frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 23 January 2023
DOI 10.3389/fgene.2023.1044555

Check for updates

OPEN ACCESS

EDITED BY
Manuel Corpas,
Cambridge Precision Medicine,
United Kingdom

REVIEWED BY
I. King Jordan,
Georgia Institute of Technology,
United States
Amy L. Non,
University of California, San Diego,
United States

*CORRESPONDENCE
Anna C. F. Lewis,
✉ annalewis@fas.harvard.edu

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics,
a section of the journal
Frontiers in Genetics

RECEIVED 14 September 2022
ACCEPTED 10 January 2023
PUBLISHED 23 January 2023

CITATION
Dauda B, Molina SJ, Allen DS, Fuentes A, Ghosh N, Mauro M, Neale BM, Panofsky A, Sohail M, Zhang SR and Lewis ACF (2023), Ancestry: How researchers use it and what they mean by it.
Front. Genet. 14:1044555.
doi: 10.3389/fgene.2023.1044555

Ancestry: How researchers use it and what they mean by it

Bege Dauda[1], Santiago J. Molina[2], Danielle S. Allen[3], Agustin Fuentes[4], Nayanika Ghosh[5], Madelyn Mauro[3], Benjamin M. Neale[6,7,8], Aaron Panofsky[9,10,11], Mashaal Sohail[12], Sarah R. Zhang[13] and Anna C. F. Lewis[3,14]*

[1]Center for Global Genomics and Health Equity, University of Pennsylvania, Philadelphia, PA, United States, [2]Department of Sociology, Northwestern University, Evanston, IL, United States, [3]Edmond & Lily Safra Center for Ethics, Harvard University, Cambridge, MA, United States, [4]Department of Anthropology, Princeton University, Princeton, NJ, United States, [5]Department of the History of Science, Harvard University, Cambridge, MA, United States, [6]Broad Institute of Harvard and MIT, Cambridge, MA, United States, [7]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, United States, [8]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, United States, [9]Institute for Society & Genetics, University of California, Los Angeles, Los Angeles, CA, United States, [10]Department of Public Policy, University of California, Los Angeles, Los Angeles, CA, United States, [11]Department of Sociology, University of California, Los Angeles, Los Angeles, CA, United States, [12]Centro de Ciencias Genomicas (CCG), Universidad Nacional Autonoma de Mexico (UNAM), Cuernavaca, Morelos, Mexico, [13]University of California, Berkeley, Berkeley, CA, United States, [14]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States

Background: Ancestry is often viewed as a more objective and less objectionable population descriptor than race or ethnicity. Perhaps reflecting this, usage of the term "ancestry" is rapidly growing in genetics research, with ancestry groups referenced in many situations. The appropriate usage of population descriptors in genetics research is an ongoing source of debate. Sound normative guidance should rest on an empirical understanding of current usage; in the case of ancestry, questions about how researchers use the concept, and what they mean by it, remain unanswered.

Methods: Systematic literature analysis of 205 articles at least tangentially related to human health from diverse disciplines that use the concept of ancestry, and semi-structured interviews with 44 lead authors of some of those articles.

Results: Ancestry is relied on to structure research questions and key methodological approaches. Yet researchers struggle to define it, and/or offer diverse definitions. For some ancestry is a genetic concept, but for many—including geneticists—ancestry is only tangentially related to genetics. For some interviewees, ancestry is explicitly equated to ethnicity; for others it is explicitly distanced from it. Ancestry is operationalized using multiple data types (including genetic variation and self-reported identities), though for a large fraction of articles (26%) it is impossible to tell which data types were used. Across the literature and interviews there is no consistent understanding of how ancestry relates to genetic concepts (including genetic ancestry and population structure), nor how these genetic concepts relate to each other. Beyond this conceptual confusion, practices related to summarizing patterns of genetic variation often rest on uninterrogated conventions. Continental labels are by far the most common type of label applied to ancestry groups. We observed many instances of slippage between reference to ancestry groups and racial groups.

Conclusion: Ancestry is in practice a highly ambiguous concept, and far from an objective counterpart to race or ethnicity. It is not uniquely a "biological" construct, and it does not represent a "safe haven" for researchers seeking to avoid evoking race or ethnicity in their work. Distinguishing genetic ancestry from ancestry more broadly will be a necessary part of providing conceptual clarity.

KEYWORDS

population descriptors, ancestry, genetic ancestry, race, ethnicity, population labeling

Introduction

The use of population descriptors is currently under the spotlight, both in genetics research specifically and across biomedical research more broadly (National Academies 2021; Vyas, Eisenstein, and Jones 2021; Khan et al., 2022). Social scientists have studied how race and ethnicity are used, but have paid much less attention to ancestry. Understanding how researchers conceptualize and use ancestry matters—both to genetics and to biomedicine more broadly—for several reasons. First, it is a key concept drawn upon in decisions about who we study and why (Popejoy and Fullerton 2016; Bentley, Callier, and Rotimi 2017). Second, because it plays a key role in making sure methodologies yield robust and replicable results (Martin et al., 2019; Peterson et al., 2019). Third, because further reliance on genetic ancestry is part of the proposed solution to the use of race in biomedicine (Borrell et al., 2021; Oni-Orisan et al., 2021). Fourth, decisions made in research directly impact translational work, and ultimately medical practice (Popejoy et al., 2018). And finally, understanding this concept is important because it is a key frame offered for understanding biological differences between groups of humans—including those that could be driving race-based health disparities (Batai, Hooker, and Kittles 2021). This is an ethically fraught topic with a long and unpleasant history (Reardon 2005; Roberts 2011; Bliss 2020a). The stakes are hence high to ensure that concepts originating in genetics do not result in repetition of past atrocities stemming from the categorization of humans into a small number of biological types (Mathieson and Scally 2020; Lewis et al., 2022).

A better understanding of how researchers use ancestry can also help provide raw material for normative recommendations about how the concept should be used. The National Academies of Science, Engineering and Medicine (NASEM) are currently convening a taskforce on the appropriate usage of race, ethnicity and ancestry in genetics research (National Academies 2021). The appropriate usage of race, ethnicity, and ancestry is not a new topic: a scoping review found over 100 articles offering relevant normative guidance published since 2000 (Mauro et al., 2022). A consistent theme in these recommendations has been the need for transparency by researchers, including why they are using population categories, and how any population categories used are defined (Mauro et al., 2022). The use of race and ethnicity have been the main focus of these normative debates to date, with ancestry receiving relatively little scrutiny. For example, guidance from the American Medical Association has solely focused on race and ethnicity (Flanagin et al., 2021). It has been explicitly suggested that ancestry is the least controversial of the population descriptors (Lee, Mountain, and Koenig 2001), and this assumption seems to drive much of the move away from race and ethnicity categories. It is also seen as the most objective classifier. Pointing out how race and ethnicity categories are broad, imprecise, and ambiguous, Borrell et al. write, “In contrast, ancestry is a fixed characteristic of the genome” (Borrell et al., 2021). Recent content analysis of research articles published in the *American Journal of Human Genetics* has shown that the term “ancestry” is increasingly used in genetics research (Byeon et al., 2021). As Wagner et al. have argued, the increasing focus on ancestry as a way to “frame human difference” should be a motivating factor to bring more attention to its use (Wagner et al., 2017).

There have been empirical insights into researchers’ use of ancestry. Articles using ancestry (compared to race or ethnicity) were the least likely to provide a rationale for its use (Ali-Khan et al., 2011). Interviews with health researchers have demonstrated the degree of confusion amongst researchers about the interrelationships between race and ethnicity and genetic differences between populations (Baer et al., 2013). Drawing on content analysis of articles published in *Nature Genetics* and interviews, Panofsky and Bliss explore geneticists’ use of population labels, demonstrating the increasing use of continental labels, which they argue are fundamentally ambiguous because they “blur racial and geographic understandings of population difference” (Panofsky and Bliss 2017). In an ethnography of geneticists’ use of principal components analysis (PCA) to capture genetic ancestry, Fujimura and Rajagopalan argue that while there are opportunities to update how the field thinks about human biological difference, race and ethnicity nonetheless enter into the concept of ancestry (Fujimura and Rajagopalan 2011). Focusing on biomedical articles using the terms “black”, “African” and “African American” Duello et al. find that most studies do not give a rationale for their focus on these populations, and conclude “we infer the authors of these studies believe African ancestry denotes a biological ‘race’ of people of common descent who share DNA unique from the rest of mankind” (Duello et al., 2021).

Researchers across multiple fields employ the concept of ancestry; it is not straightforwardly a concept that is “owned” by genetics. The concept of ancestry is often employed in everyday conversation, carrying sociocultural implications outside of its usage in genetics and health research. This use across contexts gives many opportunities for miscommunication about what ancestry is and is not, and what we can learn about or from it. Genetics has a special role in our understanding of ancestry, but, as mentioned above, this concept then diffuses out to translational research, to the practice of medicine, and to popular conceptions about the human family tree.

In this study, we employ a mixed methodology—a systematic literature analysis and semi-structured interviews—to offer a comprehensive examination of how ancestry is used by researchers. Because diverse domains use the concept of ancestry, and because they are of mutual relevance to each other, we include research from multiple disciplines. We seek to answer five questions about researchers’ use of ancestry. First, what types of research use ancestry? Second, when and why does ancestry enter the research process, i.e., what are the use cases for the concept? Third, what does ancestry mean to researchers, i.e., what definitions do researchers offer for the concept? Fourth, how is ancestry operationalized, i.e., how is this abstract concept made into a measurable observation? And finally, what types of population labels are used for ancestry categories? Answers to the first two questions help indicate just how important the concept of ancestry is in structuring both research questions and methodologies; answers to the remaining questions shed light on whether ancestry as currently conceptualized and operationalized can bear this heavy weight.

