeshi, Sri Lankan Sinhalese
South East Asian	Includes individuals who either self-report or have been described by authors as South East Asian or one of the sub-populations from this region (e.g., Vietnamese). This category also includes individuals who genetically cluster with reference populations from this region, for example, 1000 Genomes KHV population. We note that East Asian and South East Asian populations are often conflated. However, recent studies indicate a unique genetic background for South East Asian populations	Thai, Malay
Sub-Saharan African	Includes individuals who either self-report or have been described by authors as Sub-Saharan African or one of the sub-populations from this region (e.g., Yoruban). This category also includes individuals who genetically cluster with reference populations from this region, for example, 1000 Genomes and/or HapMap ESN, LWK, GWD, MSL, MKK, and YRI populations	Yoruban, Gambian

Ancestry categories are assigned to samples with distinct and well-defined patterns of genetic variation, in addition to individuals with inferred relatedness to these samples. A full list of GWAS Catalog sample descriptions assigned to each category can be found in Additional file 3: Table S2

Improving data analysis and assessing diversity

Taking advantage of this fully curated and well described data set, we performed a new and enhanced survey of the ancestral background of Catalog samples. Similar analyses have been previously performed [1, 2]. However, these have focused exclusively on the detailed descriptions, which are more complex and heterogeneous. Our analysis uses, for the first time, categories and goes beyond individuals to studies, associations, traits, and change over time.

As previously reported [2], we found that the majority (78 %) of individuals in the Catalog are exclusively of European ancestry (Fig. 2a), followed by individuals of East Asian descent (9 %). The disproportionate focus on Europeans was more prevalent in the earlier years of the Catalog (86 % of individuals in studies published between 2005 and 2010; 76 % between 2011 and 2016), with a notable increase in African (0.8 to 2.8 %, 3.5-fold increase), Hispanic or Latin American (0.1 to 1.2 %, ninefold increase) and Middle Eastern (0.01 to 0.08 %, sevenfold increase) samples in the last 5 years (Fig. 2b). Despite this trend, however, these non-European, non-Asian groups combined account for less than 4 % of the Catalog's individuals. We observed a similar result when analyzing GWA studies. Almost 50 % of all studies exclusively analyze European ancestry individuals, and an additional 25 % of studies analyze multiple ancestries, including individuals of European descent (Fig. 2c).

Interestingly, when we focused on the number of associations contributed by each category, we noted a disparity with respect to the distribution observed when analyzing individuals (Fig. 2d). This was particularly pronounced for studies including African or Hispanic or Latin American samples, many of which are African-admixed [17]. African ancestries comprise 2.4 % of individuals but contribute 7 % of associations. Similarly, only 1.3 % of individuals in the Catalog are Hispanic or Latin American, yet they contribute 4.3 % of associations. The opposite effect was seen in Europeans, with 78 % of individuals yet only 54 % of associations.

Our ability to observe this disproportionate yield of associations is directly correlated with the use of categories in our analysis. The benefits of our framework, however, extend beyond assessing diversity to the pursuit of scientific questions. Using our categories, we were able to identify diseases or traits that have been analyzed in a large number of ancestral backgrounds and use this information to search for loci and variants that generalize across ancestries as well as loci or variants that may have ancestry-specific impact. For example, we found that type 2 diabetes has been analyzed in multiple ancestral backgrounds (29 distinct detailed descriptions and 12

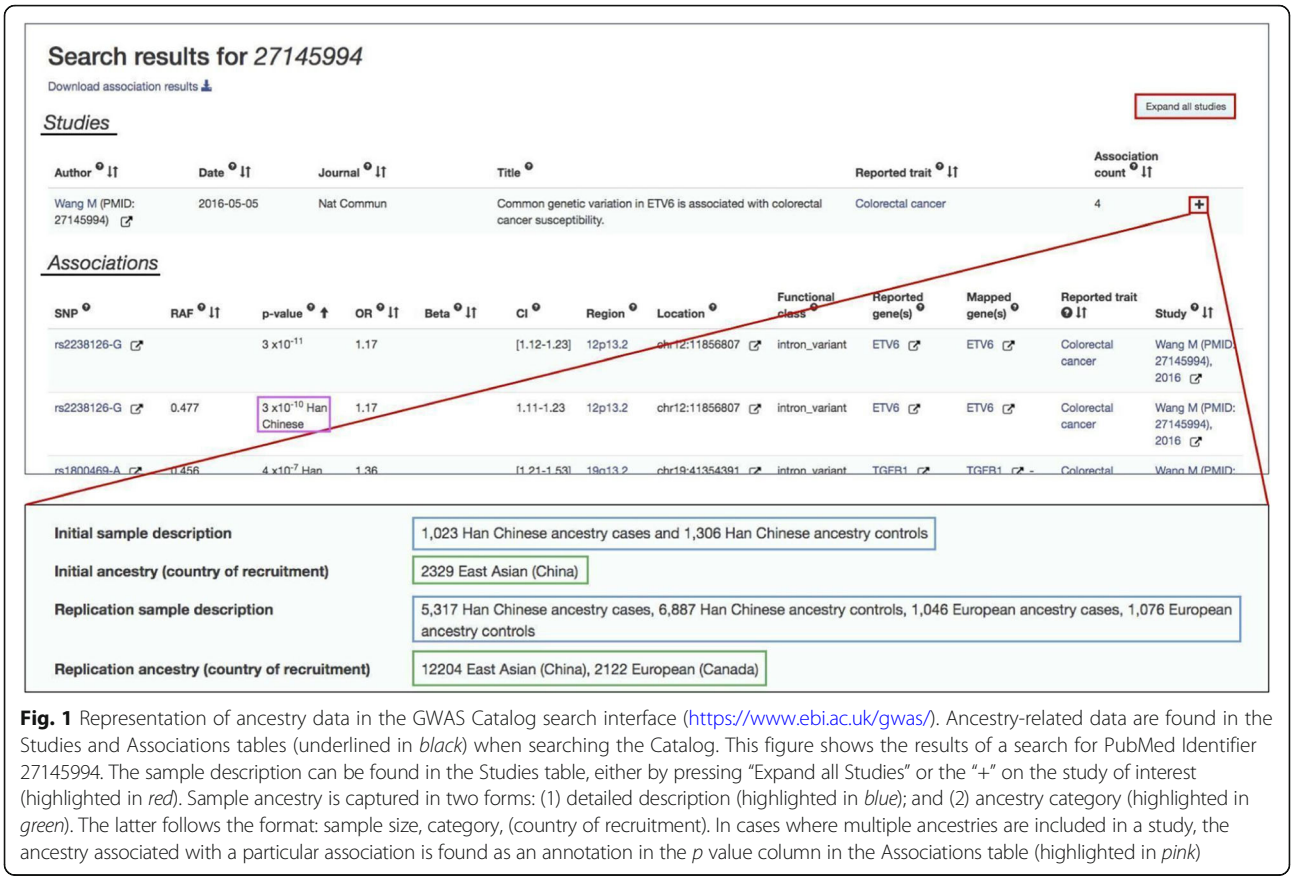


Fig. 1 Representation of ancestry data in the GWAS Catalog search interface (https://www.ebi.ac.uk/gwas/). Ancestry-related data are found in the Studies and Associations tables (underlined in *black*) when searching the Catalog. This figure shows the results of a search for PubMed Identifier 27145994. The sample description can be found in the Studies table, either by pressing "Expand all Studies" or the "+" on the study of interest (highlighted in *red*). Sample ancestry is captured in two forms: (1) detailed description (highlighted in *blue*); and (2) ancestry category (highlighted in *green*). The latter follows the format: sample size, category, (country of recruitment). In cases where multiple ancestries are included in a study, the ancestry associated with a particular association is found as an annotation in the *p* value column in the Associations table (highlighted in *pink*)

categories across 52 studies and 610 associations). We then reviewed all loci associated with this disease and found that some (for example, 10q25.2) appear to generalize across many ancestral groups and others seem limited to a small number (for example, 4p16.3 primarily in Asians). The assignment of our categories to the 1000 Genomes and HapMap project populations enables a more focused review of ancestry-specific LD and allele frequency information for these loci, and this, in turn, can inform study designs aimed at fine mapping and the identification of causal variants. This process also allows the identification of clear gaps in the data, such as particular ancestral backgrounds that have yet to be analyzed.

Application beyond the GWAS Catalog

To encourage widespread adoption of the framework, we here pursue three approaches.

Recommendations for authors

Curation of GWAS publications revealed inconsistent and ambiguous reporting of ancestry data, with a significant percentage of studies (~ 4 %) not reporting any relevant information at all. Therefore, we provide a set of specific recommendations for authors, summarized in (Table 2), that require minimal additional burden, and, if implemented, will improve the quality of reporting and have a positive impact on the interpretation of published results, data re-use, and reproducibility.

We recommend that authors make every effort to generate a detailed description for each distinct set of individuals included in their studies. Authors should also note a corresponding category by assessing whether the genetic diversity of each distinct set is representative of one of the known populations listed and defined in Table 1. Where possible, we recommend authors assess the ancestry using genomic methods (Additional file 1: Box S1), as this will aid the classification process. If authors have no knowledge about the ancestry of the participants, are not able to infer it, or cannot share it due to confidentiality concerns, we suggest noting this explicitly in the publication.

In general, terms that pertain to an individual's ethnocultural background should be avoided, unless this provides additional information regarding the genealogy of the samples. In such cases a descriptor that accurately reflects the underlying genetics should also be provided. For example, when describing "Punjabi Sikh" participants,

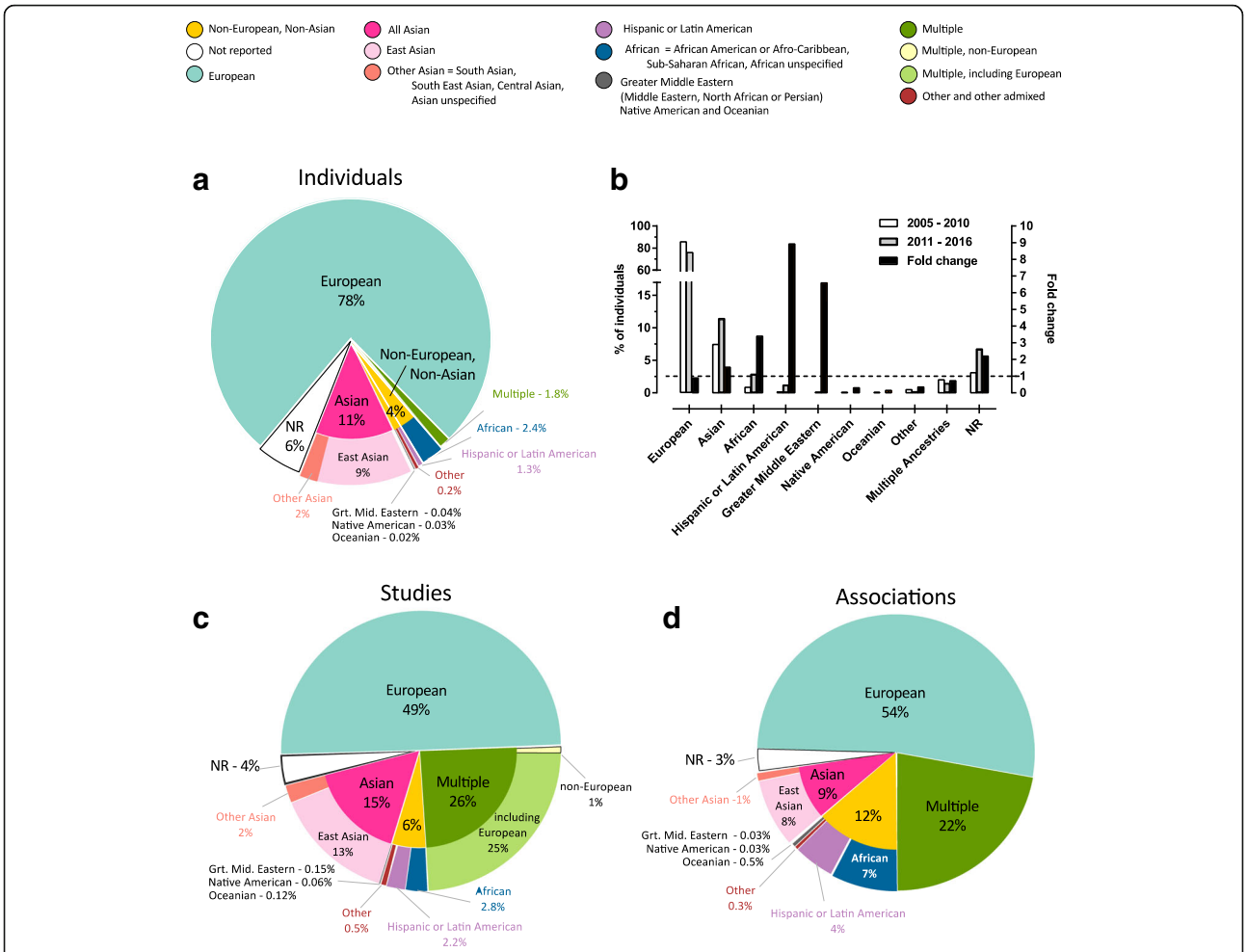


Fig. 2 Ancestry category distribution in the GWAS Catalog. This figure summarizes the distribution of ancestry categories in percentages, of individuals (N = 110,291,046; **a**), individuals over time (N = 110,291,046; **b**), studies (N = 4,655; **c**), and associations (N = 60,970; **d**). The largest category in all panels is European (*aqua*). At the level of individuals (**a**), the largest non-European category is Asian (*bright pink*), with East Asian (*light pink*) accounting for the majority. Non-European, Non-Asian categories together (*yellow*) comprise 4 % of individuals, and for 6 % (*white*) of samples no ancestry category could be specified. **b** The distribution of individuals in percentages, included in the 915 studies published between 2005 and 2010 compared to the distribution of individuals included in the 2905 studies published between 2011 and 2016. **d** The disproportionate contribution of associations from African (*blue*) and Hispanic/Latin American (*purple*) categories, when compared to the percentage of individuals (**a**, *blue*, *purple*, respectively) and studies (**b**, *blue*, *purple*, respectively)

authors should also describe the samples as "South Asian" or "Punjabi Sikh South Asian" rather than simply "Punjabi Sikh" or "Sikh".

Particular care should be taken to note if a sample derives from a founder or genetically isolated population. Given their homogeneity and reduced genetic variation, these populations are especially well-suited for GWAS [18] and are increasingly used as sample sources. When describing isolates, the broader genetic background within which the population clusters should also be indicated. For example, Old Order Amish participants should be described as "Old Order Amish population isolate individuals of European descent", for example.

While describing admixed populations can be challenging due to varying levels of admixture, every effort should be made to explicitly note whether the sample is admixed and the ancestral backgrounds that contribute to admixture. For example "Hispanics/Latinos are ethnically heterogeneous, with admixture of European, West African, and Amerindian ancestral populations", as stated in Hodonsky et al. [19].

