Genetic similarity and genetic ancestry groups

Graham Coop

Center for Population Biology and Department of Evolution and Ecology,

University of California, Davis

July 2022

I was asked to provide a commentary to the National Academies of Sciences, Engineering, and Medicine "Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research" committee. This piece expands on my remarks from that talk. As a population geneticist, I focus on the use of genetic analysis to provide sample descriptors for human genetic samples, in particular the use of "genetic ancestry groups".

While they come with a number of challenges, genetic sample descriptors are likely often unavoidable in practice, as scientists need them in order to combine and communicate findings. However, genetic sample labels are clearly a source of confusion, with slippage both in scope and between genetic versus social labels. One common genetic sample descriptor that researchers use is "genetic ancestry group": for example, labelling individuals living in the United States as having "European genetic ancestry" or "African genetic ancestry". I'll argue that these terms are imprecise and potentially misleading and that, for most applications, researchers simply mean genetic similarity or relatedness to some predefined set of samples. Given the issues associated with such labels, I believe that much of human genetics research should move away from using the term "genetic ancestry groups", and towards using more readily interpretable statements about genetic similarity (and relatedness) for sample descriptions. These statements are often nearly equivalent in terms of the information they contain, but simple statements about genetic similarity/relatedness are a more accurate statement of what population-genetic methods are providing, and importantly such language carries far less baggage in terms of its implicit depiction of the structure of human groups.

This is not a new argument; see Mathieson and Scally (2020) and Lewis *et al.* (2022) for recent discussions about the need to be more precise about what we mean by genetic ancestry. However, here I actively call for the field of human genetics to move away from using genetic ancestry groups as a sample descriptor. In most applications, human geneticists are actually concerned with controlling for genetic similarity, geography, and environments in their comparisons rather than some vague notion of ancestral populations. A number of subfields of human genetics, including human population genetics and genetic anthropology, are more directly concerned with understanding history through reconstructing various aspects of genetic ancestors; in my view, this is largely a related but distinct enterprise from providing the field of human genetics with useful genetic sample descriptors.

In what follows, I'll first begin with a few words on human genetic variation. I'll then turn towards why genetic sample descriptors will likely be necessary for some time to come. I'll discuss the topic of genetic ancestors and why in practice genetic ancestry description labels used in human genetics boil down to genetic similarity. I'll briefly touch on some of the issues and responses raised by genetic ancestry labels, before finishing with an outline of why genetic similarity labels are a better set of descriptors for the broad field of human genetics.

Genetic variation

Genetic populations are modelling constructs. As in other scientific endeavours, when developing mathematical models of evolution and statistical tools to analyze data, we have to simplify reality. In such models, we often assume that there are well-defined populations, which are modelled as persisting in particular areas, and that represent somewhat separate and well-defined evolutionary lineages. Such models can be incredibly useful in helping shape our intuition and analysis but can also constrain our thinking and lead to confusion.

Well-delimited genetic populations are rarely found within species in the real world. Certainly, such populations are very rare in humans. We can of course talk of "populations" consisting of everyone in the UK, Finland, or whole continents such as Africa, but these are obviously not populations in the sense that is typically meant in evolutionary genetics, in which individuals within a group share closer genealogical relationships to each other than any other individual outside the group. Even if we take subsets of people within these regions, e.g., "people who self-identify as White in the UK" or "people whose four grandparents all come from the UK", these groups will not form well delimited genetic subsets, as the UK is not a homogeneous population with respect to migration in and out of the UK (Leslie *et al.*, 2015). With sparser sampling of a limited number of geographic locations, we can sometimes identify groups that are closer to better defined "genetic populations", but these simply reflect the limitations of our sampling.

The reality is that we're all related to each other to varying extents, in a complex web of genealogical relations that form an unimaginably complicated family tree. As a result, genetic variation varies fairly smoothly among individuals, often in ways that are correlated with environments. Patterns of human genetic variation are shaped by geographic distance, geographical barriers, as well as broad-scale population movements. It might be tempting to think that the fairly continuous nature of modern human variation reflects admixture only over the past few hundred years. However, human groups of the past have often been geographically widespread and ephemeral, frequently forming, only to rapidly collapse together with other nearby and sometimes much more distant groups. This is a point that advances in ancient DNA technology have repeatedly made abundantly clear (Skoglund and Mathieson, 2018).

Much of the common genetic variation found within groups is shared across human groups (Lewontin, 1972; for a visualization see Biddanda *et al.*, 2020). Rarer genetic variation is usually more localized. In fact, individual rare variants are better thought of not as properties of groups that bear any resemblance to the ancestry groupings used in genetics research, but rather as features of extended families within such groups. Local adaptation

is sometimes highlighted as a reason that particular phenotypic outcomes might be common in some groups and rare in others (i.e. due to combinations of alleles varying substantially in frequencies across groups). Although a growing number of convincing cases of genetic variants shaped by local adaptation have come to light, these are still a tiny minority of the many loci associated with phenotypic and disease outcomes.