Materials and methods

Study design

This study employed two different methodologies to understand the use of ancestry by researchers: a systematic literature analysis and semi-structured interviews. The systematic literature analysis was performed on an original dataset composed of a sample of research articles in the population sciences. We designed this corpus of original research articles to capture as much of the diversity of the ways that ancestry is currently used by researchers as possible, in terms of divergent research questions and methodologies across disciplines. Given the motivation of our work to inform the use of ancestry and genetic ancestry across the biomedical sciences, we constrained the articles in our corpus to have at least some tangential relevance to human health. The study also aimed at diversity in terms of publication journals such that not only articles published in high impact factor journals were selected for the study, as this might represent a selection bias towards particular types of studies (e.g., Large N). We also identified a subset of these articles as engaging with the concept of ancestry particularly closely, and invited the first and last authors of these articles to participate in a semi-structured interview. This multi-method approach was designed to allow us to develop a robust understanding of researchers' use of ancestry, with the systematic literature analysis revealing patterns of usage, and the interviews allowing us to understand why we observed these patterns. This strategy yielded 205 articles and 44 interviewees.

Article inclusion strategy

All searches were based on Web Of Science (WOS) and conducted in February 2021. We restricted all searches to articles that contained "ancestry" in the title, abstract, or keywords, and then deployed two search strategies. First, we restricted articles to those concerning certain phenotypes, published from 2019 on. This focus on phenotypes ensured that we obtained diversity along other dimensions that we cared about, specifically research methodology. We chose a range of different types of conditions, all of which (like most health conditions) present known health disparities: COVID-19, prostate cancer, chronic kidney disease, and schizophrenia. These articles were filtered to just retain original research articles where the phenotype of interest was a central explanandum of the article. Second, additional searches were conducted in order to obtain a sufficient number of anthropology, social science, and public health articles. The start year was adjusted to ensure an adequate sample size; this meant starting in 2010 for sociology (extending back to 2010 was necessary to achieve sufficient articles to analyze), 2015 for anthropology, and 2020 for public health. For the anthropology and sociology searches, we also required that the term "health" appear in the title, abstract, or keywords. The results from all three searches were then filtered to retain articles that had either a connection to "health" broadly defined, or to human evolution, or to the characterization of human populations. The number of articles from each of the searches before and after filtering are given in Supplementary Table S1. The search strings and details of filtering are given in the Supplementary Material. The complete list of included articles is included as a Supplementary Material.

TABLE 1 Field and Country/Region of first author's primary affiliation of the articles in our corpus. The countries within the "Others" category are Australia, Canada, New Zealand, Jamaica and Suriname. The US was not aggregated into a region because of the country's dominance in the discourse around population descriptors in research, likely due to the historic antecedents on the use of race in the country.

		Number of articles (*N* = 205)
Field	Anthropology	20
	Biology	54
	Medicine	65
	Public Health	55
	Sociology	11
Country/Region	Africa	9
	Asia	20
	Europe	35
	Latin America	15
	U.S.	101
	Others	25

The search results were exported, and the PDFs of the articles and their Supplementary Information were downloaded. We used the abstract, journal, and affiliations of first and last authors to assign a primary subfield and field to each article (fields were anthropology, biology, medicine, public health, and sociology, see Supplementary Material for methodology details). The number of articles per field is given in Table 1, and per subfield in Supplementary Table S2. We also assigned a Country/Region to an article based on the Country/Region of the first author's primary affiliation, see Table 1.

Interviewee recruitment and interview guide

Based on reading the articles, we identified 97 that engaged with the concept of ancestry most closely. This was typically because they either used ancestry to frame or motivate their research question, because ancestry was evoked centrally in their methodology, or the term "ancestry" frequently occurred in the text. In order to ensure we heard from researchers at multiple career stages, we invited the first and last authors of these articles to participate. In a small handful of cases, interviewees recommended we contact a middle author or other close collaborator to interview. This yielded 190 names, 166 of which we found emails for. Of these 166, we interviewed 44 (27% response rate). The majority (29) were based in the United States, with 7 based in Europe, 6 in Central and South America, one in Canada, and one in India. The interviewees were assigned a subfield based on their training as inferred through their professional biographies, publication record, and in conversation during the interview. We achieved disciplinary diversity within our interviewees: 7 from anthropology, 14 from biology, 11 from medicine, 6 from public health, and 6 from sociology.

In the interviews, we utilized the article by virtue of which the participant was recruited to probe in greater depth how they use and think about ancestry. The semi-structured interview guide covered five areas: their background and the focus of their work; the justifications for their choices related to ancestry, and how they understood the limitations of these; conceptual questions, including "What does ancestry mean to you?"; publishing and the mechanisms of funding; and their views on the status of the field.

Interviews were 1 hour long, and conducted by two interviewers, one with a biology background (AL) and one sociologist (SM). The interviews were conducted on a video conferencing platform, recorded, and auto-transcribed. These transcripts were then updated based on the recordings. Identifying information was removed. The interview study was deemed exempt by the Harvard University-Area Committee on the Use of Human Subjects (protocol ID IRB21-0496).

Data analysis

The documents—the full text of the 205 articles and the 44 anonymized transcripts—were first coded. This involves identifying, highlighting, and annotating the appropriate sections of text corresponding to a set of codes. Around 3,000 sections of text from the articles were tagged with 27 codes, and about 2,300 to 34 codes from the interview transcripts. The list of codes covered are given in Supplementary Tables S3, S4, and further details of the development of these lists of codes and details of the coding process are given in the Supplementary Material.

On a code-by-code basis we then analyzed the sections of coded text for emergent themes. In addition to this qualitative analysis, we established two features of the articles which enable quantitative presentation of results. First, for those articles that operationalize ancestry, we coded the data type(s) they use to do so. Second, we categorized the type of population labels used in the articles. We used the types of population labels as previously analyzed in Panofsky and Bliss (Panofsky and Bliss 2017), with some minor adaptations: see Supplementary Table S5 for these types, with examples. We stress that our corpus of articles is not representative of any clearly defined set of literature, and hence that our results do not generalize to all population sciences. Instead this curated dataset is enriched to identify salient variables that shape researchers' conceptions of ancestry and to describe a wide variety of uses of the concept in scientific work.

Results

What types of research use ancestry?

The 205 articles in our corpus had diverse research aims, which we group into five categories, giving examples.

The most commonly represented category was articles aiming to understand traits and outcomes. This includes identifying genetic variation linked to traits (often but not always *via* Genome Wide Association Studies (GWAS) (e.g., (Legge et al., 2019; Lin et al., 2019; Du et al., 2020))), identifying causal influences on a trait using Mendelian Randomization [e.g., (Jordan et al., 2019; Howe et al., 2020)], identifying the interplay between genetic and environmental factors [e.g., (X. Chen et al., 2019; Ding et al., 2020)], understanding the impact of structural determinants of health [e.g., (Thayer et al., 2017; Whaley 2020)], understanding the molecular mechanisms that contribute to a trait/health outcome [e.g., (Emami et al., 2019; Gohlke et al., 2019)], and controlling for genetics in understanding social traits (Boardman et al., 2010).

A second set of articles, and the second most represented category, aimed at understanding between-group differences in traits/outcomes. Some of these articles compare traits/outcomes between those of different population categories [e.g., (Weitz, Garruto, and Chin 2016; Wong et al., 2019)] or those with different percentages of a particular ancestry category [e.g., (Grizzle et al., 2019; Fritz et al., 2020)]. Some articles compare trait-associated genetic variation between ancestry groups [e.g., (Koga et al., 2020; Liu et al., 2020)]. Some of these articles explicitly couch their efforts in terms of understanding health disparities [e.g., (Boulter et al., 2015; Marden et al., 2016)].

A third set of articles focuses on understanding genetic structure. These articles aimed to describe and infer population history, to understand evolutionary processes [e.g., (Macholdt et al., 2015)], and gain insight into how present-day genetic diversity is shaped [e.g., (Leishangthem et al., 2020; Zhao et al., 2020)].

A fourth set of articles aimed to understand social identities. This includes understanding how ancestry and genetic ancestry relate to categorical frameworks used in the present such as race [e.g., (Liebler 2016; Paredes 2017)], and understanding what factors influence these social identities (Hunley et al., 2017). It also involves trying to gain insight into the life histories and lived experiences of those who lived in the past, particularly enslaved individuals [e.g., (Wasterlain, Costa, and Ferreira 2018; Fleskes et al., 2021)].

A final set of articles evoking ancestry were aimed at directly improving the provision of healthcare. This includes: improving patient/participant engagement for example by understanding the views of those of diverse ancestries (Menzies et al., 2020; Saad et al., 2020); developing clinical tools for example by consideration of incorporation of genetic ancestry as a variable [e.g., (Haas Pizarro et al., 2020; Canter et al., 2019)]; studying the impact of genetic testing, for example by reporting the diagnosis rate by ancestry group (Groopman, 2019); and enabling quality control, for example by comparing genetically inferred ancestry to self-reported data for cell lines (Hooker et al., 2019).

Some articles conducted research that spanned these categories. For example, many of the articles aimed at understanding traits (e.g., a GWAS) additionally include a between-group comparison (e.g., comparing the frequency of an identified variant across ancestry groups). We observed that many articles do not clearly lay out their aims, with the relationship between research question and research motivation somewhat diffuse.

How does ancestry enter the research process?

Ancestry can enter the research process at multiple stages, representing different use cases for the concept.

As seen in the previous section, ancestry can be used to frame the research question when the focus is on the relevance of ancestry to a trait or outcome. The motivation for this is the assumption that ancestry reflects distinctive patterns of genetic difference, and that using it as a key variable will help identify whether genetic factors could be contributing to differences in incidence and prevalence rates of disease, or differences in traits, between groups [e.g., (Yuan et al., 2020)]. This can be couched explicitly in terms of understanding whether genetics plays a role in health disparities [e.g., (Apprey et al., 2019)].