Ancestry-specific ontology

To facilitate application to bioinformatics resources, we developed and released an ancestry-specific ontology based on our curated GWAS Catalog descriptions. We have defined terms, identified synonyms, and established hierarchical relationships between all curated terms and categories. The use of this ontology in any search interface will enable users to

Table 2 Recommendations for authors reporting ancestry data in publications. These recommendations were generated by expert curators following a detailed review of the over 3200 GWAS publications included in the Catalog

1. Provide detailed information for each distinct group of samples,
   a. Ancestry descriptors should be as granular as possible (e.g., Yoruban instead of Sub-Saharan African, Japanese instead of Asian).
   b. Avoid using country or citizenship as a substitute for ancestry.
   c. Avoid using geographic descriptors that are part of a cohort name as a substitute for ancestry (e.g., TwinsUK cannot be assumed to be European ancestry).
   d. If a population self-identifies using sociocultural descriptors, clearly provide information about the underlying genetics or genealogy (e.g., Old Order Amish individuals of European descent)
   e. If samples were derived from an isolated or founder population with limited genetic heterogeneity, clearly state the genetic ancestry within which this sub-population falls.
   f. Every effort should be made to explicitly note whether the population is admixed and the ancestral backgrounds that contribute to admixture.
   g. If available, genetic genealogy or ancestry of grandparents or parents should be included.
2. Report the method used to determine the ancestry of participants (for example, self-reported, inferred by genomic methods, or a combination of both)
   a. Where possible, use genomic methods to confirm self-reported ancestry or to infer the ancestry of samples.
   b. If inferred, indicate the analytical procedure utilized. See Additional file 1: Box S1 for a description of commonly used methods.
3. Assign an ancestry category for each distinct group of samples. See Table 1 for a list of ancestry categories. Refer to Additional file 3: Table S2 for a list of descriptors in use in the Catalog with their category assignments.
4. Provide the sample size for each distinct group of samples included in the analysis.
5. Provide country of recruitment.
6. If ancestry information is not available due to confidentiality, or any other concerns, note this in the publication.

perform more powerful and precise ancestry-related queries [20]. We aim to integrate it into the GWAS Catalog website in the near future. The ancestry ontology [21] can be browsed and downloaded (manuscript in preparation).

Application to reference populations

The HapMap [7] and 1000 Genomes [6] projects have collated a number of widely used reference populations and delivered a comprehensive survey of human genetic variation. The application of our framework to these populations, therefore, provides huge integration potential, especially with any samples described using these references in PCA and other analyses. For all HapMap and 1000 Genomes phase 3 populations, we assigned ancestry category, country of recruitment, country of origin, and a detailed description, if provided by each project (Additional file 5: Table S4).

Discussion

Summary

In this report, we describe a framework for the standardized representation of ancestry data from genomics studies. Our method provides structure to unstructured data, enabling robust searching across large datasets and integration across resources.

Limitations of the framework

Despite the successful application of our method to GWAS Catalog samples and to commonly used reference populations, there are challenges. We are aware of the sensitivities surrounding the topics of ancestry, race and ethnicity, and the difficulties that arise when trying to classify the global human population. Due to evolution and patterns of migration, the ancestry of a particular population is complex. However, it is both possible and useful to generate standardized terminology and to classify individuals into informative groupings. Reference populations or ancestry informative markers [22] that allow populations to be distinguished have been characterized, and methods have been developed to adjust for population stratification and separate samples into clusters. Practically, the classification of samples into categories facilitates data integration and allows robust searches, which is an essential component of databases such as the GWAS Catalog. Also, as we demonstrate in our survey of Catalog ancestry data, the use of categories can greatly facilitate further analyses by, for example, reducing the complexity of data sets.

We recognize that as more cohorts from diverse populations are characterized, there might arise a need to create additional categories or sub-categories. Also, it is likely that admixture will increase in the future, due to migration, for example, resulting in samples that could be described using multiple categories. The classification of admixed samples is particularly challenging. The degree and type of admixture may vary within the population, and the accuracy of classification requires well-defined reference samples, which are lacking for some groups. In an effort to address this, we have created categories to represent admixed groups that are known (for example, “Hispanic or Latin American”) and emerging (for example, “Other admixed ancestries”). We have also included, and recommended inclusion of, information regarding the populations that contribute to admixture. We note that since the vast majority of admixed Catalog samples can be classified as either “Hispanic or Latin American” or “African American or Afro-Caribbean”, we felt it was sufficient to create one category to include all other forms of admixture.

However, we recognize that as the community moves towards increased characterization of these groups, using genomic methods, for instance, our admixed categories are likely to become more precise and granular over time.

Assessing diversity in genomics

Several reports have been published urging the scientific community to ensure that individuals from all backgrounds benefit from advances in the field of genomics [1, 2]. However, this requires the establishment of metrics and proper tracking of ancestry data over time. As evidenced by our new survey of ancestral backgrounds, we believe the widespread implementation of our framework, especially the use of standardized language and categories, can yield important benefits in this area.

There are, however, limitations to the use of categories to track diversity in the Catalog. Considering that some cohorts have been included in numerous studies, some individuals are represented multiple times. The impact of this is the skewing of results towards commonly used or publicly available cohorts, which are likely of European or Asian descent. Also, associations identified in multi-ancestry studies, for example, "trans-ethnic" discoveries or multi-ethnic replications, could not be described using one category, resulting in a disproportionate number of "multiple" ancestry associations (1.8 % individuals, 22 % associations; Fig. 2d). This may contribute to the reduced proportion of associations attributed to European populations, since the vast majority of "multiple" ancestry studies include Europeans (Fig. 2c).

While the general bias towards inclusion of European ancestry samples in GWA studies has been previously reported, the disparity in the yield of associations derived from African and Hispanic or Latin American populations is a novel observation. We suggest that the higher degree of genetic diversity and reduced linkage disequilibrium (LD) in African [23] and African-admixed populations offers an explanation for this result. Shorter LD blocks in African populations facilitate the separation of nearby but independent signals in a way that is more challenging in populations with shorter LD blocks, such as Europeans and Asians. Also, as the number of individuals from African and Hispanic or Latin American populations has grown over the years, the power to discover additional disease-associated variants by leveraging the increased genetic diversity in these populations has improved.

The benefit of including diverse populations has been articulated, and extends throughout the translational research spectrum, from GWAS discovery efforts to genomic medicine. For example, studies including multiple populations may aid in fine mapping of existing signals or in identifying population-specific functional variation [6, 24]. Also, variant interpretation for genomic medicine in ancestrally diverse or admixed populations relies on the availability of non-European variation information, with potentially serious clinical consequences if such data are not available [25]. While we are encouraged by the trend we have seen in recent years towards increased diversity, we note that there are still very clear gaps as some groups continue to be underserved or ignored. We strongly urge the scientific community to expand their efforts to assemble and analyze cohorts, including especially underrepresented communities.

Human genomics studies, including GWAS, have been enormously successful [3, 5, 26]. However, the ability to properly interpret and query the generalizability of results across populations requires clarity about the ancestry of samples. Therefore, we have provided a framework for the standardized representation of ancestry. We believe widespread adoption will enable the scientific community to investigate the generalizability of genotype–trait associations across diverse populations, to identify associations more prevalent in specific ancestries, to identify novel variants with clinical implications, and to help pinpoint causative variants, thus increasing our understanding of common diseases.

Methods

GWAS Catalog data curation

GWAS Catalog eligibility criteria and general curation methods can be found on the GWAS Catalog website. Curation of ancestry data from the literature was performed according to the Ancestry Extraction Guidelines outlined in Additional file 1: Supplementary Methods.

GWAS Catalog ancestry analysis

To determine the distribution of individuals, associations and traits by ancestry category, we first downloaded all Catalog data in tabular form [16]. All data included in these analyses were curated from GWA studies published between 2005 and the end of 2016, with a release date of July 18 2017. The data can be found on the Catalog's FTP site [27] (gwas-catalog-associations_ontology-annotated.tsv, gwas-catalog-ancestry.tsv, gwas-catalog-studies_ontology-associated.tsv, and gwas-efo-trait-mappings.tsv).

1000 Genomes and HapMap Project population ancestry assignment

Information describing the 1000 Genomes [6] phase 3 and HapMap Project [7] phase 3 populations was taken from the Coriell Institute website [28]. Ancestry information, including ancestry category, country of recruitment, country of origin, and additional information, was assigned to each population following the GWAS Catalog Ancestry Extraction Guidelines mentioned above.

Additional files

Additional file 1: Figure S1. Detailed sample description displayed in the internal GWAS Catalog curation interface. **Figure S2. a** Structured ancestry and recruitment information displayed in the internal GWAS Catalog curation interface. **b** GWAS Catalog ancestry and recruitment data entry page of internal curation interface. **Supplementary Box 1.** Genomic methods of ancestry determination. **Figure S3.** Distribution of studies by ancestry category focused on Catalog traits with highest number of studies in the Catalog. **Figure S4.** Methods of ancestry ascertainment used in a subset of publications included in the GWAS Catalog. **Supplementary References**. (DOCX 893 kb)

Additional file 2: Table S1. GWAS Catalog countries of recruitment for which no ancestry information was provided. (XLSX 77 kb)

Additional file 3: Table S2. GWAS Catalog detailed descriptions with ancestry category assignments. (XLSX 77 kb)

Additional file 4: Table S3. Specific examples to illustrate the application of the framework to the GWAS Catalog. (XLSX 70 kb)

Additional file 5: Table S4. HapMap Project and 1000 Genomes Project populations with assigned ancestry category. (XLSX 27 kb)

Acknowledgements
The authors wish to thank all GWAS Catalog users and authors of studies included in the Catalog. We also thank Chris Gignoux for his expert review of the genomic methods of ancestry determination discussed in this manuscript, Kira Harvey for early discussions and assistance with assessing genomic methods, and Teri Manolio for valuable discussion.

Funding
Research reported in this publication was supported by the National Human Genome Research Institute and the National Institute of General Medical Sciences of the National Institutes of Health under award numbers U41-HG007823 and U41-HG006104. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported by the European Molecular Biology Laboratory. L.A.H., P.H., and H.J. are employees of the National Human Genome Research Institute.

Availability of data and materials
The datasets generated and/or analyzed during the current study are available on the NHGRI-EBI GWAS Catalog search interface [4] and in spreadsheet form [16].

Authors' contributions
JM, JALM, PH, HAJ, and LAH conceived this study and developed the ancestry framework. JM, JALM, EHB, AB, MC, PH, LWH, HAJ, ACM, AM, and LAH performed curation of ancestry data of GWAS Catalog publications. JM, JALM, MC, TB, and LAH analyzed the distribution of ancestry categories in the Catalog and interpreted the data. LWH, JALM, LAH, and JM assessed the methods of ancestry determination utilized in GWAS Catalog studies and interpreted the data. ACM and JM generated the figures. JM, JALM, and LWH generated the tables. EH, DW, CM, and TB developed the GWAS Catalog curation and search interfaces. DW created the ancestry ontology, with contributions from JM, JALM, and EHB. All authors contributed to writing and review of the final manuscript, with JM, JALM, and LAH playing the key roles. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
PF is a member of the Scientific Advisory Board of Omicia, Inc.

Publisher's Note

Author details
[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [2]Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892-9305, USA.

Received: 21 August 2017 Accepted: 19 January 2018
Published online: 15 February 2018

References
1. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009;25:489–94.
2. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538:161–4.
3. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45:D896–901.
4. GWAS Catalog. http://www.ebi.ac.uk/gwas/. Accessed 4 Aug 2017.
5. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–6.
6. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
7. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52–8.
8. Onengut-Gumuscu S, Chen W-M, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015;47:381–6.
9. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet. 2011;89:607–18.
10. Pal LR, Moult J. Genetic basis of common human disease: insight into the role of missense SNPs from genome-wide association studies. J Mol Biol. 2015;427:2271–89.
11. Mullen J, Cockell SJ, Woollard P, Wipat A. An integrated data driven approach to drug repositioning using gene-disease associations. PLoS One. 2016;11:e0155811.
12. UNSD—Methodology. https://unstats.un.org/unsd/methodology/m49/. Accessed 4 Aug 2017.
13. The World Factbook—Central Intelligence Agency. https://www.cia.gov/library/publications/resources/the-world-factbook/index.html. Accessed 4 Aug 2017.
14. GWAS Catalog. http://www.ebi.ac.uk/gwas/search?query=22391508. Accessed 14 Aug 2017.
15. Jiang R, French JE, Stober VP, Kang-Sickel J-CC, Zou F, Nylander-French LA. Single-nucleotide polymorphisms associated with skin naphthyl–keratin adduct levels in workers exposed to naphthalene. Environ Health Perspect. 2012;120:857–64.
16. GWAS Catalog. http://www.ebi.ac.uk/gwas/docs/file-downloads. Accessed 14 Aug 2017.
17. Adhikari K, Mendoza-Revilla J, Chacón-Duque JC, Fuentes-Guajardo M, Ruiz-Linares A. Admixture in Latin America. Curr Opin Genet Dev. 2016;41:106–14.
18. Cronin S, Berger S, Ding J, Schymick JC, Washecka N, Hernandez DG, et al. A genome-wide association study of sporadic ALS in a homogenous Irish population. Hum Mol Genet. 2008;17:768–74.
19. Hodonsky CJ, Jain D, Schick UM, Morrison JV, Brown L, McHugh CP, et al. Genome-wide association study of red blood cell traits in Hispanics/Latinos: the Hispanic Community Health Study/Study of Latinos. PLoS Genet. 2017;13:e1006760.
20. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.
21. Ancestry Ontology. http://www.ebi.ac.uk/ols/ontologies/ancestro. Accessed 2017 Aug 14.

22. Paschou P, Lewis J, Javed A, Drineas P. Ancestry informative markers for fine-scale individual assignment to worldwide populations. J Med Genet. 2010;47:835–47.
23. Campbell MC, Hirbo JB, Townsend JP, Tishkoff SA. The peopling of the African continent and the diaspora into the new world. Curr Opin Genet Dev. 2014;29:120–32.
24. Asimit JL, Hatzikotoulas K, McCarthy M, Morris AP, Zeggini E. Trans-ethnic study design approaches for fine-mapping. Eur J Hum Genet EJHG. 2016;24:1330–6.
25. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016;375:655–65.
26. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
27. Index of/pub/databases/gwas/releases/2017/07/18/. ftp://ftp.ebi.ac.uk/pub/databases/gwas/releases/2017/07/18/. Accessed 14 Aug 2017.
28. Coriell Biorepository. https://catalog.coriell.org/. Accessed 14 Aug 2017.
29. Huoponen K, Schurr TG, Chen Y, Wallace DC. Mitochondrial DNA variation in an aboriginal Australian population: evidence for genetic isolation and regional differentiation. Hum Immunol. 2001;62:954–69.
30. Nagle N, Ballantyne KN, van Oven M, Tyler-Smith C, Xue Y, Taylor D, et al. Antiquity and diversity of aboriginal Australian Y-chromosomes. Am J Phys Anthropol. 2016;159:367–81.
31. Martínez-Cruz B, Vitalis R, Ségurel L, Austerlitz F, Georges M, Théry S, et al. In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. Eur J Hum Genet. 2011;19:216–23.
32. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016;48:1071–6.
33. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, et al. Genomic insights into the ancestry and demographic history of South America. PLoS Genet. 2015;11:e1005602.
34. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012;488:370–4.
35. Kayser M. The human genetic history of Oceania: near and remote views of dispersal. Curr Biol. 2010;20:R194–201.

