HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2023 April 27.

Published in final edited form as:

Science. 2022 April 15; 376(6590): 250–252. doi:10.1126/science.abm7530.

Getting genetic ancestry right for science and society

Anna C. F. Lewis[1,2], **Santiago J. Molina**[3], **Paul S Appelbaum**[4,5], **Bege Dauda**[6,7], **Anna Di Rienzo**[8], **Agustin Fuentes**[9], **Stephanie M. Fullerton**[10], **Nanibaa' A. Garrison**[11,12,13], **Nayanika Ghosh**[14], **Evelynn M. Hammonds**[14,15], **David S. Jones**[14,16], **Eimear E. Kenny**[17,18], **Peter Kraft**[19], **Sandra S.-J. Lee**[20], **Madelyn Mauro**[1], **John Novembre**[8], **Aaron Panofsky**[11,21,22], **Mashaal Sohail**[23], **Benjamin M. Neale**[24,25,26], **Danielle S. Allen**[1]

[1]Edmond J Safra Center for Ethics, Harvard University, Cambridge, MA, USA.

[2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

[3]Department of Sociology, Northwestern University, Evanston, IL, USA.

[4]Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

[5]New York State Psychiatric Institute, New York, NY, USA

[6]Center for Global Genomics and Health Equity, University of Pennsylvania, Philadelphia, PA, USA

[7]Institute of Clinical Bioethics, Saint Joseph's University, Philadelphia, PA, USA

[8]Department of Human Genetics, University of Chicago, Chicago, IL, USA

[9]Department of Anthropology, Princeton University, Princeton, NJ, USA.

[10]Department of Bioethics & Humanities, University of Washington School of Medicine, Seattle, WA, USA

[11]Institute for Society & Genetics, University of California, Los Angeles, CA, USA

[12]Institute for Precision Health, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

[13]Division of General Internal Medicine & Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

[14]Department of the History of Science, Harvard University, Cambridge, MA, USA

[15]Hutchins Center for African and African American Research, Harvard University, Cambridge, MA, USA

[16]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, USA

[17]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[18]Department of Medicine and Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[19]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[20]Division of Ethics, Department of Medical Humanities and Ethics, Columbia University, New York, NY, USA

[21]Department of Public Policy, University of California, Los Angeles, CA, USA

[22]Department of Sociology, University of California, Los Angeles, CA, USA

[23]Centro de Ciencias Genómicas (CCG), Universidad Nacional Autónoma de México (UNAM), Cuernavaca, Morelos, México

[24]Broad Institute of Harvard and MIT, Cambridge, MA, USA.

[25]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

[26]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.

Abstract

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

Glaring health disparities have reinvigorated debate about the relevance of race to health, including how race should and should not be used as a variable in research and biomedicine (1). Following a long history of race being treated as a biological variable, there is now broad agreement that racial classifications are a product of historically contingent social, economic, and political processes. Many institutions have thus been re-examining their use of race and racism, and stating intentions about how race should be used going forward. One common proposal is to use genetic concepts — in particular genetic ancestry and population categories — as a replacement for race (5). However, the use of ancestry categories has technical limitations, fails to adequately capture human genetic diversity and demographic history, and risks retaining one of the most problematic aspects of race—an essentialist link to biology—by allowing genetic ancestry categories to stand in its place.

The process of racialization entails a dynamic cognitive process of identification on the basis of phenotype that is often highly context dependent. While research has found genetic variation correlated with phenotypes that have been historically used to assign race categories, such as skin pigmentation or hair texture, it is the case that such genetic correlates are not distributed in a manner that correspond to racially-defined groups. Race is a socio-political construct rather than a biological one. For example, in the United States, immigrants from southern and eastern Europe only began to be classified as “white” on the census in the 20th century (2); the American Indian/Alaska Native census category reflects colonizing histories and federal policies (3). As such, social scientists and others have argued that the strongest case for using race is limited to tracking the impact of racism on health outcomes, rather than as a proxy for anything biological (4).

Genetic ancestry, one of the main proposed alternatives to using race, is of relevance to statistical and population geneticists, epidemiologists, public health practitioners, physicians and patients. In particular, it has renewed relevance for the clinical application of genetic technology because the accuracy of genetic risk scores varies across ancestries (6). Genetic

ancestry and population categories are also relevant to the general public, as demonstrated by the tens of millions of individuals who have paid for ancestry reports from consumer companies. Across these different domains, a dominant description of genetic ancestry is associated with continents as meaningful groupings. Within genetics research, continental ancestry categories have become the most common type of group label (7).

Similarly, consumer genetics products give customers a report with data based on a percentage of these continental groups from which an individual can trace their "ancestry." Systems of racial classification have historically regarded continents as meaningful group boundaries; thus it is not surprising that racial categories and continental ancestry categories are often confounded. Whenever continental ancestry categories are used, the risk is high that a misconception of race as a biological attribute will re-enter through the backdoor (8). Insufficiently nuanced thinking about continental categories, genetic ancestry, and racial groups can lead to the conflation of the three.