Ancestry is also used to state the research question when it is seen as defining the population of interest. Two justifications for choice of population of interest were particularly common in our data. First, the lack of pre-existing research in that population, either noting the under-representation of those of certain ancestries in research generally, or the absence of a particular type of study in a particular population. To further strengthen the justification for using the named ancestral group in the study, some articles point to the consequences of not doing this work, for example to "*exacerbate existing health disparities*" (Harlemon et al., 2020). The second common justification given was the high prevalence of a phenotypic trait in that ancestry group/population.

Ancestry can also enter at the analysis stage. A notable example is to account for confounding in a GWAS, either in identifying suitable controls in a case-control study, or to account for population structure using principal components. While most use "ancestry" language explicitly to describe this process, some do not [e.g., (Ohi et al., 2020)], or circumscribe this use more carefully as "heterogeneity that is correlated with ancestry" (Franceschini and Morris 2020). Another example where ancestry is treated as something to be "controlled for" is in controlling for the influence of genetics when understanding a trait (Boardman et al., 2010), to help better understand the influence of other variables on that trait. Another example is in admixture mapping, which explicitly models aspects of genetic structure, using a process often referred to as local ancestry inference, in order to decrease bias in the estimates of the effect size of genotype-phenotype correlations [e.g., (Du et al., 2020)].

Finally, ancestry is evoked at the reporting stage of the research process, both in the statement of results and of the conclusions that follow from them [e.g., (Yoshikawa et al., 2020; Darst et al., 2020)]. This is true even when ancestry does not explicitly enter the research aim, for example in statements of how the results may or may not generalize to different populations (Tonon et al., 2019; Brhane et al., 2020).

What does ancestry mean to researchers?

We've seen that ancestry is used to frame research questions and as a core part of research methodologies. What do researchers mean by it? Many researchers—all of whom were selected to interview because their work closely engaged with the concept of ancestry—struggled to answer the question "What does ancestry mean to you?". While the majority did offer definitions, often after pauses, some were not able to: "Yeah it is hard, I do not know in fact." We observed that interviewees who were the most engaged in the concept were the most uncertain about it, and/or who gave the most expansive, multipart definitions. An example of the former is an interviewee who remarked that defining ancestry was "like trying to catch smoke". An example of the latter was an interviewee who emphasized that the more one thinks about the concept, the more expansive the conceptualization becomes: "it is not a simple (question), to answer, because there are of course different ways of thinking about it... each project, to some extent, helps you to rethink what (ancestry) means". Consistent with ancestry being a concept from everyday life, many interviewees drew from their own personal stories in the answers that they gave.

We identified two dimensions along which answers differed. The first is what ancestry is a property of: DNA; the individual; their family/kin; a population. The second is the criteria by which ancestry is shared: geographical origin; genealogical connections; culture; biology. Examples are given in Table 2. The qualitative methodology we employed is not suitable to quantify the answers we received, but we note that the answers we received were well distributed along both of these dimensions. Of note, geneticists did not all just give definitions in terms of genetic information. The "culture" category encompasses shared ethnicity, but also more specifically, shared narrative. For example, one interviewee described ancestry as "intergenerational transmission of who I am, what my family is, what we do, who we are."

Many individuals gave multipart answers spanning different cells in the matrix represented by Table 2. For example, "there's two aspects to it, one sort of the social kind of ancestry that people know about, that they're talking about, and the other one is then what genetics actually shows." Or, "Mostly genetic heritage, I mean you know literal inheritance... and something more like culture or religion or ethnicity or identity or family or community or group".

Some interviewees were keen to emphasize the distinction of ancestry from race and/or ethnicity. For example, "race is a self identified construct and ancestry is biology", or "it is very far from concepts like, race or ethnicity or something like that." But others directly related the concepts, for example "African American ancestry - that racial group that, you know, would essentially have common genotypic and phenotypic characteristics." In some cases there was a direct conflation which was masked by language, "Previously I used ethnicity. But then my mentor told me that nowadays people use ancestry.", or as a term to be used "instead of race, because we cannot use, like 'mixed race' because... the connotation is not good."

How is ancestry operationalized?

Before it is used, ancestry first has to be operationalized, i.e. a process has to be defined to ascribe an ancestry to individuals.

Out of the 205 articles examined, 56 (27%) did not operationalize ancestry in their methodology or analysis. Many articles use the word "ancestry" haphazardly and interchange the term with other population descriptors. For example, an article states "More than half of our patients are from African ancestry" (Arleo et al., 2021), but uses race groups throughout. In these cases, the authors seem to be grasping for a term that is suitably inclusive to cover a range of type of variation, or to use a word viewed as unobjectionable (in comparison to race or ethnicity), or to use a term that sounds more objective. Some use the term specifically to draw attention to the lack of data outside of European ancestry populations (Guan et al., 2020). The term "ancestry" appears as a label for ethnic groups in two main cases: "indigenous ancestry" (Marziali et al., 2021) and "mixed ancestry" (Hill et al., 2020). Finally, some articles only contained mention of "ancestry" in the keywords (and not in the main body of the text), using the "ancestry group" MESH terms (Dina 2022).

TABLE 2 Categorizing the definitions of ancestry offered by researchers whose work closely engages the concept by a) what ancestry is a property of, and b) the criteria by which individuals share ancestry.

	Geographical origin	Genealogical connections	Culture	Biology
DNA	"The geographic origin of genetic variation"	"The genetic information that we inherit from our ancestors"		
Individual	"Direct roots... those who may have been born, lived" (in a particular place)		"What people themselves say when we collect samples, and we asked them. . . what ethno-linguistic group they belong to"	"Ancestry is related to your biological background. . . your biological roots"
Family/ Kin	"Where your more recent ancestors came from geographically"	"We know it as genealogy"	"Full family tree.... not only what country but what's their role in the country, and how does that relate to my identity"	
Population	"The history of the distribution of populations."	"It's just a bunch of people more sharing a common ancestor than another group of people."		"Population differences in genotype that occurred over long periods of time, based on human migration"

TABLE 3 Types of data used to operationalize ancestry in articles

Type of data used to operationalize ancestry		N (%)	N (%)
Genetic			57 (52)
	Just genetic	53 (48)	
	Multiple operationalizations: genetic and not specified	4 (4)	
Non-genetic			38 (35)
	Dental morphology	1 (1)	
	Geographic	9 (8)	
	Language	1 (1)	
	Self report	20 (18)	
	Surname	4 (4)	
	Intersect of non-genetic data types	3 (3)	
Both genetic and non-genetic			15 (14)
	Intersect of genetic and self report	11 (10)	
	Multiple operationalizations: genetic and non-genetic	4 (4)	
TOTAL		110 (100)	110 (100)

Of the remaining 149 articles that did operationalize ancestry, for 39 (26%) of these we could not determine what type of information was used in this operationalization (e.g., self report, geographic, genetic). For example, when articles simply refer to "men of African ancestry" (Lam et al., 2019; Walavalkar et al., 2020). In some of these cases, it seems that the choice of population descriptor was not carefully made. For example, one article describes its sample as racially diverse, but demonstrates this using *European ancestry*, *African ancestry* and *other* categories (Gur et al., 2019). In another paper, the authors refer to *European*, *African* and *Mixed ancestries* in some places, but in others use *Mixed ancestry*, *Black African*, and *Caucasian* (Passchier et al., 2020). Some researchers use ancestry and ethnicity synonymously (Franceschini and Morris 2020).

Of the 110 articles where it was possible to tell what type of data was used to operationalize ancestry, 72 (65%) used genetic information to do so, 53 (49%) used non-genetic data, and 15 (14%) used both genetic and non-genetic data, see Table 3. Of the 38 articles that used exclusively non-genetic data to operationalize ancestry, 14 were genetics papers. The non-genetic data types used to operationalize ancestry were: self-reported information, geographical information, language spoken, surnames, dental morphology, and sometimes the intersection of more than one of these sources. Eleven articles (10%) used an intersection of genetic and non-genetic data types. Eight articles (8%) operationalized ancestry in more than one way in the same paper.

For those articles for which it was possible to determine the type of information used to operationalize ancestry, this information was often hard to find. For example, in Harrison et al., the main text referred to their Supplementary Material, which in turn pointed to a preprint. In the Supplementary Material of the preprint the details of how ancestry was operationalized are given (Harrison et al., 2020).

Most of the articles that used self-identified information to operationalize ancestry are consistent with participants being asked to report their race and/or ethnicity rather than their ancestry. Suggestive evidence of this includes interchangeable use of ethnicity and ancestry, reference to "racial ancestry" (Gendy et al., 2019; Kaur et al., 2019), and demonstration of "racial diversity" using ancestry categories (Dupont et al., 2020). A notable source of data where individuals are actually asked to self-report their ancestry is the American Community Survey (ACS), administered by the US Census Bureau. It currently asks, "What is your ancestry or ethnic origin? (For example: Italian, Jamaican, African American, Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, and so on.)". The ACS was used by six articles in our corpus: four from public health and two from sociology.

There was also diversity in how geographic data was used to operationalize ancestry: birth country of parents (Hamilton et al., 2018; Yamasato et al., 2020); birth country of paternal grandfather (Groeger et al., 2017); all of mother, father, and four grandparents born in a particular geographical area (Martínez-Magaña et al., 2019); or, no clarity beyond the mention of a country (Chen et al., 2019).

Ancestry was also operationalized using surnames, including: inferred place of origin of the surname of the individual (Bakhtiari 2020); whether an individual's two surnames (one inherited from each parent) in Mexico were deemed Mayan (Azcorra et al., 2016); the fraction of the surnames of each of an individual's two parents deemed Andean in Peru (Pomeroy et al., 2015).