The British Journal of Sociology 2015 Volume 66 Issue 1

A post-genomic surprise. The molecular reinscription of race in science, law and medicine

Troy Duster

Abstract

The completion of the first draft of the Human Genome Map in 2000 was widely heralded as the promise and future of genetics-based medicines and therapies – so much so that pundits began referring to the new century as 'The Century of Genetics'. Moreover, definitive assertions about the overwhelming similarities of all humans' DNA (99.9 per cent) by the leaders of the Human Genome Project were trumpeted as the end of racial thinking about racial taxonomies of human genetic differences. But the first decade of the new century brought unwelcomed surprises. First, gene therapies turned out to be far more complicated than any had anticipated – and instead the pharmaceutical industry turned to a focus on drugs that might be 'related' to population differences based upon genetic markers. While the *language* of 'personalized medicine' dominated this frame, research on racially and ethnically designated populations differential responsiveness to drugs dominated *the empirical work* in the field. Ancestry testing and 'admixture research' would play an important role in a new kind of molecular reification of racial categories. Moreover, the capacity of the super-computer to map differences reverberated into personal identification that would affect both the criminal justice system and forensic science, and generate new levels of concern about personal privacy. Social scientists in general, and sociologists in particular, have been caught short by these developments – relying mainly on assertions that racial categories are socially constructed, regionally and historically contingent, and politically arbitrary. While these assertions are true, the imprimatur of scientific legitimacy has shifted the burden, since now 'admixture research' can claim that its results get at the 'reality' of human differentiation, not the admittedly flawed social constructions of racial categories. Yet what was missing from this framing of the problem: 'admixture research' is itself based upon socially constructed categories of race.

Keywords: Reductionism; racial admixture; genetics; reification; post-genomic

Duster (Warren Institute on Law and Social Policy, University of California, Berkeley and Department of Sociology, New York University) (Corresponding author email: troy_duster@berkeley.edu)
 ISSN 0007-1315 print/1468-4446 online.
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA on behalf of the LSE. DOI: 10.1111/1468-4446.12118

Introduction

Precisely at the dawn of the twenty-first century – as if on cue, the first draft of the Human Genome Map was completed, providing two kinds of hope for the new century. The first was explicitly about potential medical advances – the promise and prediction that the completed map would spur the development of new kinds of therapies that would increase health and reduce the ravages of a wide variety of diseases. The second hope was more a diffuse political aspiration, loudly trumpeted at the oft-cited White House news conference marking the event in June 2000. That was when US President Clinton, UK Prime Minister Tony Blair and the two molecular geneticists who had led the public and private sector human genome projects all agreed that: *At the level of the DNA, there is no such thing as race*.

Indeed, Paul Gilroy (2000) published a monograph that very year in which he argued that the findings from human genomics would put to rest the idea that there are important biological and genetic bases for differentiating human populations. However, relevant to this pronouncement about the 'end of race' as biological, Mark Twain once famously quipped about a newspaper article that reported that he had died, 'the news of my death has been greatly exaggerated'. So it has been with racial and ethnic categories in biology and human genetics. Indeed, there is substantial evidence that developments in several fields of inquiry and practice related to molecular genetics (pharmacogenomics, pharmacotoxicology, clinical genetics, personalized medicine and forensic science) have actually served to re-inscribe race as a biological category (Toom 2014; TallBear 2014, 2013; Kahn 2013; Roberts 2011; Bolnick 2008; Fullwiley 2008, 2007; Duster 2006b, 2005).

In early 2014, former *New York Times* science writer Nicholas Wade (2014) published a book that stirred predictable controversy. Wade claimed that much of human intelligence, and thus much of what he alleged to be corresponding human achievement of the last several centuries can be explained by differences between races. Europeans in general, and Jews in particular, he argued, achieved so much more than other racial and ethnically designated populations because of their bio-genetic makeup. It took less than four months for a group of 139 scientists, population geneticists and evolutionary biologists, to sign a document denouncing Wade's conclusions. The document, published by the *New York Times Sunday Book Review* (2014) stated that the work of these scientists did not lend support to Wade's thesis about racial differences, intelligence, and achievement. Wade does indeed make huge and unwarranted speculative leaps about the unfolding of colonial empires and the ascendancy of economic and political systems – based upon what he asserts to be the biological and genetic composition of different races. No research has been able to identify a functional genomics that would code for such a complex phenotype as human intelligence. None the less, one should not be lulled into

the false conclusion that the new human molecular genetics has been a battering ram undermining the idea of a biological basis of racial categories, or even a neutral bystander on matters of race. Indeed, as I will demonstrate, scientists from these fields have played an important (and sometimes) unwitting role in resuscitating the idea of race as biological, even genetic.

Background and context

No one could have predicted that the Department of Defense's commissioning of a secure high-speed communication network for military use would 'spin-off' into the Worldwide Web of the contemporary internet.[1] While the Defense Department achieved its goal, this achievement has been dwarfed by the scale of the social, economic, and political consequences of the way the internet has developed. From daily commercial transactions to downloads of streaming videos, from distance learning to search engines that can mine the Library of Congress, from e-mail to news media clips to blogs, from Facebook and Twitter to Instagram – and this list could go on for several pages. The Internet is such a dominant feature of the lives of so many that on a worldwide scale, if banks or hotels or restaurants do not have their own websites, they are consigned to outlier status in the backwaters of an integrated global network.

While no one could have predicted the speed and penetration of these developments, if we stopped and thought about it, there are known synergistic ingredients to the infrastructure that, placed onto the conveyor belt of modern life, we could have expected dramatic changes. In a parallel fashion, one of the deeply consequential spin-offs of the mapping and sequencing of the entire human genome has been the use of these technologies to identify individuals at the level of the micro-chip. This capacity for identification of millions, even billions across the globe, is bound to have even more 'spin-offs' that will ultimately dwarf the original intentionality of the early advocates of a genome map. The first of those is just now emerging – the subtle, sometimes inadvertent re-inscribing of race at the molecular level. A second development – the use of markers for individual identification, and for claims to 'authenticity' for group membership – has substantially penetrated the criminal justice system with forensic uses, and has either advanced or subverted claims to tribal membership. Another use is the collection of DNA from suspects or arrestees in pre-trial circumstances to increase the DNA database, which in turn is designed to help law enforcement determine whether there are matches between the DNA samples of those suspects or arrestees and tissue samples left at some unsolved crime: the net to catch the guilty. In societies in which there is notable racial diversity, there have been race-based DNA dragnets. These involve situations in which DNA evidence is left at the crime scene, and the suspect is thought to be of a particular race – and the police then ask for DNA samples of those in the surrounding area who fit the racial profile of the

suspect (Boeschenstein 2006; Hanson 2004). In a more subtle and far-reaching development, convicted felons report that they are more persuaded by the prosecutorial claim of DNA evidence against them than even the prosecutors – who often lie about the possession of such evidence (Prainsack and Machado 2012). These developments are a long way from the original goals of the mapping and sequence of the human genome.

The rationale behind the Human Genome Project, and for the Haplotype Map Project which followed, has always been the search for ways to improve our health. There have already been some health benefits, and there will certainly be more. None the less, in this paper I will point to mounting evidence that the inadvertent and unintended spin-offs (into domains far removed from health concerns and clinical medical applications) of the revolution in human molecular biology will dwarf the health achievements.

Context for the ongoing dilemma

Two contradictory magnetic poles pull medical research on humans in opposite directions – producing a tension that will never be resolved.[2] On the one hand, there is a universalizing impulse – based upon a legitimate assumption that human bodies are sufficiently similar that vaccines, catheters, pasteurizing processes and tranquilizers that work in one population will work in others. On the other hand, and unless and until research protocols establish and confirm specific similarities across populations, there is sufficient human variation that targeting medicines for specific populations *can be a legitimate, even vital empirically driven task*. The theoretical question, of course, is why a particular population or sub-population is to be so targeted? Because of folk theories about different groups' biological difference? – or because of their social and political standing? Age, gender and race leap to the forefront. The history of research on ailments as disparate as breast and prostate cancer (Rothenberg 1997; Wailoo 2011), heart disease (Cooper et al. 2005) and syphilis (Jones 1981; Reverby 2009) provides strong evidence that the answer is not either/or – but *both*. So, on what grounds do we choose one strategy over the other?

And it is precisely on this point that Steven Epstein (2007) raises the most fundamental question:

> Out of all the ways by which people differ from one another, why should it be assumed that sex and gender, race and ethnicity, and age are the attributes of identity that are most medically meaningful? Why these markers of identity and not others? (Epstein 2007: 10)

The answer is profoundly social and political, economic and cultural. The USA is the only country in the world that, as public health policy, does not operate on the assumption of the single standard human.

Moreover, by highlighting certain categories, there is the unassailable truth that other categories are thereby ignored. But more to the theoretical point – because each of the categories noted above has a potential or real biological base in either scientific or common sense understandings (Schütz 1962) when scientists report findings indicating differences, the danger is that these findings can seductively divert policy-makers from seeking alternative interventions that could better address health disparities (Krieger 2011).

In the history of science and medicine, it is a very recent requirement that researchers requesting federal funding and drug manufacturers trying to obtain regulatory approval for their company's products 'are now enjoined to include women, racial and ethnic minorities, children, and the elderly as research subjects in many forms of clinical research' (Epstein 2007: 5).

This shift has occurred only in the last two and a half decades, beginning with regulations that were developed first in 1986 in the USA – a direct consequence of sharp challenges to the presumption that findings derived from the study of any single group, such as middle-aged white men, might be generalized to other populations. Once again, it is important to re-state the relatively unique feature of this development as it applies mainly to the USA (Epstein 2007: 7). The rest of the world has continued to act upon the presupposition of the standard human, at least until now. As we shall see, that too is about to change.

When is difference just difference, and when is difference something that inexorably stratifies a population? The answer lies in immediate history, context and setting – in particular, whether there have been social meanings attributed to that differentiation. The authors of an often-cited piece in *Genome Biolo*gy seem to acknowledge this when they say:

> Finally, we believe that identifying genetic differences between racial and ethnic groups, be they for random genetic markers, genes that lead to disease susceptibility or variation in drug response, is scientifically appropriate. What is not scientific is a value system attached to any such findings. Great abuse has occurred in the past with such notions as 'genetic superiority' of one particular group over another. The notion of superiority is not scientific, only political, and can only be used for political purposes. (Risch et al. 2002: 11)

But while the sentiment is admirable, this formulation constitutes a fundamentally flawed notion of a firewall between 'science' and 'politics'. All societies make sharp differentiations among their members that permit stratifying some groups over others. When humans create categories such as 'caste' or 'ethnic group' or 'race' – those taxonomies are political, and they are stratified in the most basic meaning of hierarchy: power-based differential access to resources. These three categories routinely pre-date and pre-figure scientific inquiry, but profoundly and routinely configure that inquiry. Over time, the

interaction between living at the top or bottom of a stratified hierarchy produces systematized different access to the rawest human needs. This means that there will be a feed-back loop to various health and illness outcomes to those different 'populations' (i.e., so stratified). If that seems abstract, here is a poignant example of that feedback loop.

Syngenta is one of the world's leading agri-business companies, with over 25,000 employees in nearly 100 countries across the globe. According to its official website, the company is dedicated to increase crop productivity through scientific advances, and to 'protect the environment and improve health and quality of life'. Syngenta has a plant in St. Gabriel, Louisiana, where it manufactures a crop-enhancing product called atrazine. But atrazine has an unfortunate side effect – it 'demasculinizes and feminizes' vertebrate animals who are exposed to it by inducing aromatase. When humans are exposed to atrazine for sustained periods, they are at much increased risk for certain cancers. The production facility in St. Gabriel has a prostate cancer rate 8.4 times higher among factory workers exposed to atrazine, and it just so happens that this plant is located in a community that is more than 80 per cent African American (Hayes 2010:3768).

These sharply different rates of prostate cancer between whites and blacks can be studied scientifically by geneticists trying to understand 'population differences' through a uni-dimensional genetic prism, but with no understanding of the larger context in which humans are exposed to environmental insults – as in the first part of the formulation by Risch et al. (2002). But we can also study the systemic pattern of African Americans living close to toxic waste dumps across the whole country (Bullard 2000; Sze 2007). That is also available for systematic empirical investigation and testable formulations – otherwise known as science. Why should the decontextualized genetic inquiry of differing prostate cancer rates between Americans of European and recent African descent be characterized as apolitical 'science' – while the rate of their increased risk to exposure to atrazine is seen as 'political' science? The answer is lodged in current culturally framed notions of the hierarchy of science. Being completely ahistorical and apolitical, we could take a sample of two different populations of whites and blacks in the contemporary USA, and we would find differences in their rates of hypertension. While there is some debate about the extent of the gap, blacks do tend to have somewhat higher rates than whites. But as Richard Cooper and his colleagues (Cooper et al. 2005) have shown, cross-cultural data demonstrate that this is not evidence for a biological difference between the races. This study examined hypertension prevalence rates among 85,000 subjects. It was explicitly designed to compare racial differences, sampling whites from eight surveys completed in Europe, the US, and Canada – and contrasting these results with those of a sample of three surveys among blacks from Africa, the Caribbean, and the US. The data from Brazil, Trinidad and Cuba show a significantly smaller racial disparity in blood

pressure than found in North America, and then most tellingly the authors of the study conclude:

> These data demonstrate that the consistent emphasis given to the genetic elements of the racial contrasts may be a distraction from the more relevant issue of defining and intervening on the preventable causes of hypertension, which are likely to have a similar impact regardless of ethnic and racial background. (Cooper et al. 2005)

Yet the Cooper study, which involved more than 80,000 subjects across eight nations, was not taken as seriously as the study of 1,056 African American subjects in a solely US-based study of hypertension (Roberts 2011). Indeed, the FDA approved a drug designed by African Americans with hypertension the spring of the very same year, after the Cooper study was published (Kahn 2013; Cooper et al. 2005). Since that decision was demonstrably more about economics, patenting and politics than about science, it is a naïve to think that these factors can be neatly parsed and isolated from each other.

Two quite separate and seemingly unrelated developments have converged in the last decade to concretize, literally, the cascading reification of race and ethnicity in science, medicine and law. A useful metaphor comes from an understanding of the two elements that must come to together to explain how an epoxy glue is made: One part is a crystalline phenol, either synthesized or made from organic resins. The second part is a catalyst or hardener. If the two parts are kept apart, nothing bonds. But when the two are stirred together, the bonding can produce a superglue that is remarkably hard-and-fast. So it is with the two-part formula that has crystallized an early twenty-first century re-inscription of race.

Molecular admixture: part 1 of the infrastructure of the new reification of race and ethnicity

How did we come to this current situation, in which there is an increasing acceptance of the idea that 'the reality of race' lies in the capacity of the computer to determine proportional ancestry at the molecular level? In short, what is the fast and unfolding history of a technology that has swept across fields as diverse as clinical medicine, ancestry testing, pharmaceutical industry trials for new drug developments, and molecular photo-fitting in forensics? The inventor of the technology is Mark Shriver of Pennsylvania State University. In his first paper on the topic, published in 1997, Shriver entitled the work 'Ethnic-Affiliation Estimation by Use of Population-Specific DNA Markers'. However, 'population specific markers' are so rare that the focus shifted to the relative frequency of markers when comparing populations. Shriver would later revise this framing, and in subsequent papers, refer only to something he would call as Ancestry Informative Markers. In a conversation which I still

vividly remember from our earliest exchanges on this topic, Shriver told me that the phenotype of race was too arbitrary and unable to signal or signify the real genetic make-up of the individual. He understood quite correctly that in different societies, and even in different decades of the same society, the definition of races can be fluid and arbitrary. For example, he knew that the census categories changed several times over the course of a few decades, whether in the USA, Brazil, or other nations categorizing their citizens by race. But the new markers, he insisted, would get at the underlying and thus genuine genetic composition of the individual. So, he was saying that he could use the computer analysis to determine one's 'real' genetic make-up, because the social designation of race and ethnicity is dependent upon such factors as conjectural imputations and arbitrary categorizations by others. In 2001 the BBC aired an extraordinary film, *Motherland: A Genetic Journey*, in which Shriver's work was featured 'authenticating' proportional ancestry. By 2004, he was using AIMS to tell college students just how wrong they were in assuming that the categories that they had been living with reflected 'the reality' of their race or ethnicity (Daily 2005).