A FLATTENED NOTION OF ANCESTRY

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

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

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

The concern about use of categories goes beyond these technical limitations. Imposing categories on genetic ancestry fails to adequately capture human genetic diversity and what we know of human demographic history. A standard way to visualize patterns of genetic similarity is by plotting results of principal component analysis of genetic variation data, a

technique that reduces the dimensionality of that data. Most genetic analyses use data from reference populations to contextualize a study's data. The most commonly used reference data were created by sampling individuals from a few dozen places spread across the globe. If individuals from these populations are graphed in this manner, distinct clusters roughly representing continental categories are visible (see Figure 1). Indeed, a prominent early result was that genetic ancestry was remarkably concordant with continental origins when ascertaining for individuals whose four grandparents were from the recruitment sites (11). But newly assembled datasets show that if people are sampled differently, such as individuals living in New York City, it becomes clear how impoverished this view of a structure of distinct clusters is (see Figure 1) (12). The clearly separated clusters of reference population individuals, corresponding to different continental groups, merge into a background of continuous genetic variation. This is consistent with what we know of human demographic history, in which mass migration and constant mixing across groups have been the norm. The impact of these histories leads to different structures of genetic variation in different parts of the world. Such studies illustrate just how inappropriate use of discrete continental categories can be, particularly when information framed as genetic ancestry can potentially influence medical care.

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

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

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

Yet this historical notion of genetic ancestry is flattened when just one set of categories is used. In the case of continental ancestry categories, their use reflects the assumption that at some specific point in time humans were mostly divided into homogeneous groups by the natural geographical barriers between continents. This is a gross oversimplification of human history. It also obscures other time slices where different categories would be relevant, for example ~50,000 years ago Homo Sapiens and Neanderthal categories, or ~5,000 years ago “Steppe-related”, “European” hunter-gatherer and “Near Eastern” farmer categories in Europe (13), or ~500 years ago when waves of migration and the slave trade were forging new patterns of human genetic diversity in the Americas.

A MORE COMPLEX NOTION OF ANCESTRY

What are the implications for researchers who want to invoke genetic ancestry? They should first ask whether they need to impose categories at all to answer their research question. There are many situations where categorization has been thought essential but has subsequently been shown to be avoidable, for example in correcting for population stratification in genome wide association studies (14). In cases where genetic ancestry categories can be avoided, they should be avoided. If researchers are able to justify a scientific need to impose categories, they should next think about whether they have to provide labels (be it geographic, ethnic, linguistic, or other) to the groupings they impose. If they do need to provide labels, they should give the scientific justification for that choice and show that they have considered potential disadvantages of imposing these labels. Additionally, researchers should use multiple types of categories, reflecting the fact that genetic ancestry is a historical concept: we all have multiple ancestries depending on the time horizon considered. No individual has a single “ancestry”; the plural should always be used. Different geographical resolutions — for example, “Yoruban” versus “West African” — can serve as proxies for different time slices. Ancestry categories from different time points may be of medical relevance. The incorporation of ancient DNA information can also allow for probing different time slices, though the promise of this approach will depend on how much ancient DNA can actually be recovered and analyzed. The use of continental ancestry categories as a proxy for one of the time slices considered must be particularly carefully justified, because of the conflation of continental ancestry categories with racial groupings. Additionally, future work should find better ways to conceptualize the genetic ancestry of individuals whose recent ancestors come from distant parts of the ARG.

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

Science is reductive, and a model that uses simple continental categories has been useful in starting the process of understanding human genetic diversity. But all models have their

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

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

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

To aid this transition, a solid empirical understanding of how and why different fields use and operationalize the concept of ancestry is needed. To ensure this more complex notion of ancestry is then used in practice will require systems-level change. New computational tools and data structures will be required, for example a wider variety of proxies for genetic ancestry that do not impose categories, as well as easily accessible software tools to enable use of ancestry categories representing multiple time horizons. Further development and adoption of methodologies that directly estimate the ARG should be encouraged. Educational materials will need to be developed for scientists and physicians. Scientists of all stripes who engage in research that employs biological categories for humans should not work in isolation but as part of interdisciplinary teams, ideally including engagement with impacted communities. In support of these efforts journal editors should set standards, professional societies should publish best practices, and funders should carefully consider which research agendas they will support. It is paramount, as these organizations rightly critique the use of race as a biological variable, that use of continental ancestry categories does not become the new default. The U.S. National Academies of Sciences, Engineering

and Medicine recently formed an ad hoc committee, “Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research”; we are hopeful this represents an opportunity for consideration and consolidation of the points raised here.