The articles in our corpus that used genetic data to operationalize ancestry used a variety of methodologies to do so. Some used analysis of Ancestry Informative Markers (AIMs), which are genetic variants specifically picked because they have very large frequency differences in different groups, where the groups of interest are chosen during design of the AIMs. The majority of articles used standard genotype chips and one of a small handful of methodologies. One common method is Principal Component Analysis (PCA), a dimension reduction technique that defines a space in which the first dimension (PC) captures the most variance in the inputted data, the second the next most, etc. Another common method is a version of the STRUCTURE/ADMIXTURE algorithms, which assign each individual a percentage of ancestry in different variation distribution clusters referred to as populations. These algorithms can be run in an unsupervised fashion, where the user provides only the number of populations (referred to as the "k" parameter). It can also be run in a supervised fashion, where the user defines the populations by providing reference genetic data for each. Of course, the labels for these populations are previously defined using some other form of data, typically geographic or self-identified. Methodologies other than these two were sometimes used, notably Multidimensional Scaling e.g., (Maciukiewicz et al., 2019).

As previously mentioned, many articles use PCA in their analysis; most of these refer to an individual's location in Principal Component space as their ancestry, though exceptions include referring to this as ethnicity (Yoshikawa et al., 2020). In interviews, amongst those closest to population genetics, the relationship between PCs, population structure, and genetic ancestry was couched in diverse ways: ancestry as the "interpretation" of population structure (viewed as PCs) or as "arising out of" population structure; ancestry as "a lot more than" population structure, (interpreted as PCs); PCs as "indicators" of ancestry. Some articles view use of PCs as a quality control step to "confirm" participants' self-reported ancestry (Maciukiewicz et al., 2019) or ethnicity (Yoshikawa et al., 2020).

When PCs were used to define a population, this was typically done through drawing ellipses ("bubbles") in PC space. Our interviewees reported that this was mostly either an "eyeball" process, or simply following a rule of thumb because that is what another paper had done. When PCs were used to control for confounding, the number used to do so varied widely, and, when asked, interviewees explained they were either choosing a number based on what a previous paper had used, or followed a rule of thumb (such as the "elbow rule", whereby visual inspection for a dropoff in the amount of variance explained with additional PCs is used to define a cutoff (Cattell 1966)). Few were able to further justify why this was appropriate. While most of our interviewees used PCs somewhat blindly, others drew attention to issues with the use of PCs, including that they depend entirely on the data inputted, that they are hard to interpret and it is not clear what they are actually picking up on, and that they can pick up on non "real" population structure, due to relatedness or QC issues. One interviewee pointed out that many databases are making PCs for their data available, increasingly enabling the "off the shelf" use of this way to operationalize ancestry.

With the use of statistical software packages such as STRUCTURE/ADMIXTURE, which characterize ancestry of an individual in terms of their percentage similarity with different populations, many interviewees showed a greater awareness of the ways it could be arbitrary, reflecting the fact that choices must be made to run it (either by providing reference data of the populations of interest, or by providing the k parameter). For example, one interviewee described the extent to which interesting observations could be made about their data as k varied (e.g., 3 vs. 5 populations using the same data). Some researchers, again those closest to population genetics, offered additional cautions about the use of this methodology to operationalize ancestry, including that it is very easy to over interpret the results, that newer admixture can be confused with identity by descent, and that there never have been "pure" populations. This last observation refers to the fact that the underlying population theoretic model motivating the design of STRUCTURE/ADMIXTURE models an individual as deriving proportions of ancestry from well-defined ancestral populations, but any such populations are themselves mixtures of other populations.

Examples of the "Genetic Intersect Self Report" operationalizations are found in articles that used the United Kingdom Biobank "White British" data (Harrison et al., 2020) (Kolin et al., 2020), which uses a "bubble" in genetic principal component space combined with a filter on an ethnicity category, and others with self-reported information that was further analyzed using multi-dimensional scaling (MDS) in PLINK (Avinum et al., 2020).

We observed inconsistent use of terminology to describe ancestry operationalized using genetic data. Articles were very mixed about whether they referred to this as "genetic ancestry", with about equal numbers systematically using "genetic ancestry" or using just "ancestry", and with many others using a mixture of these two. Other terms in use include "population ancestry", "genomic ancestry", "ancestral population structure", and "DNA ancestry". Additionally, some articles distinguished global from local ancestry, but meant different things by this distinction. For some, "global ancestry" means the percentage of ancestral populations inferred by

TABLE 4 Types of data used to operationalize ancestry by articles in different fields. Note that the 8 articles that operationalized ancestry in more than one way appear more than once in this table.

Data type	Primary field					
	Anthropology N (%)	Biology N (%)	Medicine N (%)	Public health N (%)	Sociology N (%)	Total N (%)
Genetic	4 (29)	22 (45)	22 (40)	12 (39)	1 (13)	61 (39)
Genetic Intersect Non-genetic	–	4 (8)	3 (5)	4 (13)	–	11 (7)
Non-genetic	8 (57)	8 (16)	11 (20)	9 (29)	6 (75)	42 (27)
Not specified	2 (14)	15 (31)	19 (35)	6 (19)	1 (13)	43 (27)
Total	14 (100)	49 (100)	55 (100)	31 (100)	8 (100)	157 (100)

TABLE 5 Types of population labels employed by articles that use different types of data to operationalize ancestry. Because continental and continental region categories are those geographical categories most likely to be conflated with racial categories, we separate out those articles that used a mixed set of labels into those that represented a mixture between these types of label (continent, continental region, and race) from other types of mixtures.

Population labels used in the operationalization of ancestry	Type(s) of data used to operationalize ancestry N (%)				
	Genetic	Genetic and non-genetic	Non-genetic	Not specified	Total
Continent	29 (48)	4 (36)	15 (36)	24 (56)	72 (46)
Continental region	2 (3)	1 (9)	4 (10)	1 (2)	8 (5)
Country		1 (9)	3 (7)		4 (3)
Ethnicity	1 (2)		6 (14)	3 (7)	10 (6)
Mixed: just continent, continental region, race	9 (15)	2 (18)	1 (2)	6 (14)	17 (11)
Mixed: not just continent, continental region, race	7 (11)		7 (17)	7 (16)	21 (15)
No labels used	12 (20)	1 (9)		1 (2)	14 (9)
Others	1 (2)	2 (18)	3 (7)	1 (2)	7 (4)
Race			3 (7)		3 (2)
Total	61 (100)	11 (100)	42 (100)	43 (100)	157 (100)

programs like ADMIXTURE (typically, but not always, continental ancestral populations) whereas "local ancestry" involves assigning population labels to sections of chromosome (others used "chromosomal ancestry" for this (Chen et al., 2020)). For other researchers, "global ancestry" means continental ancestry, and "local ancestry" indicates finer-resolution categories, e.g., country-level ancestry.

We observed a similar lack of clarity around how the term "population" is used by researchers. For some, a population is a model from population genetics implying random mating. For others, drawing from statistics, data from a sample should be chosen such that it is representative of a population. But for most, usage was more diffuse, implying simply a group of people with something (anything) in common, or as one interviewee put it, simply "large N" (where N is the sample size).

We investigated how the data type used to operationalize ancestry varies across fields of study: anthropology, biology, medicine, public health, and sociology, see Table 4. All fields of study used both genetic and other data types to operationalize ancestry. Genetic data is the most common data type used in biology, (22 operationalizations, 45% of those in biology), medicine (22, 40%), and public health (12, 39%). Of the 157 operationalizations, 119 appeared in articles that used genetic data, and it was not the case that they exclusively used genetic data to do so. Only 61 (51%) used exclusively genetic information, 11 (9%) used genetic and non-genetic information, 18 (15%) used exclusively non-genetic information (both geographic and self report), and 29 (24%) did not specify which type of data was used. The rates of not specifying what data were used to operationalize ancestry were particularly high in biology (15, 31%) and medicine (19, 35%).

Throughout our analysis of the operationalization of ancestry we observed many sources of conflation between ancestry and race and ethnicity. As mentioned, it seems likely that many of the participants who were listed as self-reporting their ancestry actually self-reported their race or ethnicity. This is likely also true for the articles where it was not specified what type of data they used to operationalize ancestry. Other articles referred to their ancestry categories inferred from genetic data as ethnicities (Bani-Fatemi et al., 2019). One article refers to "race assigning" using either self report or tertile of genetically inferred West African ancestry (Gohlke et al., 2019).

What types of population labels are used for ancestry categories?

Just as there is diversity in what type of data is used to operationalize ancestry, there is also diversity in the types of population labels used to describe the resulting categories (see Table 5). Continental ancestry is the most used population label (68 articles, 43%). This is followed by the labels representing mixed types (not just continent, continental region, race) (24, 15%). The labels in the "Other" category were mostly *White British ancestry*, a label from the United Kingdom Biobank.

A combined analysis of population labels and type of data used to operationalize ancestry (see Table 5) reveals that no matter what type of data is used (and when it was not clear what type of data was used), continental ancestry categories are the most common. This usage was highest (56%) when no indication is given of what type of data is used to operationalize ancestry. For all data types, most other types of labels besides continental are also used.

Among the articles which use ancestry categories, most types of population labels are used by articles from all the fields of study (see Supplementary Table S6). Continental ancestry is the most used population label in biology (23 articles, 47%) medicine (28, 51%), and public health (15, 48%). In sociology, ethnicity was the most popular label type (3, 38%). Anthropology was marked by a spread of different label types.

An analysis of how population labels are used across the authors' country/region of institutional affiliation (see Supplementary Table S7) indicates that for those based everywhere save Asia, Continental ancestry labels are the most commonly used. In Asia, researchers most often used Mixed labels—a combination of continent and a country is typical (e.g., European and Japanese).

We observed several instances of grand generalizations from samples from populations with ethnic or country labels to continental groups, for example from Japanese ancestry to Asian ancestry (Brhane et al., 2020), and from French ancestry to European ancestry (Tonon et al., 2019).

Discussion

In this discussion we first summarize our main findings, and then reflect on the normative consequences of our findings for the type of research we included in this empirical work. We start by noting that our results are limited by the types of articles, and subsequently, researchers, that we included. There are some types of research questions, namely those that do not have even a tangential relationship to human health, that we do not cover.