Current technology permits us to link via DNA analysis to only two specific lines. On the Y chromosome, one's father's father's DNA, going back as far as we can locate the genetic material, can be determined with a high degree of certainty. That is how Thomas Jefferson – or one of his brothers – was definitively linked to Sally Hemings' offspring (Gordon-Reed 2008). On the female side, mitochondrial DNA (mtDNA) can link one's mother's mother's mother going back as far as we can garner the DNA. So, while we have 64 great-great-great-great-grandparents, the technology allows us to locate only two of those 64, if we're going back six generations, as our real legacy and genetic link to the past. But what of the other 62? Those links are equal contributors to our genetic makeup, and we ignore them only because we do not have access to them.

Sometimes putative links to ancestry (or lack of same) have significant financial repercussions. The Black Seminoles have struggled with this very question – of whether to use DNA analysis to 'authenticate' their relationship to the Seminole Indian Tribe. The reason is straightforward and serious: money. The federal government, pursuant to a land-settlement claim, made an award to Seminole Indians in 1976, poised to distribute upward of $60 million. In 2000, the Seminole Nation of Oklahoma amended its constitution so that members needed to show 'one-eighth Seminole blood' (Johnston 2003). The Black Seminoles could use either Y-chromosome analysis or mitochondrial DNA (mtDNA) to link themselves through very thin chains back on two edges of the genealogical axis (mother's mother's mother, etc.; or father's father's father, etc.), but that would miss all other grandparents (14 of 16, 30 of 32, 62 of 64). The stakes are even higher for the Florida Seminoles. In 2006, the tribe purchased the entire Hard Rock Café chain for approximately one billion

dollars. If you were offered a genetic ancestry test of either Y-Chromosome or mtDNA analysis, would you really want to engage the probabilistic Russian-roulette type gamble? Kim TallBear puts it this way:

> . . . genetic population categories themselves are not even consistently defined. For example, a scientist may draw blood from enrolled members at the Turtle Mountain Band of Chippewa Indians reservation in North Dakota and call her sample a 'Turtle Mountain Chippewa' sample. At the same time, she may have obtained 'Sioux' samples from multiple other scientists and physicians who took them at multiple sites (on multiple reservations or in urban Indian Health Service facilities) over the course of many years. In the Turtle Mountain Chippewa instance, we have a 'population' circumscribed by a federally-recognized tribal boundary. In the 'Sioux' instance, we have a population circumscribed by a broader ethnic designation spanning multiple tribes. Histories of politics *inhere in* the samples. (TallBear 2014)

Unlike Y DNA or mtDNA tests, the of technology Ancestry-Informative Markers (AIMS) examines a group's relative share of genetic markers found on the autosomes – the non-gender chromosomes inherited from both parents. As noted, AIMS are overwhelmingly shared across all human populations, it is therefore not their absolute presence or absence, but their rate of incidence, or frequency, that is usually being analysed. This is especially true when it comes to claims about the distinctiveness of continental populations. How did some markers come to represent ancestral populations of Africa, Europe, and Native America? Because the companies marketing ancestry tests hold proprietary interests in their techniques, most do not make them available for possible scientific replication, and their modeling constructs are therefore typically undisclosed. Thus, we are usually left to speculate about the threshold level of frequency that is used to determine the grounds for inclusion or exclusion, as well as what counts as a 'pure' referent population.

In one lab that permitted its procedures to be studied by a medical anthropologist, ancestry percentages were generated by formulas that compare the relative frequency of markers (44 in total) between selected populations of recent European, African, and Native American descent (Fullwiley 2008). All those in the defined group were tested for the frequency of markers that the researchers hoped would provide relative distinguishability. Recall that the frequency at which each marker appears in each group is noted – and whole continents are never sampled. Finally, the researchers compare marker frequencies between the three groups to come up with values which, when taken together, yield a probability result about ancestral percentages. This procedure generates the baseline for the statistically-based notion of a 100 per cent pure European (or African, etc.), so that when you send in your DNA from the saliva swab, and it turns out that you have one-third of the markers that have

been designated as 'European' – you are told that you are 33 per cent European. It is by this statistical legerdemain that we have come to generate the first half of the two elements that go into the molecular reinscription of race in contemporary human genetics (TallBear 2014; Fullwiley 2007).

There are a number of deeply problematic, even flawed assumptions behind that percentage claim. What is this 'reference population' that has become the measuring stick by which we inform people of their 'per cent ancestry to a putatively pure continental population' (read 'race' here)? Let's re-examine such a result if reported back to someone of recent African descent. First, more than 700 million people currently inhabit the African Continent – and human geneticists have known for decades that this is the continent with the greatest amount of genetic variation on the globe. One important reason for this variation has been noted by Pilar Ossorio:

> For many regions of the human genome, there are more variants found among people of Africa than found among people in the rest of the world. This is probably because humans have resided in Africa for much longer than we have resided any place else in the world, so our species had time to accumulate genetic changes within the people in Africa. (Ossorio 2009).

A scientifically valid random sampling of even one per cent of this population would require a prohibitively expensive research programme – a database of seven million. So instead researchers have settled for 'opportunity samples' or 'convenience samples' – namely, a few hundred here or there, or even thousands that have been collected for a variety of reasons. No attempt has ever been made to take theoretically driven or random samples from African tribes such as the Lua, Kikiyu, Ibo, Hauser, Bantu, Zulu (with all the linguistic, cultural and political complexities of defining the boundaries of such groups), not to mention the thousands of language groups spread across the continent. How then, can we have any sense of reliability or validity for a claim that says someone is 80 per cent African – when the baseline for that claim is based upon the transparent scaffolding of chance – not purposive sampling?

Yet, when taken together, we are told that these markers appear to yield sufficiently distinctive patterns in those continental populations tested. So now we see how a specific pattern of genetic markers on each of a set of chromosomes that have a higher frequency in the 'Native Americans' sampled becomes established as a 'Native American' ancestry reference. (The fact that there are more than 480 different populations of the Tribal Council – the vast majority of which have never been sampled – is no small matter here, but that is not the focus of a separate critique I am about to make.) The problem is that millions of people around the globe will have a similar pattern – that is, they'll share similar base-pair changes at the genomic points under scrutiny. This means that someone from Bulgaria whose ancestors go back to the fifteenth century could (and sometime does) map as partly 'Native American', although

no direct ancestry is responsible for the shared genetic material. There is an overwhelming tendency for those who do AIMs analysis with the purpose of claims about ancestry to arbitrarily reduce all such possibilities of shared genotypes to 'inherited direct ancestry'. In so doing, the process relies excessively on the idea of 100 per cent purity, a condition that could never have existed in human populations.

While this is a huge problem, yet another issue looms even larger. If a computer program produces an outcome indicating that 35 per cent or more of a particular genetic marker exists in population A (let's call them East Asian), while 35 per cent or less occur in population B (let's call them European), the researcher may use that marker to say that someone is from East Asian ancestry. To make matters even more complicated, claims about how a test subject's patterns of genetic variation map to continents of origin and to populations where particular genetic variants arose, require that the researchers have 'reference populations'. The public needs to understand that these reference populations comprise relatively small groups of contemporary people. Those groups sampled may have migrated over several centuries, and thus these researchers must make many untested assumptions in using these contemporary groups to stand as proxies for populations from centuries ago, whether putatively representing a continent, a region, or a linguistic, ethnic or tribal group. To construct tractable mathematical models and computer programs, researchers bracket these assumptions about ancient migrations, reproductive practices, and the demographic effects of historical events such as plagues and famines. Given these intractable barriers to even low-level probabilistic reliability, geneticists are on demonstrably thin ice telling people that they do or don't have ancestors from a particular population.

Thus, instead of asserting that someone has no Native American ancestry, the most truthful statement would be: It is possible that while the Native American groups we sampled did not share your pattern of markers, others might since these markers do not exclusively belong to any one group of our existing racial, ethnic, linguistic, or tribal typologies. But computer-generated data provide an appearance of precision that is dangerously seductive and equally misleading. We cannot conclude that an individual has a close affinity to a particular ethnic or racial group or local geographical population simply because their DNA markers match that population.

> Such a conclusion would require demonstrating that the DNA sequence is not present in other places, it would require demonstrating that the gene pool of that ethnic group or local population had been close and immobile for centuries and millennia . . . (Weiss and Long 2009)

Despite these caveats about how much remains black-boxed assumptions, the last decade has witnessed an explosion of articles published in science

journals, attempting to explain the role of genetic admixture in diabetes, asthma, obesity and a number of other health outcomes (Moreno-Estrada et al. 2014; Fernandez and Shriver 2004). Here is an example of the kind of framing language used in such studies:

> There have been several recent reports of significant associations between genetic admixture and obesity-related traits in admixed populations. For example, Williams and colleagues reported a significant negative association between European admixture with body mass index (BMI kg/m^2) and fasting glucose measurement, suggesting genetic susceptibility for both diabetes and obesity in a sample of nondiabetic Pima Indians. (Fernandez and Shriver 2004)

We shall return to this matter of competing explanations of the high rate of diabetes among the Pima, but first we need to set the stage for the second element, the 'other half' of the two parts that synergistically interact as a 'gel' hardening the molecular reinscription process.

Navigating with race while trying to navigate around race

A significant wing of the Biological Sciences has found an unusual and effective way around the problem of confronting the matter of 'race as a biological category'. The strategy is to *not* deal with race in a full-scale case-control design, but to 'back into' a clinical study that was never designed to test whether race plays any role, only to discover *ex post facto* that the race of the clinical population, however defined, played a role in drug efficacy. Simply deploying racial categories of already collected data sets (by race) is one strategy that permits researchers to conveniently circumnavigate the problem of having to define terms. After all, to do a case control study would require the researcher to define terms and to specify the boundaries of the relevant populations. For 'race' this would be a knotty problem these days, with self-report of racial designation the primary criterion for classification.

There is yet a more subtle method of navigating around the problem of defining race, and it has become an increasingly standard operating procedure – the deployment of the idea of 'admixture'. Of course the irony resides in the routine practice of treating four continental ancestral populations as the basis of admixture – and these four populations align with Africa, Europe, Asia, and the Americas before the arrival of Europeans and have the putative 'purity' prior to admixture (Fernandez and Shriver 2004; TallBear 2014; Fullwiley 2006).

The second part, or: the 'good intentions' path from health concerns to molecular reinscription of race

In the last decade, there has been a peculiar and fateful irony in the convergence of the desire (and pressure) to use genetics to improve our health, and the decision by the US Congress to require that the National Institutes of Health record data and engage in research to lessen the health disparities between racial and ethnic groups. In 2000, Congress passed the Minority Health and Health Disparities Research and Education Act of 2000 #106-525, which mandated the National Institutes of Health to support research on health disparities between groups categorized by race and ethnicity.

As a direct consequence, the last decade has generated a sharp increase in articles that report health disparities between members of the majority white population and the various groups racially and ethnically designated. That was to be expected. Moreover, since the National Human Genome Research Institute is a branch of the National Institutes of Health, it would follow that research on human genetics would enter the fray, with scientists poised and ready to assert the unique contribution of molecular genetic differences to an explanation of these health disparities. For example, because the rate of prostate cancer in African Americans is more than double that of white Americans, it was inevitable that some would attempt to explain this through the lens of genetics. This in turn would lead down the path that would serve to rescue old racial taxonomies and their relationship to genetic profiles and genetic conditions. It was not expected that, in so doing, this strategy would inadvertently resuscitate the idea that genetic differences between those we place in racial categories might well explain different health outcomes. This is territory fraught with minefields for obvious reasons, dating back to the eugenics movement in the USA and its promulgation and extension into Nazi Germany (Proctor 1999; Kuhl 1994; Reilly 1991; Kevles 1985; Ludmerer 1972; Haller 1963).

To navigate around such problems, research scientists have developed two strategies. The first is to use only an *ex post facto* deployment of race and ethnicity, after data have been collected in large clinical trials in which race was not a defining category of selection. Two recent examples of this are the coming to market of racialized drugs, BiDil (designed for hypertension relief and congestive heart failure amelioration in African Americans) and Iressa (a late stage drug for treatment of lung cancer in Asians) (Kahn 2013; Sun 2012; Roberts 2011). The second strategy of navigation around the explosive combination of race-genetics-disease is to propose a large-scale research project aimed at determining the degree of genetic vs. environmental sources of health problems. Of course this formulation presumes that one could successfully 'partial out' the genetic from the environmental. Yet, it is one of the most

fundamental axioms of the contemporary life sciences that there is an interactional element at the core of the relationship between genes-and-environment. To assert that some human health condition, trait, or behaviour is, say, 60 per cent genetic and 40 per cent environmental is to express a static and archaic version that is entirely a statistical artifact of this static assumption. Such a statistical technique of partialing out denies or obscures the profoundly interactional problem of race, genes, and disease – now generating one of the most vexing and visceral debates in contemporary science. In order to understand why, we must first get a handle on what it means to 'isolate' race as a variable (or constant) in a research protocol.

Reifying biological race and ethnicity to explain diabetes and obesity

> There is no race which is so subject to diabetes as the Jews', wrote W. H. Thomas in 1904 . . . a New York physician who was voicing an almost universally held belief in the USA that of all the 'races', Jews had the greatest likelihood of developing diabetes. At the same time, most members of the medical community considered the prevalence of diabetes among Blacks to be unusually low. In the words of a Johns Hopkins physician in 1898, 'Diabetes is a rare disease in the colored race. (Tuchman 2011:24)

A century later, things have changed dramatically. Jews are now routinely categorized together with other Americans of European descent as white – and whites have less than half the rate of diabetes than African Americans. The link between Jews and diabetes had its origins in the European medical literature, and most particularly in the late nineteenth century writings of Joseph Seegen of Vienna. Tuchman reports that

> after Seegen noted in 1870 that roughly one quarter of his 140 diabetes patients were Jewish, other studies started appearing alleging that Jews died of diabetes at a rate between two and six times higher than the rest of the population. (Tuchman 2011: 25)

In the German literature, diabetes even came to be known as the *Judenkrankheit*, or 'Jewish disease'.

> When J. G. Wilson, a physician with the US Public Health Service, tried to understand why the diabetes mortality rate in New York City had tripled between 1889 and 1910, he compared the rapid growth in the city's Jewish population with the rise in the diabetes mortality rate. For Wilson, the correlation between these two sets of data was sufficient to demonstrate causation.

> To explain why Jews experienced such a high rate of diabetes, Wilson turned to racial traits, claiming that 'some hereditary defect' made the Jews more prone to develop the disease. He did not elaborate on the nature of the 'defect', but others pointed to the supposedly sensitive nervous system of the Jews. For Osler, it was the Jews' particularly 'neurotic temperament'; for the author of an article in the widely read Collier's Magazine, it was the Jews' 'racial tendency to corpulence. (Tuchman 2011: 25)

Although the Pima Indians of Arizona have long since replaced Jews as the group with the highest reported risk of diabetes, the method of recording a snapshot of corpulence (now cast as body-mass-index) continues the tradition of collecting cross-sectional data on the physical characteristics of the target population.