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

Acknowledgements

Funding:

NIMH administrative supplement 5000747-5500001474 to 3R37MH107649-06S1. Author Contributions: ACFL: Writing – original draft; BMN and DA: Supervision; All authors: Conceptualization and Writing – review & editing. B.M.N and D.S.A. contributed equally to this work. Competing interests: ACFL owns stock in Fabric Genomics; EK has received personal fees from Regeneron Pharmaceuticals, 23&Me, and Illumina, and serves on the advisory boards for Encompass Biosciences and Galateo Bio; BMN is a member of the scientific advisory board at Deep Genomics and RBNC Therapeutics, Member of the scientific advisory committee at Milken and a consultant for Camp4 Therapeutics and Merck.

REFERENCES AND NOTES

1. Vyas DA, Eisenstein LG, Jones DS, Engl N. J. Med. 383, 874–882 (2020).
2. Roediger DR, Working Toward Whiteness: How America’s Immigrants Became White: The Strange Journey from Ellis Island to the Suburbs (Basic Books, New York, Text is Free of Markings edition., 2005).
3. Haozous EA, Strickland CJ, Palacios JF, Solomon TGA, Environ J. Public Health. 2014, e321604 (2014).
4. Fujimura JH, Duster T, Rajagopalan R, Soc. Stud. Sci. 38, 643–656 (2008). [PubMed: 19227816]
5. Oni-Orisan A, Mavura Y, Banda Y, Thornton TA, Sebro R, Engl N. J. Med., in press, doi:10.1056/NEJMms2031080.
6. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ, Nat. Genet. 51, 584 (2019). [PubMed: 30926966]
7. Panofsky A, Bliss C, Am. Sociol. Rev. 82, 59–87 (2017).
8. Duster T, Backdoor to Eugenics (Routledge, New York, 2nd edition., 2003).
9. Mathieson I, Scally A, What is ancestry? PLOS Genet. 16, e1008624 (2020). [PubMed: 32150538]
10. Lawson DJ, van Dorp L, Falush D, Nat. Commun. 9, 3258 (2018). [PubMed: 30108219]
11. Rosenberg NA et al., Science 298, 2381–2385 (2002). [PubMed: 12493913]
12. Belbin GM et al., Cell. 184, 2068–2083.e11 (2021). [PubMed: 33861964]
13. Olalde I et al., Nature 555, 190–196 (2018). [PubMed: 29466337]
14. Wojcik GL et al., Nature 570, 514–518 (2019). [PubMed: 31217584]
15. Sung W-C, in Asian Biotech: Ethics and Communities of Fate, A. Chen Ong, N., Eds. (Duke University Press, 2010; https://read.dukeupress.edu/books/book/1481/chapter/170875/Chinese-dnaGenomics-and-Bionation), pp. 263–288.

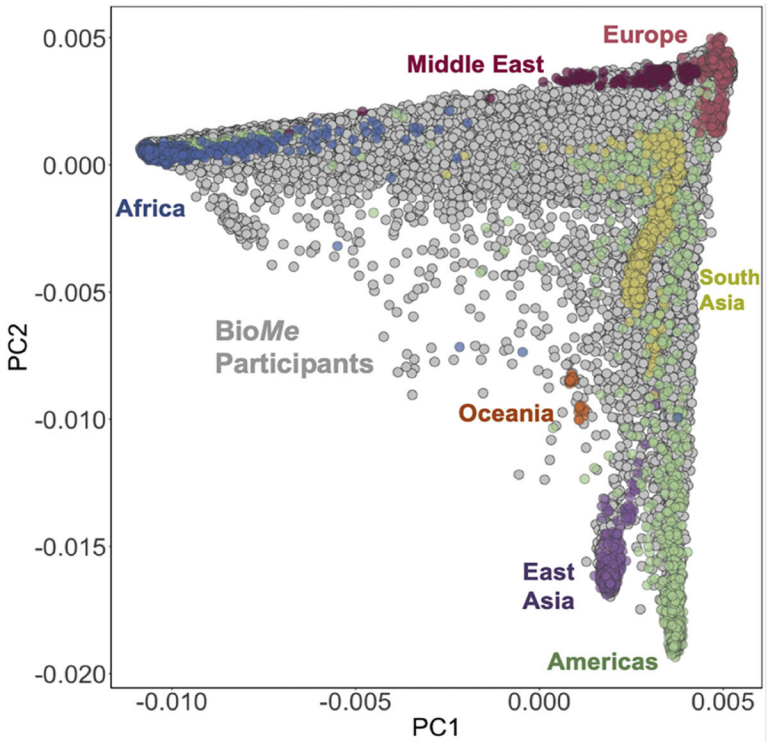

Figure 1. The continuous, category free, nature of genetic variation.

Colored dots (N=4149) are reference panel individuals from 87 populations representing ancestry from 7 continental or subcontinental regions projected onto the first two principal components of genetic similarity. Gray dots (N=31705) are participants from BioMe, a diverse biobank based in New York City. Clearly delineated continental ancestry categories, the islands of color, are shown to be a by-product of sampling strategy. They are not reflective of the diversity in this real-world dataset, made evident by the continuous sea of gray. Reproduced/modified from (12).