Our results indicate that ancestry is a concept that is drawn upon in multiple key ways across a broad range of types of research. This is particularly true when it comes to understanding traits and outcomes, from blood pressure to income. For researchers seeking to identify genetic variation linked to these traits, ancestry is evoked as a central part of the methodology, as something to be controlled for. For researchers interested in understanding social outcomes who are not interested in identifying genetic variants, ancestry is again something to be controlled for. For other researchers, ancestry shapes their research question; they hope that studying how a trait varies with ancestry can enable the identification of genetic contributions to between-group differences, including health disparities. Having established that researchers evoke ancestry in key ways, understanding what the concept means to them becomes central.

We demonstrate a huge diversity of understandings of ancestry amongst researchers. Several observations stand out. First, the concept of ancestry encompasses much more than genetics. Many of our interviewees, including geneticists, gave definitions that were broader than what could be inferred from genetic data, stressing for example the narrative aspect of ancestry. The example of "indigenous ancestry" also illustrates that ancestry is more than genetics; several indigenous groups have explicitly rejected genetics as relevant to questions of ancestry, in favor of cultural affiliation (TallBear 2013). Attempting to secure the term "ancestry" to refer to genetic variation—as done in e.g., (Mersha and Tilahun, 2015)—is not a viable strategy. Second, there is an absence of agreement on what is core to the concept of ancestry, for example, whether it has anything to do with geography. Third, many researchers, who were selected to be part of our sample precisely because their work engaged closely with the concept of ancestry, struggled to define it. Fourth, while some researchers stressed that ancestry was fundamentally different from the social constructs of race and ethnicity, we observed both in articles and in interviews frequent slippage between ancestry, race and/or ethnicity. Our results, which highlight the ways in which ancestry is ambiguous, can be read in the light of the broader literature on ambiguity in scientific concepts (reviewed in (Panofsky and Bliss 2017)). Some of this literature highlights positive roles that ambiguity can play, while the majority of the literature highlights negative functions of ambiguity, and the advantages that flow from standardization, see for example (Timmermans and Epstein 2010).

We also report a huge diversity of operationalizations of the concept of ancestry. The process of operationalization involves taking a definition of an abstract concept and making it measurable. Given the diversity of definitions of ancestry, it is perhaps not surprising there are so many operationalizations. The fundamental ambiguity of the concept, as discussed above, is also evidenced by our result that a quarter of all papers that operationalize ancestry fail to state anything about how this was done, for example whether inclusion criteria were based on self-reported information, geography, or genetics. This was particularly the case in Biology and Medicine, with Anthropology and Sociology articles more frequently stating what type of data was used to operationalize ancestry, perhaps reflecting greater sensitivity in these fields to the ways in which these categories reflect decisions made by researchers. The failure to specify what type of data is used to operationalize ancestry allows for the introduction of further ambiguity between genetics and social identities, particularly given that researchers often use "ancestry" language (rather than "genetic ancestry") when genetic data has been used to operationalize the concept, and particularly when ancestry is operationalized more than once in the same article, sometimes using genetic data and sometimes not.

Our results also highlight that practices associated with using genetic data to operationalize ancestry rest on unclear conventions, and that there is a lack of clarity concerning the relationships between the different key concepts. This is particularly true for the use of Principal Components. PCs are sometimes referred to as ancestry, genetic ancestry, population structure, or some more qualified term, suggesting that they "capture" or "correlate" with one of the above. Researchers often justified their choices about the use of PCs by reference to prior papers, without being able to give their own justifications of why those choices were appropriate. This may be concerning, given the growing usage of PCs not only within statistical genetics but by those seeking "off the shelf" solutions to "control for genetics" (Boardman et al., 2010). It should also be concerning given that results can depend critically on choices made (Elhaik 2022).

The predominance of continental ancestry labels, particularly when genetic data is used and when it is not specified what type of data is used, reinforces concerns that the turn to genetic ancestry may just essentialize the quasi-racial groupings represented by these categories (Bliss 2020b; Lewis et al., 2022). On the other hand, the diversity of population labels in use helps demonstrate that researchers have a wide range of choices when it comes both to their operationalization of ancestry, and to their framing of their results.

In repeated guidelines for the use of population descriptors, transparency of what is meant by the concepts and how they are operationalized is presented as a minimum bar (Mauro et al., 2022). Our results show that, in the case of ancestry specifically, current research is very far from meeting this bar. Our results also help indicate why the goal of transparency may be hard to achieve: there is a huge diversity of ideas underlying the concept, which leads to a deep-running ambiguity about where the concept sits in relation to biology on the one hand and social identities on the other. The existing empirical work discussed in the background demonstrates the close links between racial ways of thinking and the way genetic ancestry is being conceptualized. By focusing our work not just on genetic ancestry, but on ancestry more broadly, we demonstrate additional ways in which this conflation happens.

The conflation between genetic and social ways of conceptualizing human difference—aided by the highly ambiguous term "ancestry"—is problematic because the concepts are importantly different (Cerdeña, Grubbs, and Non 2022). As Ian Hacking has pointed out, whereas there is a sense in which all concepts are socially constructed, some concepts are "interactive kinds," in which the entities being classified know they are being classified, and act and are treated differently based on the ways that they are classified (Hacking 2001). People learn to treat people categorized by a structural system differently, depending on their category. This is true of race, which is a construct invented by white Europeans to secure their racial privilege (Omi and Howardt 2015). Genetic ancestry refers to how DNA is passed down through the human family tree (Mathieson and Scally 2020; Lewis et al., 2022). This human family tree has a complex structure; it is after all shaped by everything that shapes who has children with whom, which includes amongst other things geography and cultural practices. There are of course correlations between race and ethnicity and patterns of genetic variation, including at the continental level. But genetic variation is continuous, not categorical and not best represented by continental categories (Lewis et al., 2022). And whereas racial labels are a result of sociopolitical processes, it is researchers who choose to impose categories on genetic data, and then to attach labels to those categories. As Bonham et al. write, "It is critical to avoid creating fictitious, discrete genomic groups while recognizing that self-identified race and ethnicity are highly associated with genetic ancestry at the continental and population level" (Bonham, Green, and Pérez-Stable 2018).

The ambiguity that use of the term "ancestry" provides is almost certainly helpful to researchers in some ways (Panofsky and Bliss 2017). But it also confuses our attempts to gain genuine understanding of the dynamics and processes in the creation of health outcomes. The motivation for much genetic research is to contribute to a causal understanding of why health-related traits are distributed the way they are. Genetic ancestry cannot be a causal factor in such analysis, it can only act as a proxy for underlying genetic variants that are playing a causal role. Racism can play a causal role, through many mechanisms that are being elucidated (see e.g., (Krieger 2021)). Inadequate reflection of the relation between these systems for capturing human difference can lead to incorrect conclusions. For example, attempts to explain differences in health outcomes based on differences in genetic ancestry are confounded by racism (Boulter et al., 2015). To make progress in accurately understanding the distribution of health outcomes, the different roles of genetic variation and social and environmental factors must be carefully considered.

While we agree with the dozens of other commentators on the importance of researchers transparently describing exactly who was studied, how they were classified, and why, we advocate for the following additional considerations.

(1) Use of the term "ancestry" by itself should be avoided. Rather, this term should always be qualified, for example as "genealogical ancestry" or "genetic ancestry".
(2) Use of the term "population" should be avoided. It has no agreed upon meaning and only serves to make something sound more scientific than it in fact is. This is particularly true in genetics research when the term is apt to be confused with the term in population genetics theory for a group of individuals who are mating at random. Use of the term "group" is to be preferred, because it correctly draws attention to the question "by virtue of what are these individuals being grouped together" and because it evokes less scientific authority.
(3) Operationalizations of genetic ancestry that reflect the continuous nature of genetic variation, such as PCA, should be encouraged wherever possible (Duello et al., 2021; Lewis et al., 2022).
(4) If genetic ancestry must be operationalized using categories, multiple sets of categories should be used, to reflect the fact that one can carve up the human family tree in multiple ways (Lewis et al., 2022).
(5) Genetically inferred continental ancestry categories should only be used if necessary (Panofsky and Bliss 2017; Lewis et al., 2022).
(6) Researchers need to make themselves familiar with the key limitations of the tools they use.

The hope of some is that ancestry represents the biological, objective counterpart to the social constructs of race and ethnicity, and that a turn to ancestry categories could help us get away from the bad science and damaging history of previous classification systems and practices. Our investigation of how ancestry is actually used indicates it is a far cry from the objective and straightforward concept hoped for. Nor is it a uniquely "biological construct". Indeed, the concept is fundamentally ambiguous, and not more conceptually clear than race or ethnicity in practice. By just moving to a new term, there is a danger that research in this area fails to address the known issues in the uses of race and ethnicity as population descriptors. Part of the problem is precisely that some scientists are searching for a more objective term, thus treating this set of issues as purely semantic when in fact the problems run deeper: we need more careful attention to the different roles of genetic variants (which are not uniformly distributed) on the one hand and the myriad other contributors to health outcomes on the other. The move to "ancestry" just confuses the issue. Given the central role ancestry plays in genetics research and beyond, these deep-seated issues with how ancestry is conceptualized and operationalized should raise concern, and highlight the importance of strategies that will advance conceptual clarity.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Harvard University-Area Committee on the Use of Human Subjects. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

BD: coded articles for systematic literature analysis; SM: conducted interviews and coded interviews; NG: transcribed interviews; SZ: transcribed interviews and coded interviews; AL: conducted interviews, coded interviews, coded articles, wrote original draft. All authors: Conceptualized project, and reviewed and edited manuscript.

Funding

NIMH administrative supplement 5000747-5500001474 to 3R37MH107649-06S1. AP acknowledges support from the Carnegie Foundation.