The question of how best to approach a strategy to increase the health of the disenfranchised and economically distressed is a very hotly contested issue, mainly because of the overlap of poverty, illness, ethnicity and race (Keller et al. 2012). This overlap has led some to the conclusion that there is something basically different in the bio-genetic make-up of different groups that might best explain health disparities. The Pima Indians have the highest rate of diabetes of any population ever studied, and they have become the subject of intense scrutiny and research as to why. 'More than half of the Pima older than 35 years of age have the disease, and the prevalence rates reach a peak of 86 per cent in women aged 55 to 64' (Johnson, Nowatzki and Coons1996). The prevalence rate increased by 42 per cent in the decade between 1967 and 1977 (Carter et al.1989). Here is an excerpt from an account of the approach supported by the National Institutes of Health that I have termed elsewhere 'looking inside the body' for answers:

Beginning in 1983 and continuing for 10 years, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) studied the genetic codes of almost 300 non-diabetic Pima Indians in great detail:

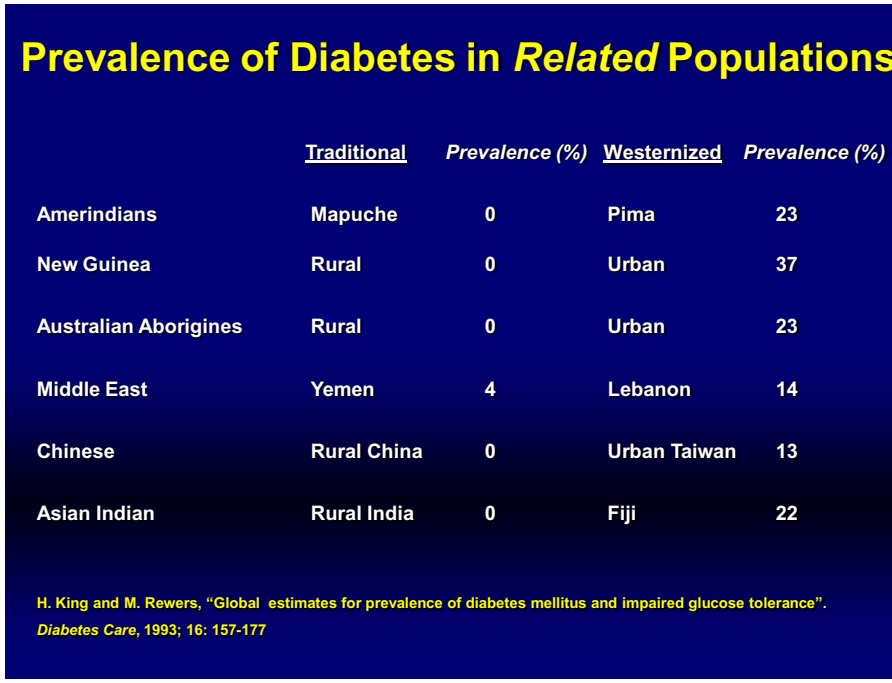
> We looked at body composition, how well a person produced insulin, how well that person's cells responded to insulin, and other factors. After a number of years, some of the volunteers developed diabetes and we were able to determine that insulin resistance and obesity were major predictors of disease', Dr. Bogardus explained.
>
> Because diabetes is such a complex disease, Dr. Bogardus and his staff are attempting to narrow their search by first looking for the genetic causes of physical conditions that can lead to diabetes, such as the genes that influence a person's cells to secrete less and respond less to insulin that is needed to regulate blood sugar.

> In 1993, they identified a gene called FABP2 that may contribute to insulin resistance. This gene makes an intestinal fatty acid binding protein using one of two amino acids. When the gene makes the protein with threonine, one of those amino acids, the body seems to absorb more fatty acids from the fat in meals. NIH scientists think that could lead to a higher level of certain fats and fatty acids in the blood, which could contribute to insulin resistance. (NIDDK 1996: 3)

On the matter of potential known effective interventions, this approach does not have (indeed, could not have) much of a track record. For at least the last half century, we have known that the rate of Type II diabetes among Native Americans is more than double that of whites in the USA. Until quite recently, the dominant theory among geneticists who have approached this topic has strongly suggested that Native Americans are far more likely to possess genes that enable fat-hoarding, sometimes labeled 'thrifty genes'. They hypothesized that these genes were conducive to adaptation because the ancestors had a need to survive during cycles of famine. However, since this group now lives in a world in which they can routinely ingest foods with high fat and sugar content, these putatively formerly protective/adaptive genes are now placing this population at greater risk for diabetes.

A recent study suggests that it was the high-fibre diet that 'locked in to place' the thrifty genes, not the adaptive mechanisms generated by famine cycles (Reinhard et al. 2012). Notice that in both accounts, genes are playing a dominant role in explaining 'health disparities'. Yet we have strong evidence that diet has far more analytic explanatory power, across all groups, when addressing the diabetes crisis that has struck across the globe since 1980. The fastest rate of increase is in India, which the World Health Organization has called the diabetes capital of world. Current estimates suggest that at least 35 million suffer now, and best estimates predict this figure will double in the next decade (Siegel, Venkat Narayan and Kinra 2008). In the last three decades, the growing middle classes of India have experienced an exponential increase in rates of diabetes, and 'thrifty genes' have less to do with this than the new capacity of the newly well-to-do to consume high levels of sugar in the countless ceremonies and ritual dinner celebrations that they can now afford.

To return to the extraordinarily high rate of diabetes among the Pima, an alternative to the genetic approach sets the analytic frame in a broader socio-historical context. Those who approach the matter from this angle have a very different view of how to think about diabetes prevention and treatment. In the graphic below on 'prevalence of diabetes in related populations' note the striking pattern of urban versus rural dwelling among six populations across the globe. Those who live in urban areas and consume a



westernized diet have a very high rate of diabetes, but those who have lived in 'traditional' sites where they practice 'traditional culture' hardly experience any diabetes:

Of course what is most striking about this table is that this pattern holds true for every group sampled, across a wide swath of the globe. Did they all have thrifty genes, and if so, what intervention is implied other than a dramatic shift in diet? Or from another perspective, since the sharp increase in Type II diabetes has come about in the last three decades – just in pure scientific logic, far more of the variance is explained by a systematic empirical investigation of shifting patterns of nutritional intake.

In a summary of the problems encountered when trying to explain sharp rates of difference between the tribes studied, the authors concluded:

> We have no data on relative rates of obesity or lifestyle differences that might explain the different rates of diabetes between the Pueblo tribes and the tribes of the Athabascan groups . . . Whether tribes of the Athabascan language group carry a genetic risk of diabetes different from that of the Pima and Pueblo Indians is unknown. (Carter et al. 1989)

A few years after this paper was published, complaining that there were no accounts of lifestyle among the various groups, a study was undertaken to

focus on lifestyle. Pima Indians using an outpatient hospital pharmacy in Southern Arizona were invited to participate (Johnson et al. 1996). Subjects were given a self-administered questionnaire that probed for demographic information and clinically relevant variables, and then were asked to take a short form version of a health survey. Notice the inverted parallel to the situation described in the introduction – where data collected on patients' heart condition was confined to the doctor's office. But while many of the heart patients assessed at routine check-ups were deemed otherwise healthy, the Pima in the study just noted here were restricted to those known to have diabetes. And while the title of the study had in it the name 'lifestyle' – the only data collected were within the confines of a medical establishment, a hospital pharmacy visit.

These are 'snapshots' or 'freeze-frame' accounts of the condition of a population at a single point in time. What happens when we step back and try to situate the Pima Indians' health crisis around diabetes within a larger socio-cultural context, literally situating this group outside the hospital – and investigating instead the 'natural setting' in which their lives have been shaped. A good ethnography begins with a socio-cultural history of the group being studied, and we can learn much that has direct relevance to their current high rates of diabetes from just a brief overview of that history.

At the end of the nineteenth century, the Pima were known to be superb farmers, self-sustaining and independent (DeJong 2007: 59). They had lived for centuries near the free-flowing Gila River, which supplied ample water for their agricultural needs. Indeed, they called themselves Akimel O'tham, or River People. However, the expansion of white settlers westward would dramatically change that lifestyle and force them to abandon water-intensive crops. By the turn of the century, thousands of these white settlers lived next to the Pima Reservation, and began competing directly for irrigation rights. As early as 1877, the Desert Land Act required that an applicant 'required bona fide application of water to the land to obtain title'. Ultimately dams were built that re-directed water away from traditional Pima farms, forcing residents to either abandon the area or shift to a different source of livelihood (DeJong 2007: 48). In 1902, a health survey found only a single case of diabetes among the Pima. Three decades later, the number had increased to more than 500. Then during the 1930s, the Coolidge Dam was completed. Although it was heavily touted to bring water to all, within a few years it became clear that the Pima were not to be the beneficiaries. There was certainly not enough water directed their way to restore traditional farming. Poverty was taking a heavy toll, with early deaths rising precipitously among the Pima, and the federal government embarked upon a programme to provide free government surplus food to the community.

Thus begins the substantive and compelling account of the dramatic increase of diabetes in this population. It was free food, but it was saturated

with a diabetic's nightmare: refined white flour, processed cheese, lard, candy and chips, refined sugar, grape juice, and lots of macaroni. Anthropologists monitoring the dietary circumstances noted that the diet of the Pima from an earlier times consisted of wild plants and game animals, when

> they used such foods as seeds, buds, fruits and joints of various cacti; seeds of the mesquite, ironwood, palo verde, amaranth, salt bush, lambsquarter, horsebean and squash; acorns and other wild nuts; . . . roots and bulbs of the sandroot (wild potato) . . . deer, antelope, ..rabbits, quail, dove, wild ducks, wild turkey. (Mark 1960: 46)

Fast forward to the middle of the twentieth century, when this diet was completely obliterated (Hackenberg 1962) – and in its place the Pima received boxes piled upon boxes of processed macaroni and cheese, where the larger the family size, the more entitlement to those free boxes of food. In the 1890s, the dietary intake of fat was 15 per cent, but by the 1990s it had nearly tripled to an eye-popping 40 per cent (NIDDKD 1996:19).

National genomic sovereignty

In the opening section, I noted that the history of why and how the USA has been in the vanguard of a movement to tailor therapies to particular sub-populations within its borders. That is changing rapidly around globe, with the emergence of something Benjamin (2009) has called 'National Genomic Sovereignty'

One of the most striking developments of the last few years has been the move by several governments to take strong protective 'ownership' of the DNA of their own populations – a move designed to protect from possible bio-piracy from the pharmaceutical industry in Western countries. This 'national genomic sovereignty' represents a pathway the very opposite of the universal notion of human DNA envisaged at the inception of the Human Genome Project:

> On the surface, this policy frame asserts a deeply nationalist sentiment of self-determination in a time of increasing globalization. It implicitly 'brands' national populations as biologically distinct from other populations, 'naturalizing' nation-state boundaries to ensure that less powerful countries receive the economic and medical benefits that may result from population genomics. (Benjamin 2009: 341)

Mexico amended its General Health Law in 2008 to make 'the sampling of genetic material and its transport outside of Mexico without prior approval . . . illegal' (Séguin et al. 2008: 6).

> The Genomic Sovereignty amendment states that Mexican-derived human genome data are the property of Mexico's government, and prohibits and

> penalizes its collection and utilization in research without prior government approval. It seeks to prevent other nations from analyzing Mexican genetic material, especially when results can be patented, and comes with a formidable bite in the form of prison time and lost wages. (Benjamin 2009: 344)

Mexico may be in the vanguard in being so explicitly asserting its commitment to national 'genomic sovereignty', but the nation is hardly alone. India, China, Thailand and South Africa have all issued policy statements or passed legislation designed to develop national genomics infrastructure to benefit their populations (Séguin et al. 2008).

In 2009, the HUGO Pan-Asian SNP Consortium, an international research team led by Edison Liu of the Genome Institute of Singapore, mapped genetic variation and migration patterns in 73 Asian populations, with data coming from 11 Asian countries: Japan, Korea, China, Taiwan, Singapore, Thailand, Indonesia, Philippines, Malaysia, Thailand, and India. The results – which included a summary statement that 'there is substantial genetic proximity of SEA [Southeast Asian] and EA [East Asian] populations' – were published in the journal *Science* (Hugo Pan-Asian Consortium 2009). In the same year, the Iressa Pan-Asian study (IPASS) was carried out by researchers in Hong Kong, mainland China, Thailand, Taiwan, and Japan with the participation of 87 centres in 9 countries in Asia (Sun 2012; Mok et al. 2009). This study was the result of previous research suggesting that Asian populations have a different, more positive response to this cancer drug, than do other populations.

The segue to forensics and criminal justice and 'molecular race'

There is a yet more ominous and troubling element of the reliance upon DNA analysis to determine who we are in terms of lineage, identity, and identification. The very technology that purports to tells us what proportion of our ancestry can be linked, proportionately, to sub-Saharan Africa is the same being offered to police stations across the USA to 'predict' or 'estimate' whether the DNA left at a crime scene belongs to a white or black person. As with research of health disparities, this 'ethnic estimation' using DNA relies on a social definition of the phenotype, that is, the observable physical or biochemical characteristics of an organism, determined by both genetic makeup and environmental influences. As noted above, any molecular, population, or behavioural geneticist who uses the term 'per cent European' or 'per cent Native American' is obliged to disclose that the measuring point of 'purity' (100 per cent) is a statistical artifact that begins not with the DNA, but with a researcher's adopting the folk categories of race and ethnicity, then determining if all four grand-parents of the subject originated in the same continent-of-origin as the basis for including the subject in the database.

Racial and ethnic markers in forensic DNA – and molecular photo-fitting

In the July 8, 1995 issue of the *New Scientist* entitled, 'Genes in Black and White', some extraordinary claims were made about what it is possible to learn about socially defined categories of race from reviewing information gathered using new molecular genetic technology. In 1993, a British forensic scientist published what is perhaps the first DNA test explicitly acknowledged to provide 'intelligence information' along 'ethnic' lines for 'investigators of unsolved crimes'. Ian Evett, of the Home Office's forensic science laboratory in Birmingham, and his colleagues in the Metropolitan Police, claimed that their DNA test can distinguish between 'Caucasians' and 'Afro-Caribbeans' in the vast majority of cases.

Evett's work (Evett et al. 1993), published in the *Journal of Forensic Science Society*, drew on apparent genetic differences in three sections of human DNA. Like most stretches of human DNA used for forensic typing, each of these three regions differs widely from person to person, irrespective of race. But by looking at all three, the researchers claimed that under select circumstances it is possible to estimate the probability that someone belongs to a particular racial group. The implications of this for determining, for practical purposes, who is and who is not 'officially' a member of some racial or ethnic category are profound.

The legal and social uses of these technologies are already in use, and here are some examples: In the early 1980s several states in the USA began keeping DNA database files for sexual offenders. Three factors converged to make this a popular decision by criminal justice officials that would be backed by politicians and the public because: 1) sex offenders are those most likely to leave body tissue and fluids at the crime scene, 2) they rank among the most likely repeat offenders, and 3) their crimes are often particularly reprehensible in that they violate persons, from rape to molestation and abuse of the young and most vulnerable. Today, all fifty states in the USA store DNA samples of sex offenders, and most states do the same for convicted murderers. But by 2006 thirty-four states were storing DNA samples of all felons (Krimsky and Simoncelli 2011).

On January 5, 2006, the President of the USA signed into law HR 3402, the Department of Justice Reauthorization bill of the Violence Against Women Act of 2005. This legislation for the first time permits state and federal law enforcement officials the right to transfer DNA profiles of those merely arrested for federal crimes into the federal Combined DNA Index System (CODIS) database. Previously, only convicted felons could be included. Those DNA profiles will remain in the database unless and until those who are exonerated or never charged with the crime request that their DNA be expunged. Thus the default will be to store these profiles, and expunging requires the proactive agency (and resources) of those arrested.

But there is reason to be wary of these developments and vigilant about the uses of expanding DNA databases. Criminologists and statisticians have provided enough convincing evidence that reliability may be a systemic issue with regard to 'exact matches', leading to false 'hits' with traditional approaches (Thompson 2008). As for the possibility of using full DNA samples for forensic research, attempts to determine physical features, such as skin colour, hair texture, and eye pigment, have already been made (Fullwiley 2014, 2008). These techniques, commonly referred to as 'molecular photo-fitting', rely on 'admixture estimates' discussed earlier, and are rife with reliability issues despite their veneer of exact precision with regard to continental genetic affinity, or, put bluntly, racial diagnosis. This kind of categorizing of subjects and patients is occurring in medical and health journals, often with the idea that pharmaceuticals could be tailored to patients based on putative notions of their ancestral genetic 'admixture'. Researchers are also finding new ways to identify genetic variants related to 'admixed' populations that they believe may be 'linked' to variable complex disease conditions, such as end-stage renal disease (Kao et al 2008). Selected areas of the genome are designated to be ancestrally more typically African or European with very little attention to the complex set of historical events (migration, wars and conquest) that have shaped and reshaped determinations of putative purity.

There is a recurring theme in the biosciences, a tendency to admit that previous eras of scientific work was flawed because it was immersed in the social and political issues of a bygone era. At the end of the nineteenth century, scientists were able to distance themselves from the pro-slavery arguments of their predecessors who published scholarly articles showing how and why black's lung capacity made them suitable for slavery (Braun 2014). These were the eugenicists – but by mid-twentieth century, it was possible for the biosciences to disavow eugenics as flawed or bad science. Every era tends to assert that its science somehow levitates above the social, economic, political forces of the day – and is free from such immersion:

> The fact that an idea has been misused does not mean it should be forgotten or that it was wrong. It is vitally important to realize that horrific though these atrocities were, they were *not* based on science. They were based on the prejudices and psychopathic policies of people in power who decided to misquote science to an uneducated public to fulfill their own immoral agendas . . . it was the public's ignorance of the true facts that allowed such people to use these misrepresentations as weapons. (Italics in original). Anderson (2007: 5)

But Charles Davenport, an ardent proponent of eugenics, was one of the leading geneticists of his era, and the director of Cold Springs Harbor Laboratory. Eugenics was only called pseudo-science in hind-sight, 30–50 years later (Kevles 1985)! And so I am confidently predicting that in the year

2064, at the annual meetings of the American Association for the Advancement of Science, many will acknowledge that way back in the first part of the twenty-first century scientists were caught in the social fabric of their era, often inadvertently deeply mired in the prevailing ideas of racial admixture as the natural order. They will certainly admit that, 'way back then, half a century ago' big pharmaceutical companies shaped the direction of much of scientific research. They will point to the role of heavy investments in these companies, the push to understand the high rate of diabetes among the Pima as explained by their genetic make-up, of their proportional admixture.

As funding for research to address health disparities moves in the direction of the biological or genetic emphasis (as with diabetes and the Pima), or even to gene-environment interaction long-term studies, social scientists have three major choices. The first is to stand along the side-lines and continue to cite the mantra that race and ethnicity are socially constructed. While demonstrably true, that truth is overwhelmed by the megaphone of advancing, under-theorized admixture research and increasingly taken-for-granted scientized components (of 100 per cent statistical purity) undergirding admixture. The second choice is to accept new molecular reinscription of race and join in joint projects – uncritical and without examining the domain assumptions that have 'hardened' and gelled. The third choice is the one that I strongly recommend – namely, that social analysts go to the site of the production of knowledge, and closely examine the procedures, the domain assumptions of how race is being used in human molecular genetics, examine how heavily these assumptions are located in social, historical, and folk categories but are then transmogrified into the language of science and anointed with an imprimatur of legitimacy. Social scientists do have a few good models of how to better engage and re-balance the debates about the proper role of race in science, medicine and law. We need look no further than the works of TallBear (2013) and Fullwiley (2007; 2008), each of whom has dissected the domain assumptions behind Ancestry Informative Markers; or of Bolnick (2008), who has examined and critiqued the social assumptions behind the use of the computer program Structure; or Fujimura and Rajagopalan (2011) who have dissected the algorithms of the EIGENSTRAT technology. Here is an example of the important findings from this kind of investigative ethnography:

> . . . on the construction and implementation of EIGENSTRAT, a population genetics software technology. We illustrate how some biomedical researchers use EIGENSTRAT to avoid emphasizing populations in their search for disease-related DNA and certainly to avoid the use of race. We also show, however, how other researchers using EIGENSTRAT find it difficult to give up on geographically 'locating' DNA and designating populations; that is,

> they move from *genetic similarity* to *genetic ancestry* to *genome geography*. (Fujimura and Rajagopalan 2011)

Each of these researchers has spent time in either the labs, or in careful exegesis of the scaffolding of laboratory work, or a close re-analysis of computer-based programs and algorithms. With these kinds of empirical investigations of the architecture of thought behind the advancing biologistic reductionism of human taxonomies, we will be far better equipped to have a serious debate about the limitations and diversions that inhere in the molecular reinscription of race and ethnicity.

(Date accepted: November 2014)

Notes

1. The following few pages are a condensed excerpt from a more expanded analysis of this topic in Duster (2006)

2. The next few pages of this manuscript are adapted from my longer discussion of this issue (Duster 2014)

Bibliography

Anderson, Gail S. 2007 *Biological Influences on Criminal Behavior*, New York and London: CRC Press, Taylor & Francis.

Benjamin, Ruha 2009 'A Lab of Their Own: Genomic Sovereignty as Postcolonial Science Policy', *Policy & Society* 28(Fall): 341–55.

Boeschenstein, N. 2006 'The Charlottesville Dragnet, Part I: Just the Facts, Ma'am', Archipelago, 1-8, available at http://www.archipelago.org/vol8-2/boeschenstein.htm [last visited 9 September 9 2014].

Bolnick, Deborah 2008 'Individual Ancestry Inference and the Reification of Race as a Biological Phenomenon' in Barbara A. Koenig, Sandra Soo-Jin Lee and Sarah S. Richardson (eds) *Revisiting Race in a Genomic Age*, Piscataway NJ: Rutgers University Press.

Braun, Lundy 2014 *Breathing Race in the Machine: The Surprising Career of the Spirometer from Plantation to Genetics*, Minneapolis, MN: University of Minnesota Press.

Bullard, Robert D. 2000 *Dumping in Dixie: Race, Class, and Environmental Quality*, Boulder, Colo: Westview Press.

Carter, J., Horowitz, R., Wilson, R., Sava, S., Sinnock, P. and Gohdes, D. 1989 'Tribal Differences in Diabetes: Prevalence Among American Indians in New Mexico' Reviewed Work(s), *Public Health Reports* 104(6): 665–9.

Cooper, R.S., Wolf-Maier, K., Adeyemo, A., Luke, A., Banegas, J.R., Forrester, T.E., Giampaoli, S., Joffres, M., Kastarinen, M., Primatesta, P., Stegmayr, B. and Thamm, M. 2005 'An International Comparison Study of Blood Pressure in Populations of European vs. African Descent', *BioMed Central* 3(22) URL http://www.biomedcentral.com/1741-7015/3/2:11

Daily, Emma 2005 'DNA Tells Students They Aren't Who They Thought', *The New York Times* [online] April 13.

DeJong, D. 2007 'Abandoned Little by Little: The 1914 Pima Adjudication Survey, Water Deprivation, and Farming on the

Pima Reservation', *Agricultural History* 81(1(Winter)): 36–69.

Duster, Troy 2005 'Race and Reification in Science', *Science* 307(18): 1050–1.

Duster, Troy 2006 'The Molecular Reinscription of Race', *Patterns of Prejudice* 40: 427–41.

Duster, Troy 2014 'Social Diversity in Humans: Implications and Hidden Consequences for Biological Research' in Aravinda Chakravarty (ed.) *Perspectives on Human Variation*, Cold Spring Harbor Press.

Epstein, Steven 2007 *Inclusion: The Politics of Difference in Medical Research*, University of Chicago Press, Chicago, Illinois.

Evett, I.W. 1993 'Criminalistics: The Future of Expertise', *Journal of the Forensic Science Society* 33(3): 173–8.

Evett, I.W., Buckleton, I.S., Raymond, A. and Roberts, H. 1993 'The Evidential Value of DNA Profiles', *Journal of the Forensic Science Society* 33(4): 243–4.

Fernandez, Jose R. and Shriver, Mark D. 2004 'Using Genetic Admixture to Study the Biology of Obesity Traits and to Map Genes in Admixed Populations', *Nutrition Reviews* 62(July Issue Supplement s.2): S69–S74.

Fujimura, Joan H. and Rajagopalan, Ramya 2011 'Different Differences: The Use of "Genetic Ancestry" Versus Race in Biomedical Human Genetic Research', *Social Studies of Science* 41(1): 5–30.

Fullwiley, Duana 2006 'Biosocial Suffering: Order and illness in Urban West Africa', *BioSocieties* 1(4): 421–38.

Fullwiley, Duana 2007 'The Molecularization of Race: Institutionalizing Human Difference in Pharmacogenetics Practice', *Science as Culture* 16(1): 1–30.

Fullwiley, Duana 2008 'The Biologistical Construction of Race: Admixture Technology and the New Genetic Medicine', *Social Studies of Science* 38(5): 695–735.

Fullwiley, Duana 2014 'The "Contemporary Synthesis": When Politically Inclusive Genomic Science Relies on Biological Notions of Race', *ISIS, Journal of the History of Science Society* 105(4(December)): 803–14.

Gilroy, Paul 2000 *Against Race: Imagining Race Beyond the Color Line*, Cambridge, Mass: Belknap Press of Harvard University.

Gordon-Reed, Annette 2008 *The Hemingses of Monticello: An American Family*, New York: W.W. Norton.

Hackenberg, R. 1962 'Economic Alternatives in Arid Lands: A Case Study of the Pima and Papago Indians', *Ethnology* 1(2): 186–96.

Haller, Mark H. 1963 *Eugenics: Hereditarian Attitudes in American Thought*, New Brunswick, NJ: Rutgers University Press.

Hanson, M. 2004 'DNA Dragnet', *American Bar Association Journal* 90: 38–43.

Hayes, Tyrone B. 2010 'Diversifying the Biological Sciences: Past Efforts and Future Challenges', *Molecular Biology of the Cell* 21(15): 3767–9.

Hugo Pan-Asian Consortium 2009.

Johnson, J., Nowatzki, T.E. and Coons, Stephen Joel 1996 'Health-Related Quality of Life of Diabetic Pima Indians', *Medical Care* 34(2): 97–102.

Johnston, Josephine 2003 'Resisting a Genetic Identity: The Black Seminoles and Genetic Tests of Ancestry', *Journal of Law, Medicine and Ethics* 31(November): 262–71.

Jones, James H. 1981 *Bad Blood: The Tuskegee Syphilis Experiment*, New York: Free Press.

Kahn, Jonathan 2013 *Race in a Bottle: The Story of BiDil and Racialized Medicine in the Genomic Age*, New York: Columbia University Press.

Kao, W.H. et al. 2008 '*MYH9* is Associated with Nondiabetic End-stage Renal Disease in African Americans', *Nat Genet.* 40(10): 1185–92.

Kevles, Daniel J. 1985 *In the Name of Eugenics: Genetics and the Uses of Human Heredity*, New York: Knopf.

Keller N., Bhatia, S., Braden, J.N., Gildengorin, G., Johnson, J., Yedin, R., Tseng, T., Knapp, J.,Glaser, N., Jossan, P., Teran, S., Thodes, E.T. and Noble, J.A. 2012 'Distinguishing Type 2 Diabetes from Type 1 Diabetes in African American and Hispanic American Pediatric Patients', *PLoS ONE*

7(3): e32773. doi:10.1371/journal.pone.0032773.

King, H. and Rewers, M. 1993 'Global Estimates for Prevalence of Diabetes Mellitus and Impaired Glucose Tolerance', *Diabetes Care* 16: 157–77.

Krieger, Nancy 2011 *Epidemiology and the People's Health*, New York: Oxford University Press.

Krimsky, Sheldon and Simoncelli, Tania 2011 *Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties*, New York: Columbia University Press.

Kuhl, Stefan 1994 *The Nazi Connection: Eugenics, American Racism, and German National Socialism*, New York: Oxford University Press.

Ludmerer, Kenneth M. 1972 *Society, Genetics and American*, Baltimore and London: John Hopkins University Press.

Mark, A. 1960 'Ecological Change in the History of the Papago Indian Population', Master of Arts thesis, University of Arizona.

Mok, Tony S., Wu, Yi-hong, Thongprasert, Sumitra and Yang, Chih-Chih 2009 'Gefitinib or Carboplatin-Paclitaxel in Pulmonary Adenocarcinoma', *New England Journal of Medicine* 361(10): 947–57.

Moreno-Estrada, Andres et al. 2014 'The Genetics of Mexico Recapitulates Native American Substructure and Affects Biomedical Traits', *Science* 344(14): 1280–5.

National Institute of Diabetes and Digestive Kidney Diseases (NIDDKD) 1996 'The Pima Indians: Pathfinders for Health', Document Number NIH 95-3821, Bethesda, MD, National Institutes of Health.

New York Times, *Sunday Book Review* 2014 'Four Geneticists, Then 135 Signatures, Against Wade's Speculation', August 8.

Ossorio, Pilar N. 2009 'Myth and Mystification: The Science and Race of IQ', in S. Krimsky and K. Sloan (eds) *Race and the Genetic Revolution: Science, Myth and Culture*, New York: Columbia University Press.

Prainsack, Barbara and Machado, Helen 2012 *Tracing Technologies: Prisoners' Views in the Era of CSI*, Ashgate.

Proctor, Robert 1999 *The Nazi War on Cancer*, Princeton, NJ: Princeton University Press.

Reinhard, D.A., Konrath, S.H., Lopez, W.D. and Cameron, H.G. 2012 'Expensive Egos: Narcissistic Males Have Higher Cortisol', *PLoS ONE* 7(1): e30858. doi:10.1371/journal.pone.0030858

Reilly, Philip 1991 *The Surgical Solution: A History of Involuntary Sterilization in the United States*, Baltimore, MD: Johns Hopkins University Press.

Reverby, Susan 2009 *Examining Tuskegee: The Infamous Syphilis Experiment and its Legacy*, Chapel Hill: University of North Carolina Press.

Risch, Neil, Burchard, Esteband, Ziv, Elad and Tang, Hua 2002 'Categorizations of Humans in Biological Research: Genes, Race and Disease', *Genome Biology* 3(7): online at http://genomebiology.com/2002/3/7/comment/2007.