Conflict of interest

AL owns stock in Fabric Genomics; BN 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.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1044555/full#supplementary-material

References

Ali-Khan, S. E., Krakowski, T., Tahir, R., and Daar, A. S. (2011). The use of race, ethnicity and ancestry in human genetic research. *HUGO J.* 5 (1–4), 47–63. doi:10.1007/s11568-011-9154-5

Apprey, V., Wang, S., Tang, W., Kittles, R. A., Southerland, W. M., Ittmann, M., et al. (2019). Association of genetic ancestry with DNA methylation changes in prostate cancer disparity. *Anticancer Res.* 39 (11), 5861–5866. doi:10.21873/anticanres.13790

Arleo, T., Tong, D., Shabto, J., O'Keefe, G., and Khosroshahi, A. (2021). Clinical course and outcomes of COVID-19 in rheumatic disease patients: A case cohort study with a diverse population. *Clin. Rheumatol.* 40 (7), 2633–2642. doi:10.1007/s10067-021-05578-x

Avinum, R., Romer, A. L., and Israel, S. (2020). Vitamin D polygenic score is associated with neuroticism and the general psychopathology factor. *Prog. Neuropsychopharmacol. Biol. Psychiatry.* 100, 109912. doi:10.1016/j.pnpbp.2020.109912

Azcorra, H., Vázquez-Vázquez, A., Mendez, N., Salazar, J. C., and Datta-Banik, S. (2016). Maternal maya ancestry and birth weight in yucatan, Mexico. *Am. J. Hum. Biol.* 28 (3), 436–439. doi:10.1002/ajhb.22806

Baer, R. D., Arteaga, E., Dyer, K., Eden, A., Gross, R., Helmy, H., et al. (2013). Concepts of race and ethnicity among health researchers: Patterns and implications. *Ethn. Health* 18 (2), 211–225. doi:10.1080/13557858.2012.713091

Bakhtiari, E. (2020). Health effects of muslim racialization: Evidence from birth outcomes in California before and after september 11, 2001. *SSM - Popul. Health* 12, 100703. doi:10.1016/j.ssmph.2020.100703

Bani-Fatemi, A., Graff, A., Gerretsen, P., Dada, O. O., Kennedy, J. L., Hettige, N., et al. (2019). The effect of ethnicity and immigration on treatment resistance in schizophrenia. *Compr. Psychiatry* 89, 28–32. doi:10.1016/j.comppsych.2018.12.003

Batai, K., Hooker, S., and Kittles, R. A. (2021). Leveraging genetic ancestry to study health disparities. *Am. J. Phys. Anthropol.* 175 (2), 363–375. doi:10.1002/ajpa.24144

Bentley, A. R., Callier, S., and Rotimi, C. N. (2017). Diversity and inclusion in genomic research: Why the uneven progress? *J. Community Genet.* 8 (4), 255–266. doi:10.1007/s12687-017-0316-6

Bliss, C. (2020b). Conceptualizing race in the genomic age. *Hastings Cent. Rep.* 50, S15. doi:10.1002/hast.1151

Bliss, C. (2020a). *Race decoded: The genomic fight for social justice*. Stanford, CA: Stanford University Press.

Boardman, J. D., Blalock, C. L., Stallings, M. C., Domingue, B. W., McQueen, M. B., Crowley, T. J., et al. (2010). Ethnicity, body mass, and genome-wide data. *Biodemogr. Soc. Biol.* 56 (2), 123–136. doi:10.1080/19485565.2010.524589

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 (15), 1533–1534. doi:10.1001/jama.2018.13609

Borrell, L. N., Elhawary, J. R., Fuentes-Afflick, E., Witonsky, J., Bhakta, N., AlanWu, H. B., et al. (2021). Race and genetic ancestry in medicine - a time for reckoning with racism. *N. Engl. J. Med.* 384 (5), 474–480. doi:10.1056/NEJMms2029562

Boulter, A. C., Jacklyn QuinlanMiró-Herrans, A. T., Pearson, L. N., Todd, N. L., Gravlee, C. C., Mulligan, C. J., et al. (2015). Interaction of alu polymorphisms and novel measures of discrimination in association with blood pressure in african Americans living in tallahassee, Florida. *Hum. Biol.* 87 (4), 295–305. doi:10.13110/humanbiology.87.4.0295

Brhane, Y., Yang, P., Christiani, D. C., Liu, G., McLaughlin, J. R., Brennan, P., et al. (2020). Genetic determinants of lung cancer prognosis in never smokers: A pooled analysis in the international lung cancer consortium. *Cancer Epidemiol. Biomarkers Prev.* 29 (10), 1983–1992. doi:10.1158/1055-9965.EPI-20-0248

Byeon, Y. J. J., Islamaj, R., Yeganova, L., John Wilbur, W., Lu, Z., Brody, L. C., et al. (2021). Evolving use of ancestry, ethnicity, and race in genetics research—a Survey spanning seven decades. *Am. J. Hum. Genet.* 108 (12), 2215–2223. doi:10.1016/j.ajhg.2021.10.008

Canter, D. J., Reid, J., Latsis, M., Variano, M., Halat, S., Rajamani, S., et al. (2019). Comparison of the prognostic utility of the cell cycle progression score for predicting clinical outcomes in african American and non-african American men with localized prostate cancer. *Eur. Urol.* 75 (3), 515–522. doi:10.1016/j.eururo.2018.10.028

Cattell, R. B. (1966). The scree test for the number of factors. *Multivar. Behav. Res.* 1 (2), 245–276. doi:10.1207/s15327906mbr0102_10

Cerdeña, J. P., Grubbs, V., and Amy, L. (2022). Genomic supremacy: The harm of conflating genetic ancestry and race. *Hum. Genomics* 16 (1), 18. doi:10.1186/s40246-022-00391-2

Chen, X., Bhuiyan, I., Kuja-Halkola, R., Magnusson, P. K. E., and Svensson., P. (2019). Genetic and environmental influences on the correlations between traits of metabolic syndrome and ckd. *Clin. J. Am. Soc. Nephrol.* 14 (11), 1590–1596. doi:10.2215/CJN.11971018

Chen, Y., Sadasivan, S. M., She, R., Datta, I., Taneja, K., Chitale, D., et al. (2020). Breast and prostate cancers harbor common somatic copy number alterations that consistently differ by race and are associated with survival. *BMC Med. Genomics* 13 (1), 116–215. doi:10.1186/s12920-020-00765-2

Darst, B. F., Wan, P., Sheng, X., Bensen, J. T., Ingles, S. A., Rybicki, B. A., et al. (2020). A germline variant at 8q24 contributes to familial clustering of prostate cancer in men of African ancestry. *Eur Urol.* 78 (3), 316–320. doi:10.1016/j.eururo.2020.04.060

Dina (2022). *Race and ethnicity-related 2022 MeSH changes*. Available at: https://www.nlm.nih.gov/mesh/meshhome.html.

Ding, M., Ahmad, S., Lu, Q., Hu, Y., Bhupathiraju, S. N., Guasch-Ferré, M., et al. (2020). Additive and multiplicative interactions between genetic risk score and family history and lifestyle in relation to risk of type 2 diabetes. *Am. J. Epidemiol.* 189 (5), 445–460. doi:10.1093/AJE/KWZ251

Du, Z., Hopp, H., Sue, A., Huff, C., Weaver, B., Stern, M., et al. (2020). Ingles, Chad Huff, Xin Sheng, Brandi Weaver, Mariana Stern, et alA Genome-Wide Association Study of Prostate Cancer in Latinos. *Int. J. Cancer* 146 (7), 1819–1826. doi:10.1002/ijc.32525

Duello, T. M., Rivedal, S., Wickland, C., and Weller, A. (2021). Race and genetics versus 'race' in genetics: A systematic review of the use of african ancestry in genetic studies. *Evol. Med. Public Health* 9 (1), 232–245. doi:10.1093/emph/eoab018

Dupont, W. D., Breyer, J. P., Plummer, W. D., Chang, S. S., Cookson, M. S., Smith, J. A., et al. (2020). 8Q24 genetic variation and comprehensive haplotypes altering familial risk of prostate cancer. *Nat. Commun.* 11 (1), 1523. doi:10.1038/s41467-020-15122-1

Elhaik, Eran (2022). Principal component analyses (PCA)-Based findings in population genetic studies are highly biased and must Be reevaluated. *Sci. Rep.* 12 (1), 14683. doi:10.1038/s41598-022-14395-4

Emami, N. C., Kachuri, L., Meyers, T. J., Das, R., Hoffman, J. D., Hoffmann, T. J., et al. (2019). Association of imputed prostate cancer transcriptome with disease risk reveals novel mechanisms. *Nat. Commun.* 10, 3107. doi:10.1038/s41467-019-10808-7

Flanagin, A., Frey, T., and Stacy, L.AMA Manual of Style Committee (2021). Christiansen, and AMA manual of style CommitteeUpdated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* 326 (7), 621–627. doi:10.1001/jama.2021.13304

Fleskes, R. E., Ofunniyin, A. A., Joanna, K. G., Eric Poplin, S. M. A., Bueschgen, W. D., Juarez, C., et al. (2021). Ancestry, health, and lived experiences of enslaved africans in 18th century charleston: An osteobiographical analysis. *Am. J. Phys. Anthropol.* 175 (1), 3–24. doi:10.1002/ajpa.24149

Franceschini, N., and Morris, A. P. (2020). Genetics of kidney traits in worldwide populations: The continental origins and genetic epidemiology network (COGENT) kidney consortium. *Kidney Int.* 98 (1), 35–41. doi:10.1016/j.kint.2020.02.036

Fritz, J., Lopez-Ridaura, R., Choudhry, S., Razo, C., and Lamadrid-Figueroa, H. (2020). The association of native American genetic ancestry and high-density lipoprotein cholesterol: A representative study of a highly admixed population. *Am. J. Hum. Biol.* 32 (6), 1–11. doi:10.1002/ajhb.23426

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 (1), 5–30. doi:10.1177/0306312710379170

Gendy, M. N. S., Clement, Z., Le Foll, B., and Kennedy, J. L. (2019). Association study of OPRM1 gene in a sample of schizophrenia patients with alcohol dependence or abuse. *Can. J. Addict.* 10 (4), 30–34. doi:10.1097/CXA.0000000000000069

Gohlke, J. H., Lloyd, S. M., Basu, S., Putluri, V., Vareed, S. K., Rasaily, U., et al. (2019). Methionine-homocysteine pathway in african-American prostate cancer. *JNCI Cancer Spectr.* 3 (2), pkz019. doi:10.1093/JNCICS/PKZ019

Grizzle, W. E., Kittles, R. A., Rais-Bahrami, S., Shah, E., Adams, G. W., DeGuenther, M. S., et al. (2019). Self-identified african Americans and prostate cancer risk: West african genetic ancestry is associated with prostate cancer diagnosis and with higher gleason sum on biopsy. *Cancer Med.* 8 (16), 6915–6922. doi:10.1002/cam4.2434

Groeger, J., Opler, M., Kleinhaus, K., Perrin, M. C., Calderon-Margalit, R., and Manor, O. (2017). Live birth sex ratios and father's geographic origins in Jerusalem, 1964–1976. *Am. J. Hum. Biol.* 29 (3). doi:10.1002/ajhb.22945

Groopman, E. E., Marasa, M., Cameron-Christie, S., Petrovski, S., Aggarwal, V. S., Milo-Rasouly, H., et al. (2019). Maddalena Marasa, Sophia Cameron-Christie, Slavé Petrovski, Vimla S. Aggarwal, Hila Milo-Rasouly, Yifu Li, et alDiagnostic Utility of Exome Sequencing for Kidney Disease. *N. Engl. J. Med.* 380 (2), 142–151. doi:10.1056/nejmoa1806891

Guan, F., Zhang, T., Han, W., Zhu, L., Tong, N., Lin, H., et al. (2020). Relationship of SNAP25 variants with schizophrenia and antipsychotic-induced weight change in large-scale schizophrenia patients. *Schizophrenia Res.* 215, 250–255. doi:10.1016/j.schres.2019.09.015

Gur, R. E., Tyler, M., Moore, A. F. G. R., Barzilay, R., Roalf, D. R. M. E. C., Ruparel, K., et al. (2019). Burden of environmental adversity associated with psychopathology, maturation, and brain behavior parameters in youths. *JAMA Psychiatry* 76 (9), 966–975. doi:10.1001/jamapsychiatry.2019.0943

Haas Pizarro, M., Santos, D. C., Melo, L. G. N., Muniz, L. H., Porto, L. C., Silva, D. A., et al. (2020). Glomerular filtration rate estimated by the chronic kidney disease epidemiology collaboration (CKD-EPI) equation in type 1 diabetes based on genomic ancestry. *Diabetology Metabolic Syndrome* 12 (1), 71–77. doi:10.1186/s13098-020-00578-4

Hacking, I. (2001). *The social construction of what? 7*. Cambridge, Mass: Harvard Univ. Press.

Hamilton, T. G., and Green, T. L. (2018). From the West Indies to Africa: A universal generational decline in health among blacks in the United States. *Soc. Sci. Res.* 73, 163–174. doi:10.1016/j.ssresearch.2017.12.003

Harlemon, M., Ajayi, O., Kachambwa, P., Kim, M. S., Simonti, C. N., Quiver, M. H., et al. (2020). A custom genotyping array reveals population-level heterogeneity for the genetic risks of prostate cancer and other cancers in africa. *Cancer Res.* 80 (13), 2956–2966. doi:10.1158/0008-5472.CAN-19-2165

Harrison, S., Davies, A. R., Dickson, M., Tyrrell, J., Michael, J. G., Katikireddi, S. V., et al. (2020). The causal effects of health conditions and risk factors on social and socioeconomic outcomes: Mendelian randomization in UK Biobank. *Int. J. Epidemiol.* 49 (5), 1661–1681. doi:10.1093/ije/dyaa114

Hill, J., Delville, C. L., Auorousseau, A. M., Jonathan, D., Peer, N., Oldenburg, B., et al. (2020). Anne marie auorousseau, deborah jonathan, nasheeta peer, brian oldenburg, and andre pascal KengneDevelopment of a tool to increase physical activity among people at risk for diabetes in low-resourced communities in Cape town. *Int. J. Environ. Res. Public Health* 17 (3), 865. doi:10.3390/ijerph17030865

Hooker, S. E., Woods-Burnham, L., Bathina, M., Lloyd, S., Gorjala, P., Mitra, R., et al. (2019). Genetic ancestry analysis reveals misclassification of commonly used cancer cell lines. *Cancer Epidemiol. Biomarkers Prev.* 28 (6), 1003–1009. doi:10.1158/1055-9965.EPI-18-1132

Howe, L. D., Kanayalal, R., Harrison, S., BeaumontDavies, R. N. A. R. T. M. F., Davies, N. M., Frayling, T. M., et al. (2020). Effects of body mass index on relationship status, social contact and socio-economic position: Mendelian randomization and within-sibling study in UK Biobank. *Int. J. Epidemiol.* 49 (4), 1173–1184. doi:10.1093/ije/dyz240

Hunley, K., Edgar, H., Healy, M., Mosley, C., Cabana, G. S., and West, F. (2017). Social identity in new Mexicans of Spanish-speaking descent highlights limitations of using standardized ethnic terminology in research. *Hum. Biol.* 89 (3), 217–228. doi:10.13110/humanbiology.89.3.04

Jordan, D. M., Choi, H. K., Verbanck, M., Topless, R., Won, H. H., Nadkarni, G., et al. (2019). No causal effects of serum urate levels on the risk of chronic kidney disease: A mendelian randomization study. *PLoS Med.* 16 (1), 10027255–e1002815. doi:10.1371/journal.pmed.1002725

Kaur, H. B., Jiayun Lu, L. B. G., Maldonado, L., Logan, R., Barber, J. R., De Marzo, A. M., et al. (2019). TP53 missense mutation is associated with increased tumor-infiltrating T cells in primary prostate cancer. *Hum. Pathol.* 87, 95–102. doi:10.1016/j.humpath.2019.02.006

Khan, A. T., Stephanie, M., Gogarten, C. P., McHugh, A. M. S., Sofer, T., Bowers, M. L., et al. (2022). Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: Experiences from the NHLBI TOPMed program. *Cell Genomics* 2 (8), 100155. doi:10.1016/j.xgen.2022.100155

Koga, Y., Song, H., Chalmers, Z. R., Newberg, J., Kim, E., Carrot-Zhang, J., et al. (2020). Genomic profiling of prostate cancers from men with african and European ancestry. *Clin. Cancer Res.* 26 (17), 4651–4660. doi:10.1158/1078-0432.CCR-19-4112

Kolin, D. A., Scott, K., Christos, P. J., and Olivier, E. (2020). Clinical, regional, and genetic characteristics of covid-19 patients from UK Biobank. *PLoS ONE* 15, e0241264. doi:10.1371/journal.pone.0241264

Krieger, N. (2021). "Ecosocial theory, embodied truths, and the people's health," in *Small books, big ideas in population health* (New York, NY: Oxford University Press), 4.

Lam, M., Chen, C-Y., Li, Z., Martin, A. R., Broyis, J., Ma, X., et al. (2019). Comparative genetic architectures of schizophrenia in east asian and European populations. *Nat. Genet.* 51 (12), 1670–1678. doi:10.1038/s41588-019-0512-x

Lee, S. S., Mountain, J., and Koenig, B. A. (2001). The meanings of 'race' in the new genomics: Implications for health disparities research. *Yale J. Health Policy, Law, Ethics* 1, 33–75.

Legge, S. E., Pardiñas, A. F., Helthuis, M., Jansen, J. A., Karel, J., Knapper, S., et al. (2019). A genome-wide association study in individuals of african ancestry reveals the importance of the duffy-null genotype in the assessment of clozapine-related neutropenia. *Mol. Psychiatry* 24 (3), 328–337. doi:10.1038/s41380-018-0335-7

Leishangthem, S., Kushwaha, K. P. S., Chauhan, T., Kumawat, R. K., Chaubey, G., and Shrivastava, P. (2020). Evaluation of the genomic diversity and shared ancestry of the Meitei community of Manipur (India) with the east asian populations using autosomal STRs. *Ann. Hum. Biol.* 47 (7–8), 642–651. doi:10.1080/03014460.2020.1821772

Lewis, A. C. F., Molina, S. J., Paul, S. A., Dauda, B., Di Rienzo, A., Fuentes, A., et al. (2022). Getting genetic ancestry right for science and society. *Science* 376 (6590), 250–252. doi:10.1126/science.abm7530

Liebler, C. A. (2016). On the boundaries of race. *Sociol. Race Ethn.* 2 (4), 548–568. doi:10.1177/2332649216632546

Lin, B. M., Nadkarni, G. N., Tao, R., Graff, M., Fornage, M., Buyske, S., et al. (2019). Genetics of chronic kidney disease stages across ancestries: The PAGE study. *Front. Genet.* 10, 494. doi:10.3389/fgene.2019.00494

Liu, Y. P., Wu, X., Xia, X., Yao, J., and Wang, B. J. (2020). The genome-wide supported CACNA1C gene polymorphisms and the risk of schizophrenia: An updated meta-analysis. *BMC Med. Genet.* 21 (1), 159–212. doi:10.1186/s12881-020-01084-0

Macholdt, E., Montgomery, S., Pakendorf, B., and Stoneking, M. (2015). New insights into the history of the C-14010 lactase persistence variant in eastern and southern africa. *Am. J. Phys. Anthropol.* 156 (4), 661–664. doi:10.1002/ajpa.22675