Roberts, Dorothy 2011 *Fatal Invention: How Science, Politics and Big Business Recreate Race in the 21st Century*, New York: The New Press.

Rothenberg, Karen 1997 'Breast Cancer, the Genetic "Quick Fix" and the Jewish Community: Ethical, Legal and Social Challenges', *Health Matrix* 7(1): 97–124.

Schütz, Alfred 1962 'Common-Sense and Scientific Interpretation of Human Action' in M. Natanson (ed.) *Collected Papers I*, The Hague: Nijhoff.

Séguin, B., Hardy, B., Singer, P.A. and Daar, A.S. 2008 'Genomics, Public Health and Developing Countries: The Case of the Mexican National Institute of Genomic Medicine (INMEGEN)', *Nature Reviews Genetics* 9: S5–S9.

Shriver, Mark D., Smith, Michael W., Jin, Li et al. 1997 'Ethnic-Affiliation Estimation by Use of Population-Specific DNA Markers', *American Journal of Human Genetics* 60(4): 957–64.

Siegel, Karen, Venkat Narayan, K.M. and Kinra, Sanjay 2008 'Finding A Policy Solution To India's Diabetes Epidemic', *Health Affairs* 27(4): 1077–90.

Sun, Shirley Hsiao-Li 2012 *Population Policy and Reproduction in Singapore: Making Future Citizens*, New York: Routledge.

Sze, Julie 2007 *Noxious New York: The Racial Politics of Urban Health and Environmental Justice*. Cambridge, MA: MIT Press.

TallBear, Kim 2013 *Native American DNA: Tribal Belonging and the False Promise of Genetic Science*, Minneapolis, MN: University of Minnesota Press.

TallBear, Kim 2014 'The Emergence, Politics and Marketplace of Native American DNA' in Daniel Lee Kleinman and Kelly Moore (eds) *Routledge Handbook of Science, Technology and Society*, New York: Routledge.

Thompson, William C. 2008 'Beyond Bad Apples: Analyzing the Role of Forensic Science in Wrongful Convictions', *Southwestern Law Review* 37(4): 1027–50.

Toom, Victor 2014 'Trumping Communitarianism: Crime Control and Forensic DNA Typing and Databasing in Singapore', *East Asian Science, Technology and Society: An International Journal* 8(3): 273–96.

Tuchman, A. 2011 'Diabetes and Race: A Historical Perspective', *American Journal of Public Health* 101(1): 24–33.

Wade, Nicholas 2014 *A Troublesome Inheritance: Genes, Race and Human History*, Penguin.

Wailoo, Keith 2011 *How Cancer Crossed the Color Line*, New York: Oxford University Press.

Weiss, Kenneth M. and Long, Jeffrey C. 2009 'Non-Darwinian Estimation: My Ancestors, My Genes' Ancestors', *Genome Res.* 19: 703–10.

Genes, Race, and Population: Avoiding a Collision of Categories

Jonathan Kahn, JD, PhD

A wide array of federal mandates have a profound impact on the use of racial and ethnic categories in biomedical research, clinical practice, product development, and health policy. Current discussions over the appropriate use of racial and ethnic categories in biomedical contexts have largely focused on the practices of individual researchers.

By contrast, our discussion focuses on relations between the daily practices of biomedical professionals and federal regulatory mandates. It draws upon the legal doctrine of equal protection to move beyond such debates and to propose guidelines to address the structural forces imposed by federal regulations that mandate how data about race and ethnicity are used in biomedical research. It offers a framework to manage the tension involved in using existing federally mandated categories of race and ethnicity alongside new scientific findings about human genetic variation. (*Am J Public Health*. 2006;96:1965–1970. doi:10.2105/AJPH.2005.067926)

CURRENT DISCUSSIONS ABOUT the appropriate use of racial and ethnic categories in biomedical contexts have largely focused on the practices of individual researchers. Individual research, however, takes place within larger structural contexts that shape how and when such categories get taken up, circulated, and applied. In particular, more consideration needs to be given to the impact federal regulatory mandates and incentives upon how biomedical professionals use racial and ethnic categories. Prominent among these mandates are requirements to use the social categories of race and ethnicity provided by the Office of Management and Budget for the collection of data for publicly funded research. Use of such social categories are heading for a collision with diverse categories of population that are classified in federally maintained genetic data bases. As genetic information becomes increasingly central to an ever-widening array of biomedical enterprises, the danger of improperly confusing or conflating social categories of race and ethnicity with genetic categories of population rises accordingly. Drawing analogies to the legal doctrine of equal protection, we offer a preliminary framework to begin discussion on how best to manage or avoid such collisions.

RACE AND ETHNICITY IN BIOMEDICINE

The recent Food and Drug Administration (FDA) approval of the drug BiDil with a race-specific indication to treat heart failure only in African Americans has brought to the fore a host of issues related to the use of racial and ethnic categories in biomedical research and drug development.[1] Because the BiDil application was premised on the activity of the drug at the molecular level in the trial subjects, the FDA approval has, in effect, given the imprimatur of the federal government to the use of race as a biological category.[2] Ironically, the FDA approval was based on a trial—the African-American Heart Failure Trial (A-HeFT)—that enrolled only self-identified African Americans. The results of this single-race design therefore precluded the investigators from making any claims regarding whether BiDil works differently in self-identified African Americans than in anyone else.[3]

The race-specific design of the A-HeFT trial is inextricably linked to the fact that its sponsors obtained a race-specific patent in 2000 for the use of BiDil in African Americans.[4] In granting the patent, the US Patent and Trademark Office provided an additional federal stamp of approval on the implicit use of race as a biological category. The federally granted patent also provided a powerful commercial incentive for the race-specific design of A-HeFT.[2]

The story of BiDil is significant because it marks the first race-specific application to the FDA. More broadly, it brings into high relief a powerful dynamic whereby federal regulatory incentives and directives promote the increasing use of racial and ethnic categories in a biomedical context. In the case of efforts to address well-documented disparities in health outcomes, such use, although complicated, does not necessarily imply a biological or genetic difference between races.[5] In the context of seeking the causal molecular basis for certain diseases, as in much drug development, the use of racial and ethnic categories as surrogates for genetic markers presents more problematic issues. Some researchers believe correlations between racial/ethnic and genetic categories can serve as useful research tools[6,7]; others contest the rigor and utility of such purported correlations, arguing that they risk naturalizing race and ethnicity as somehow genetic.[8–10] When federal approval is sought for such uses, the power of the state becomes implicated in marking racial or ethnic differences as genetic.

Over the past several years, recurring controversies have arisen among scientists and biomedical professionals regarding the nature of the relation, if any, between genes and race.[11] A host of articles has been published in the attempt to help biomedical researchers clarify their use of the concepts of race and ethnicity in general[9,12,13]; some specifically relate to genetically based concepts of population.[14–16] Several biomedical journals have published policy statements or guidelines concerning the use of racial and ethnic categories.[17–19] These articles and related debates over how, when, or whether to use race and ethnicity in biomedical

research are targeted at the practices of researchers themselves.

To date, however, such articles have largely overlooked the fact that research practices involving the use of racial and ethnic categories are profoundly shaped by federal regulatory incentives and guidelines. Thus, before proceeding with further debate about their own scientific practices, biomedical researchers and clinicians need to consider more fully and systematically the role of the federal government in shaping such practices. The recent proliferation of biomedical research that uses race and ethnicity as variables did not spontaneously emerge from a sudden discovery of their relevance. Rather, from funding requests to drug approval and market protection, specific federal initiatives mandating the use of such categories have played a critical role in promoting their inclusion as variables in biomedical research.

FEDERAL MANDATES

Prominent among these federal mandates are the National Institutes of Health (NIH) Revitalization Act of 1993, which directed the NIH to develop guidelines for including women and minorities in NIH-sponsored clinical research,[20] and the Food and Drug Modernization Act of 1997, which directed the FDA to examine issues related to the inclusion of racial and ethnic groups in clinical trials of new drugs.[21] Pursuant to these mandates, the NIH and FDA have issued detailed guidelines and guidance mandating certain procedures and practices concerning the inclusion of ethnic and racial minorities in clinical trials.[22,23]

Thus, for example, the NIH "Policy on Reporting Race and Ethnicity Data" states, inter alia, that the "NIH requires all grants, contracts, and intramural projects conducting clinical research to address the Inclusion of Women and Minorities. . . . Investigators are instructed to provide plans for the total number of subjects proposed for the study and to provide the distribution by ethnic/racial categories and sex/gender."[24] Similarly, the FDA recommended that individuals or corporations submitting drug approval applications "collect race and ethnicity data for clinical study participants."[25] These mandates impose significant requirements and provide incentives to identify and collect research data according to categories of race and ethnicity.

The federally mandated racial and ethnic categories, however, are not biomedical in origin. Rather, they derive from the 1997 "Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity" published by the Office of Management and Budget (OMB).[26] These standards set forth 5 minimum categories for data on race: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White. There are 2 categories for data on ethnicity: Hispanic or Latino and Not Hispanic or Latino. These categories provide the basis for the classification of all federal data on race and ethnicity, most notably, the census.

The OMB standards, however, contain an important caveat: "The racial and ethnic categories set forth in the standards should not be interpreted as being primarily biological or genetic in reference."[26] These categories were developed to serve social, cultural, and political purposes. When the federal government requires biomedical researchers and clinicians to import these social categories into explicitly biological or genetic contexts, it is creating a structural situation in which social categories of race and ethnicity may easily become confused and may be conflated with biological and genetic categories in day-to-day practice.

GENETIC DATABASES

Since the advent of the federally sponsored Human Genome Project in 1990, increasing knowledge of genetics has been transforming biomedical research. This research, however, often involves protocols that have been designed in response to federal mandates to incorporate the social categories of race and ethnicity defined by the OMB. The protocols compel researchers and clinicians to juggle genetic categories alongside racial and ethnic categories in the same conceptual and physical space. This creates a situation that facilitates and even promotes the conflation of genetic categories of population with social categories of race and ethnicity—it is an accident waiting to happen.

Already existing federally supported genetic databases add to the confusion through their own problematic uses of racial and ethnic categories. Thus, for example, the National Institute of General Medical Science/Coriell Cell Repository maintains a Human Variation Collection of genetic samples organized into the following broad categories: North America/Caribbean, South America, Europe, Asia/Pacific, Africa, and Middle East. Within these broad categories, subdivisions are made with diverse and potentially inconsistent classifications that include White, Basque, Mexican American Community of Los Angeles, Southeast Asians (excluding Japanese and Chinese), Quechua—South Central Andes of Peru, Africans South of the Sahara, Ashkenazi Jews, Czechoslovakian, and Northern European.

Implicated in these various categories are sometimes-overlapping concepts of ethnicity, race, continental geography, regional geography, geopolitical nation-states, urban ethnic communities, religion, geographic isolation, and endogamous indigenous populations.[27] (The contingency of the Czechoslovakian category is particularly notable, because there is no longer a geopolitical entity known as Czechoslovakia.) The National Center for Biotechnology Information maintains a separate database of genetic information known as dbSNP, which similarly organizes its data into population classes that mix geography, nationality, race, and ethnicity.[28]

A major new federal initiative, the International Haplotype Map Project,[29,30] promises to exacerbate this problem. The project has devoted more than $100 million to charting blocks of genetic variation in the human genome.[31] This otherwise-laudable effort, intended to help researchers identify genetic variations related to health and disease, may inadvertently be opening the door to further confusion of racial and ethnic categories with genetic groupings. The initial phase of the project has been structured around 270 tissue samples taken from Yorubas in Nigeria, Japanese, Han Chinese, and individuals of western and northern European descent in the United States.

The resulting blocks of variation are being identified with their source population.[32] The population groups are already being characterized as representative of the broad continental population groups of Africa, Asia, and Europe.[32] The stated rationale is that although "most of the common haplotypes occur in all human populations . . . their frequencies differ among populations. Therefore, data from several populations are needed to choose tag SNPs [single nucleotide polymorphisms]."[32] One can readily see how such genetic categories are ripe for conflation with the social/bureaucratic categories of race and ethnicity promulgated by the OMB.

As population-identified genetic information increasingly comes online for use from the Haplotype Map Project and other federally maintained databases, the need to provide a structuring mechanism to keep genetic categories in a socially responsible and scientifically appropriate relation to social categories of race and ethnicity will become ever more pressing.

INTRODUCING AN EQUAL PROTECTION MODEL

The various attempts to provide guidance to researchers on how, when, and whether to use race and ethnicity in their work are important, but they are not enough. Researchers can and should be able to decide how they choose to pursue their particular research agendas. But for years now, researchers and clinicians have been working under a variety of federal mandates that influence how, when, and whether they use racial and ethnic categories in their work. The time has come to examine those mandates and focus on them—rather than on the researchers and clinicians—as targets for constructive intervention.

Previous attempts to articulate best practices for using racial and ethnic categories in biomedical research and clinical practice have largely involved discussions among social scientists, natural scientists, and medical professionals about how best to characterize and manage the social and scientific meaning of these categories. Largely absent from these considerations, however, has been an alternative approach with a long tradition of assessing how best to characterize and manage such classifications in a regulatory context: equal protection law.

Equal protection doctrine derives from the 14th Amendment to the US Constitution, which declares that "no State shall make or enforce any law which shall deny to any person within its jurisdiction the equal protection of the laws." Equal protection doctrine is used to evaluate state-mandated use of racial categories in areas such as school desegregation and affirmative action. Although this doctrine is not necessarily directly applicable to the context of federal practice guidelines or regulatory approvals,[33,34] over the decades, courts and legal commentators have devoted considerable attention to developing guidelines and standards to assess and evaluate how racial and ethnic categories may be used appropriately to achieve specific goals. Under equal protection doctrine, race is considered to be a suspect classification because of a US history of racial oppression and the structural vulnerability of racial minority groups. Therefore, the state must justify the use of a racial classification by demonstrating that the classification is "narrowly tailored to serve a compelling state interest."[35] This is called strict scrutiny. It requires a tight fit between the classification and the purpose or interest it serves to force out potentially invidious motivations behind the use of race in law.

Concepts from equal protection analysis may be adapted to a biomedical context through comparison with biomedical analogues already in use in federal regulation of racial and ethnic classifications in research and clinical practice. Thus, for example, NIH guidelines for grant applicants and contract solicitations already require the inclusion of "a description of plans to conduct analyses to detect *significant differences* in intervention effect by sex/gender, racial/ethnic groups, and relevant subpopulations, if applicable [italics added]."[36] The guidelines go on to define significant difference as "a difference that is of clinical or public health *importance*, based on *substantial* scientific data [italics added]."[36] Similarly, the guidelines require such submissions to "include a description of plans to conduct *valid analysis* by sex/gender, racial/ethnic groups, and relevant subpopulations, if applicable [italics added]."[36]

Significant difference and valid analysis, like the equal protection standard of "narrow tailoring to serve a compelling state interest," involve terms of art that have been used constructively to manage racial and ethnic categories in diverse contexts. They have been defined over time and applied through an accretion of understanding, practice, and interpretation developed by the relevant professional communities. The model of equal protection analysis can be adapted to a biomedical context by using analogous concepts such as significant difference and valid analysis to evaluate the rigor and legitimacy of uses of racial/ethnic classifications in relation to genetics.

Equal protection doctrine thus provides a useful model for developing guidelines to improve already existing and pervasive federal mandates governing the management of race and ethnicity in regulatory contexts. In addition to exposing possible invidious motives, heightened scrutiny can bring to light well-intentioned but careless or inconsistent use of racial and ethnic classifications.