Maciukiewicz, M., Gorbovskaya, I., Laughlin, C. P., Nurmi, E. L., Liebermann, J. A., Meltzer, H. Y., et al. (2019). Genome-wide association study on antipsychotic-induced weight gain in Europeans and african-Americans. *Schizophrenia Res.* 212, 204–212. doi:10.1016/j.schres.2019.07.022

Marden, J. R., Walter, S., Kaufman, J. S., and Maria Glymour, M. (2016). African ancestry, social factors, and hypertension among non-hispanic blacks in the health and retirement study. *Biodemogr. Soc. Biol.* 62 (1), 19–35. doi:10.1080/19485565.2015.1108836

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51 (4), 584–591. doi:10.1038/s41588-019-0379-x

Martínez-Magaña, J., Gonzalez-Castro, T. B., Genís-Mendoza, A. D., Tovilla-Zárate, C. A., Juárez-Rojop, I. E., Saucedo-Uribe, E., et al. (2019). Exploratory analysis of polygenic risk scores for psychiatric disorders: Applied to dual diagnosis. *Rev. Investig. Clinica; Organo Del Hosp. Enfermedades La Nutr.* 71 (5), 321–329. doi:10.24875/RIC.19003013

Marziali, M. E., Card, K. G., Taylor, M., Closson, K., Wang, L., Trigg, J., et al. (2021). Correlates of social isolation among people living with HIV in British columbia, Canada. *AIDS Care - Psychol. Socio-Medical Aspects AIDS/HIV* 33 (5), 566–574. doi:10.1080/09540121.2020.1757607

Mathieson, I., and Scally, A. (2020). What is ancestry?" Edited by jonathan flint. *PLOS Genet.* 16 (3), e1008624. doi:10.1371/journal.pgen.1008624

Mauro, M., Allen, D. S., Dauda, B., Molina, S. J., Neale, B. M., and Lewis, A. C. F. (2022). "A systematic review of guidelines for the use of race, ethnicity, and ancestry reveals widespread consensus but also points of ongoing disagreement." arXiv.

Menzies, R., Aqel, J., Abdi, I., Joseph, T., Seale, H., and Nathan, S. (2020). Why is influenza vaccine uptake so low among aboriginal adults? *Aust. N. Z. J. Public Health* 44 (4), 279–283. doi:10.1111/1753-6405.13004

Mersha, T. B., and Tilahun, A. (2015). Self-reported race/ethnicity in the age of genomic research: Its potential impact on understanding health disparities. *Hum. Genomics* 9, 1. doi:10.1186/s40246-014-0023-x

National Academies (2021). *Use of race, ethnicity, and ancestry as population descriptors in genomics research*. Washington, DC: National Academies. Available at: https://www.nationalacademies.org/our-work/useof-race-ethnicity-and-ancestry-aspopulation-descriptors-in-genomics-research.

Ohi, K., Nishizawa, D., Shimada, T., Kataoka, Y., Hasegawa, J., Shioiri, T., et al. (2020). Polygenetic risk scores for major psychiatric disorders among schizophrenia patients, their first-degree relatives, and healthy participants. *Int. J. Neuropsychopharmacol.* 23 (3), 157–164. doi:10.1093/ijnp/pyz073

Omi, M., and Howard, W. (2015). *Racial Formation in the United States*. Third edition. New York: Routledge/Taylor & Francis Group.

Oni-Orisan, A., Mavura, Y., Banda, Y., Thornton, T. A., and Sebro, R. (2021). Embracing genetic diversity to improve black health. *Debra Malina N. Engl. J. Med.* 384 (12), 1163–1167. doi:10.1056/NEJMms2031080

Panofsky, A., and Bliss, C. (2017). Ambiguity and scientific authority: Population classification in genomic science. *Am. Sociol. Rev.* 82 (1), 59–87. doi:10.1177/0003122416685812

Paredes, C. L. (2017). Mestizaje and the significance of phenotype in Guatemala. *Sociol. Race Ethn.* 3 (3), 319–337. doi:10.1177/2332649216682523

Passchier, R. V., Stein, D. J., Uhlmann, A., van der Merwe, C., and Dalvie, S. (2020). Schizophrenia polygenic risk and brain structural changes in methamphetamine-associated psychosis in a South African population. *Front. Genet.* 11, 1018–1027. doi:10.3389/fgene.2020.01018

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C-Y., Popejoy, A. B., Periyasamy, S., et al. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell* 179 (3), 589–603. doi:10.1016/j.cell.2019.08.051

Pomeroy, E., Wells, J. C. K., Stanojevic, S., Jaime Miranda, J., Moore, L. G., Cole, T. J., et al. (2015). Surname-inferred andean ancestry is associated with child stature and limb lengths at high altitude in Peru, but not at sea level. *Am. J. Hum. Biol.* 27 (6), 798–806. doi:10.1002/ajhb.22725

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 538 (7624), 161–164. doi:10.1038/538161a

Popejoy, A. B., Ritter, D. I., Crooks, K., Currey, E., Fullerton, S. M., Hindorff, L. A., et al. (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Hum. Mutat.* 39 (11), 1713–1720. doi:10.1002/humu.23644

Reardon, J. (2005). "Race to the finish: Identity and governance in an age of genomics," in *Formation series* (Princeton: Princeton University Press).

Roberts, D. (2011). *Fatal invention: How science, politics, and big business Re-create race in the twenty-first century*. New York: New Press.

Saad, F., Ayyash, M., Ayyash, M., Elhage, N., Ali, I., Makki, M., et al. (2020). Assessing knowledge, physician interactions and patient-reported barriers to colorectal cancer screening among arab Americans in dearborn, Michigan. *J. Community Health* 45 (5), 900–909. doi:10.1007/s10900-020-00807-x

TallBear, K. (2013). Genomic articulations of indigeneity. *Soc. Stud. Sci.* 43 (4), 509–533. doi:10.1177/0306312713483893

Thayer, Z. M., Irene, V., Blair, D. S. B., and Manson, S. M. (2017). Racial Discrimination Associated with Higher Diastolic Blood Pressure in a Sample of American Indian Adults: THAYER et Al. *Am. J. Phys. Anthropol.* 163 (1), 122–128. doi:10.1002/ajpa.23190

Timmermans, S., and Epstein, S. (2010). A world of standards but not a standard world: Toward a sociology of standards and standardization. *Annu. Rev. Sociol.* 36 (1), 69–89. doi:10.1146/annurev.soc.012809.102629

Tonon, L., Fromont, G., Boyault, S., Thomas, E., Ferrari, A., Sertier, A. S., et al. (2019). Mutational profile of aggressive, localised prostate cancer from african caribbean men versus European ancestry men. *Eur. Urol.* 75 (1), 11–15. doi:10.1016/j.eururo.2018.08.026

Vyas, D. A., Eisenstein, L. G., and Jones, D. S. (2021). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *Obstetrical Gynecol. Surv.* 76 (1), 5–7. doi:10.1097/01.ogx.0000725672.30764.f7

Wagner, J. K., Yu, J-H., Ifekwunigwe, J. O., Tanya, M., Harrell, M. J. B., and Royal, C. D. (2017). Anthropologists' views on race, ancestry, and genetics. *Am. J. Phys. Anthropol.* 162 (2), 318–327. doi:10.1002/ajpa.23120

Walavalkar, K., Saravanan, B., Singh, A. K., Singh Jayani, R., Nair, A., Farooq, U., et al. (2020). A rare variant of african ancestry activates 8q24 LncRNA hub by modulating cancer associated enhancer. *Nat. Commun.* 11 (1), 3598. doi:10.1038/s41467-020-17325-y

Wasterlain, S. N., Costa, A., and Ferreira, M. T. (2018). Growth faltering in a skeletal sample of enslaved nonadult africans found at lagos, Portugal (15th–17th centuries). *Int. J. Osteoarchaeol.* 28 (2), 162–169. doi:10.1002/oa.2643

Weitz, C. A., Garruto, R. M., and Ting, C. C. (2016). Larger FVC and FEV1 among Tibetans compared to han born and raised at high altitude. *Am. J. Phys. Anthropol.* 159 (2), 244–255. doi:10.1002/ajpa.22873

Whaley, A. L. (2020). Ethnicity, nativity, and the effects of stereotypes on cardiovascular health among people of african ancestry in the United States: Internal versus external sources of racism. *Ethn. Health* 0 (0), 1010–1030. doi:10.1080/13557858.2020.1847257

Wong, M., Bierman, Y., Curtis, P., Kittles, R., Mims, M., Jones, J., et al. (2019). Comparative analysis of P16 expression among african American and European American prostate cancer patients. *Prostate* 79 (11), 1274–1283. doi:10.1002/pros.23833

Yamasato, K., Chern, I., and Lee, M.-J. (2021). Racial/ethnic representation in United States and Australian obstetric research. *Matern Child Health J.* 25 (5), 841–848. doi:10.1007/s10995-020-03099-8

Yoshikawa, A., Jiang, L., and Meltzer, H. Y. (2020). A functional HTR1A polymorphism, Rs6295, predicts short-term response to lurasidone: Confirmation with meta-analysis of other antipsychotic drugs. *Pharmacogenomics J.* 20 (2), 260–270. doi:10.1038/s41397-019-0101-5

Yuan, J., Kensler, K. H., Hu, Z., Zhang, Y., Zhang, T., Jiang, J., et al. (2020). Integrative comparison of the genomic and transcriptomic landscape between prostate cancer patients of predominantly african or European genetic ancestry. *PLoS Genet.* 16 (2), 10086411–e1008726. doi:10.1371/journal.pgen.1008641

Zhao, J., Sun, J., Xia, Z., He, G., Yang, X., Guo, J., et al. (2020). Genetic substructure and admixture of Mongolians and Kazakhs inferred from genome-wide array genotyping. *Ann. Hum. Biol.* 47 (7–8), 620–628. doi:10.1080/03014460.2020.1837952