To this end, I offer the following preliminary recommendations to consider in revising relevant federal mandates to address the use of race and ethnicity in biomedical research and clinical practice. They are organized sequentially to parallel a general research plan of project conceptualization, design, and implementation. These recommendations might be thought of as a regulatory analogue to the sort of guidelines on the use of racial and ethnic categories currently being considered and adopted by some biomedical and scientific journals. They attempt to adapt or transpose the conceptual apparatus of equal protection law into the domain of biomedical research and clinical practice. I hope that they will provide the groundwork for further discussion of how federal mandates might be revised to help biomedical professionals keep genetic categories of population and social categories of race and ethnicity in a constructive relation to one another.

RECOMMENDATIONS

Federal regulations, mandates, guidelines, or other similar directives relating to federal funding, regulatory approval, or intellectual property protection for biomedical research and related products should be revised to require applications and related documents submitted to federal agencies that use or make claims on the basis of racial or ethnic categories to include the following:

Definitions

Population. Require a clear definition of the source of any population category being used, its scope, and its limits. Specify whether or to what extent shared biology or genetics is presumed to underlie the population classification chosen and the degree to which the classification also implicates nonbiological values (e.g., nationality, race/ethnicity, religion, mere proximity).

Race/ethnicity. Require a clear recognition of the requirements of the OMB revised standards regarding the selection and use of racial/ethnic categories and an explicit statement of the social basis of those categories. This may take the form of including the OMB caveat: "The racial and ethnic categories set forth in the standards should not be interpreted as being primarily biological or genetic in reference."

Rationale. The OMB revised standards establish basic categories of race and ethnicity, but they do not dictate specifically how those categories are to be used or interpreted in different contexts. Thus, in practice these categories are often merely starting points and are often elaborated upon and modified. The requirement of definition allows researchers and clinicians to adapt these categories to their own particular needs. It also ensures that from the outset such adaptation does not involve an inadvertent or inappropriate conflation of social categories of race and ethnicity with genetic population groupings.

Articulation

Population. Require articulation of the rationale for the particular population grouping(s) being used. Require articulation of the relation between the actual sample being used and the population category in which it is being placed. Specifically require articulation of the nature or degree of representativeness being asserted for the sample in relation to the population category chosen.

For genetically defined population categories, require clarification of the justification for any concurrent use of nonbiological values, such as geopolitical nation-state boundaries or cultural groupings, to specify the location of descent populations. Nation-states may used to describe geographic regions of the world from which certain populations recently descended, but such correlations must be justified and refined to clarify discontinuities between the nation-state boundaries and relevant geographic regions.

Race/ethnicity. Require articulation of and justification for any relation asserted between any population-based genetic categories and any racial/ethnic categories. In particular, where appropriate, require articulation of whether race is being used as a risk factor or as a risk marker for a particular biomedical condition.

Rationale. The federal mandates create a powerful incentive for using racial and ethnic groupings to structure data and research or trial design. Once population groups and racial/ethnic groups are defined, it is important to require a clear articulation of how and why such categories are used in the trial or research project.

Particular problems may arise where a relatively small sample size comes to stand as a proxy for successively larger groups. Thus, for example, the Haplotype Map Project sample of 45 Han Chinese in Beijing (a geographically situated ethnic group) may come to stand for all Chinese people (a historical geopolitical group) and then for all Asians (a continental group). Indeed, this has already occurred: the International Haplotype Map Project Consortium itself has referred to these samples as simply being from a part of Asia.[30] This type of sequential expansion of correlation should be explicitly justified at each step.

Tight Fit

Correlation. Require a tight fit (a) between the population, racial/ethnic, and genetic categories being used and (b) between the genetic category identified and the disease state/health issue or other biological activity being analyzed.

The tightness of the fit may be assessed by considering whether the relation is based (a) on a significant difference (or identity) between the racial/ethnic and genetic categories used and (b) on a valid analysis that connects both the relevant genetic category and its racial/ethnic correlate to the identified disease state or other biomedical condition.

Where race or ethnicity is being used as a risk factor, require a tight fit between the aspect of race or ethnicity identified as a risk factor and causal aspects of the condition.

Where race is being used as a risk marker, require an explicit articulation of the nature of the correlation asserted between the marker and the identified condition. Require the specification that such a correlation does not speak to underlying causal aspects of the condition.

Rationale. One of the most imposing challenges in using racial and ethnic categories in biomedical contexts is preventing a sort of conceptual slippage that occurs through the elaboration of excessively attenuated relations between racial and ethnic categories and purported biological and genetic correlates. Harking back to the example of the 45 Han Chinese who come to stand for all of Asia, imagine further that this group of 45 is identified as having a particular frequency of a specific genetic marker that correlates with a higher likelihood of having a particular genetic variation, which in turn further correlates with a higher likelihood of contracting a particular disease at some unspecified time in the future. This disease, in turn, may have multiple causes and be manifested in various forms with differing degrees of severity. This attenuated correlation becomes even more problematic when one realizes that the initial OMB-defined category of race itself is not tightly bounded in a social context but involves the use of proxy markers and historically contingent conceptions of racial identity that have changed substantially over time.[37]

It should also be noted that when differential health outcomes are being studied, the fit between racial/ethnic categories and biology will tend naturally

to be very tight. For example, in health-disparities research on the biomedical impact of differential access to medical care among specified African American, Hispanic, Asian, or White populations, the fit between racial/ethnic categories and the biological health outcomes would be one of almost perfect identity.

Issues of fit will become more central in assessing projects that use race and ethnicity as proxies to uncover purported underlying genetic causes of disease.

Purpose

Social significance. Require a substantial health or scientific interest to be furthered by the use of racial or ethnic categorization in this context.

Rationale. The diverse federal mandates requiring the organization of data by race and ethnicity create an incentive to use the data thereby produced—whether or not they are directly relevant to the project at hand. Requiring the furtherance of a substantial health or scientific interest ensures that correlations between racial and ethnic categories and genetic categories will not be asserted post hoc with minimal justification. The standard of substantial interest is somewhat less rigorous than the compelling interest required under equal protection law. The rationale here is to recognize that biomedical research and clinical practice generally use racial and ethnic classifications for benign purposes.

Maintenance

Consistency. These requirements must be met for each use of racial and ethnic categories throughout the relevant project or practice.

Rationale. This is another deterrent to slippage. One common pitfall of existing approaches to using racial and ethnic categories in biomedical contexts is that researchers and clinicians may issue a sort of general disclaimer up front about race and ethnicity being social categories but then proceed through the rest of the project to treat them as, in effect, primarily biological or genetic.

Caveat

Exceptions. If a researcher is unable to meet these requirements because of an inability to disentangle what are perceived to be complexly intertwined social/genetic/biological variables or categories, the application and related documents may still be submitted to the relevant federal agency if the researcher provides an explanation and prominently incorporates the OMB caveat that "the racial and ethnic categories set forth in the standards [or application] should not be interpreted as being primarily biological or genetic in reference."[26]

Rationale. As a practical matter, individual researchers may find it difficult, given the design or nature of their projects, to break racial/ethnic and genetic population categories into their social, genetic, and nongenetic biological components. This is a major undertaking, but it is also necessary. These guidelines provide incentives to work out these issues, whereas the caveat allows researchers to proceed with their projects in a more deliberate manner while this difficult work progresses.

CONCLUSIONS

These recommendations are primarily procedural in nature. They preserve scientific autonomy and allow researchers and clinicians to define and act on their own conceptions of the relevance of the OMB categories of race and ethnicity to their own work. They would apply only to applications and other related documents submitted to the federal government.

Race and ethnicity are powerful categories. They have important roles to play in understanding a wide array of health-related phenomena. They must, however, be used with care. There are significant differences between using such categories to identify disparities in health outcomes and using them as proxies to try to identify underlying genetic causes of disease. I hope that these guidelines will promote more consistent and scientifically rigorous articulation, clarification, and application of these categories when applications and related documents are submitted to relevant federal agencies. ■

About the Author

Requests for reprints should be sent to Jonathan Kahn, Hamline University School of Law, 1536 Hewitt Ave., St. Paul, MN 55104 (e-mail: jkahn01@hamline.edu).

This article was accepted September 14, 2005.

Acknowledgments

This work was supported by National Human Genome Research Institute (grant R01 HG002818–01).

Thanks to the participants in the grant working group meetings and others for their helpful comments and suggestions: Donna Arnett, Guillermo Aviles-Mendoza, Rene Bowser, Rose Brewer, Ellen Clayton, Colin Campbell, Troy Duster, Phyllis Epps, Kim Fortun, Morris Foster, Jeffrey Kahn, Vivek Kapur, Richard King, Sandra Lee, Teri Manolio, Jonathan Marks, Michael Omi, Harry Orr, Susan Parry, Dorothy Roberts, Charmaine Royal, William Toscano, Rebecca Trotsky-Sirr, Karen-Sue Taussig, Samuel Wilson, Susan Wolf. Thanks also to the members of the NHGRI Human Genetics Variation Consortium for their comments on a related presentation.

Human Participant Protection

No human participants were involved in this study.

References

1. Food and Drug Administration, "FDA Approves BiDil Heart Failure Drug for Black Patients," *FDA News,* June 23, 2005. Available at: http://www.fda.gov/bbs/topics/NEWS/2005/NEW01190.html. Accessed July 5, 2005.

2. Jonathan Kahn, "How a Drug Becomes "Ethnic": Law, Commerce, and the Production of Racial Categories in Medicine, *Yale J Health Policy Law Ethics* 4 (2004): 1–46.

3. Jonathan Kahn, "Ethnic Drugs," *Hastings Cent Report* 35, Jan-Feb (2005): 1 p following 48. For a discussion of the trail design and data supporting the BiDil application, see Anne L. Taylor, et al., "Combination of Isosorbide Dinitrate and Hydralazine in Blacks with Heart Failure," *N Engl J Med* 351 (2004): 2049–2057.

4. Cohn J, and Carson P. 2002. Methods of treating and preventing congestive heart failure with hydralazine compounds and isosorbide dinitrate or isosorbide mononitrate. US Patent 6,465,463, issued October 15, 2002.

5. Institute of Medicine, *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care* (Washington, DC: National Academies Press, 2002).

6. Esteban González Burchard, et al., "The Importance of Race and Ethnic Background in Biomedical Research and Clinical Practice," *N Engl J Med* 348 (2003): 1170–1175.

7. Armand Leroi, "A Family Tree in Every Gene," *New York Times*, sec. A, March 15, 2005.

8. Pamela Sankar, et al., "Genetic Research and Health Disparities," *JAMA* 291 (2004): 2985–2989.

9. Sandra Soo-Jin Lee, Joanna L. Mountain, and Barbara Koenig, "The Meanings of 'Race' in the New Genomics," *Yale J Health Policy Law Ethics* 1 (2001): 33–75.

10. Troy Duster, "Enhanced: Race and Reification in Science," *Science* 307 (2005): 1050–1051.

11. See, e.g., Richard S. Cooper, Jay S. Kaufman, and Ryk Ward, "Race and Genomics," *N Engl J Med* 348 (2003): 1166–1170; Esteban González Burchard, et al., "Importance of Race and Ethnic Background." In response to a *New York Times* op-ed piece on genes and race, written by evolutionary biologist Armand Leroi in March 2005, the Social Science Research Council established a Web site dedicated to discussions of race and genetics. See Social Science Research Council, "Is race real?" Available at: http://raceandgenomics.ssrc.org. Accessed September 5, 2005.

12. Judith Kaplan and Trude Bennett, "Use of Race and Ethnicity in Biomedical Publication," *JAMA* 289 (2003): 2709–2715.

13. Elizabeth G. Phimister, "Medicine and the Racial Divide," *N Engl J Med* 348 (2003): 1081–1082.

14. Pamela Sankar and Mildred Cho, "Genetics: Toward a New Vocabulary of Human Genetic Variation," *Science* 298 (2002): 1337–1338.

15. Morris Foster and Richard Sharp, "Race, Ethnicity and Genomics: Social Classifications as Proxies of Biological Heterogeneity," *Genome Res* 12 (2002): 844–850.

16. Charles Rotimi, "Are Medical and Nonmedical Uses of Large-Scale Genomic Markers Conflating Genetics and 'Race'?" *Nat Genetics* 36 (2004): S43–S47.

17. Editorial, *Nat Genetics* 24 (2000): 97.

18. Editorial, *BMJ* 312 (1996): 1094.

19. Author Guidelines, *Paediatr Perinat Epidemiol.* Available at: http://www.blackwellpublishing.com/submit.asp?ref=0269–5022&site=1. Accessed December 8, 2004.

20. NIH Revitalization Act of 1993 (PL 103–43), codified at 42 U.S.C. § 289a-1 (1993).

21. Food and Drug Modernization Act of 1997 (P.L. 105–115), codified as amended at 21 U.S.C.§ 355 (2005)

22. National Institutes of Health, "NIH Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research." Available at: http://grants.nih.gov/grants/funding/women_min/women_min.htm. Accessed September 8, 2005.

23. Food and Drug Administration, "Guidance for Industry: Collection of Race and Ethnicity Data in Clinical Trials." Available at: http://www.fda.gov/cder/guidance/5054dft.pdf. Accessed December 9, 2004.

24. National Institutes of Health, "NIH Policy on Reporting Race and Ethnicity Data: Subjects in Clinical Research." Available at: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-01–053.html. Accessed September 8, 2005.

25. Food and Drug Administration, "Guidance for Industry," p 5

26. Office of Management and Budget, "Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity." Available at: http://www.whitehouse.gov/omb/fedreg/ombdir15.html. Accessed December 9, 2004.

27. National Institute of General Medical Sciences, Coriell Institute for Medical Research, "Human Variation Collections of the NIGMS Repository." Available at: http://locus.umdnj.edu/nigms/cells/humdiv.html. Accessed November 7, 2005.

28. National Center for Biotechnology Information, "Single Nucleotide Polymorphism: Search Population Class." Available at: http://www.ncbi.nlm.nih.gov/SNP/popclass.cgi. Accessed November 7, 2005.

29. National Human Genome Research Institute, "International HapMap Project." Available at: http://www.genome.gov/10001688. Accessed December 9, 2004.

30. International HapMap Consortium, "The International HapMap Project, *Nature* 426 (2003): 789–796.

31. National Human Genome Research Institute, "International Consortium Launches Genetic Variation Mapping Project." Available at: http://www.genome.gov/10005336. Accessed December 9, 2004.

32. International HapMap Consortium, "About the International HapMap Project." Available at: http://www.hapmap.org/abouthapmap.html. Accessed July 18, 2006.

33. Charles Sullivan and Erik Lilquist, "The Law and Genetics of Racial Profiling in Medicine," *Harv Civ Rights-Civil Lib Law Rev.* 39 (2004): 391–480.

34. John Robertson, "Constitutional Issues in the Use of Pharmacogenomic Variations Associated with Race," in *Pharmacogenomics: Social, Ethical and Clinical Dimensions*, ed. Mark Rothstein, 391–318 (Hoboken, NJ: John Wiley & Sons, 2003).

35. *Grutter v. Bollinger*, 539 US 306 (2003).

36. National Institutes of Health, "NIH Policy and Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research—Amended, October, 2001." Available at: http://grants.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm. Accessed March 3, 2005.

37. Melissa Nobles, *Shades of Citizenship: Race and the Census in Modern Politics* (Palo Alto, CA: Stanford University Press, 2000).