As a consequence of the complicated genetic structure of humanity, there's no single "right" level of granularity to use for description for all questions. Human groups are structured from broad geographic scales to fine-scale patterns well below the level of a country. These fine-scale patterns are generally not well captured by discrete labels. The choice of level of granularity will depend on the specific questions being addressed. For example, the accuracy of polygenic score predictions is lower for people of Italian and Polish ancestry in the UK than for people of United Kingdom ancestry (Privé *et al.*, 2022) (the relative contributions of changes in linkage disequilibrium, specific genetic, environmental, and gene-environment interactions to such changes in accuracy are under current investigation and debate). Whether the researcher thinks a set of polygenic scores are appropriate for use in "European ancestries" or that they need much more fine-grained GWAS will depend on the questions being pursued and the prediction resolution required by any future uses. Thus, the nature of the sample descriptions needed in this example, and in many other cases in human genetics, will depend on the problem at hand.

Given the complexity of human genetic variation, all verbal descriptions of genetic structure will be incomplete and open to miscommunication. While this has often been an issue in human genetics, this issue is coming to the fore as the depth of genetic sampling of humans increases. We need to move towards terminology for genetic sample descriptors that are clear to people across fields and encourage good use.

Why then use descriptors/labels in analysis?

If human genetic structure is relatively continuous why then do we need to use (somewhat) discrete sample descriptors? Lots of valid explanations and uses have been put forward during the presentations at the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research Meetings (2022a,b,c). From my perspective, there are two major reasons for using genetic sample descriptors that I want to highlight.

Data subsetting

Researchers set out to gather large enough samples that will offer sufficient statistical power to study their chosen problem, but in practice the sampling, genotyping, phenotyping, and analysis effort of a given project will always be limited. There are obviously many historical and ongoing reasons why sampling has focused on particular groups, but one technical issue is the limitations of statistical approaches and tools. Many methods in statistical and population genetics fit statistical models that assume relatively homogeneous groups. For example, standard GWAS rely on pseudo-randomization of genotypes at a locus across genomic backgrounds and environmental causes of trait variation, and so researchers often limit themselves to more genetically well-mixed groups to better satisfy these requirements.

On the population genomics side, methods to reconstruct population history are constrained to describing the history of a small number of groups, while methods to describe more realistically continuous histories are daunting, and are as yet are underdeveloped. Because of these limitations, researchers subset their data in various ways, for example restricting GWAS samples to a particular ethnicity and geographic region. But having genotyped the individuals, they often further subset their data to a particular subset of genetic variation, e.g., a particular subregion of principal components analysis (PCA) space where many of their samples are present. Having subset the samples, descriptors are needed for these genetic subsets ("who was the method applied to?").

Some of the limitations of methods reflect historical legacies of working in modelling frameworks designed to analyze smaller genetic datasets and previous computational limitations. Methods are improving, allowing GWAS to be performed across somewhat more heterogeneous samples and more flexible models of human population genetic history. However, conceptually, it is very hard to analyze even small subsets of human diversity all at once and so it is unlikely that subsetting genomic data will stop in practice any time soon. Thus, genetic sample descriptors will remain in use.

Communication

The second and perhaps bigger reason for genetic sample descriptors is that scientists often rely on them to communicate with each other. In many human genomic analyses, assessing patterns of genetic structure is the first step. In such applications, researchers use prior population descriptors to orient results: "we see patterns, what do they reflect?". For example, what features do the major axes of variation found in a principal components analysis correspond to? We need words to describe these axes and describe our interpretations. Furthermore, as a field we often pool together different datasets, e.g., in GWAS meta-analysis, or combine together different data types, e.g., GWAS effect sizes from one group with genotype data from another. In a practical sense, we need shorthand labels to communicate to each other what we are doing.

Genetic ancestry

One common genetic sample descriptor applied to samples in human genomics is that of "genetic ancestry group", with terms like "European genetic ancestry" or "East Asian genetic ancestry" frequently appearing in papers to describe the genetics of samples of individuals based on the analysis of their genotypes. "Genetic ancestry" is an evocative concept, as made clear by the success of personal genomics companies, which capitalize on the fact that our views on ancestry and identity long predate the development of genetics.

Genetic ancestors

I first want to make the distinction between two related concepts of "genetic ancestors" and "genetic ancestry group". The former is a well-defined concept and a central part of population genetics. Your genealogical ancestors, the people from whom you are biologically

descended, are a well-defined set of people. More than a few hundred years back, each of us have tens of thousands of genealogical ancestors, a number that initially grows exponentially backward in time until it stabilizes when you are descended from everyone who left any descendants to the present day. However, with only 2 copies of your genome, which is inherited in big blocks, you can't inherit genetic material from all of these ancestors. Instead, you inherit genetic material from only a very small subset of your ancestors (Donnelly, 1983; Coop, 2013). For example, ∼450 years ago you may have more than 32,000 living genealogical ancestors, but only ∼1000 of them contributed genetic material that ended up in your genome, and the proportion of ancestors who contribute genetic material drops farther back in time. The subset of your genealogical ancestors from whom you inherited genetic material are your "genetic ancestors", a small fraction of your total ancestry.

Several thousand years back, all modern humans share all of their genealogical ancestors. Farther back than that, anyone who left any descendants in the present day (and many did) is an ancestor to all humans living today (Manrubia *et al.*, 2003; Rohde *et al.*, 2004; Coop, 2017b). What does that mean for our genetic relatedness? Well, that's complicated. You and I share a genealogical common ancestors at least as recently as the time when all modern humans share all their ancestors. However, only a limited subset of these people are our genetic ancestors, and we only inherit small fractions of our genome from any one of these common ancestors. Therefore, at a typical locus, your and my most recent genetic common ancestor lived much farther back in the past.

Genetic similarity, genealogy, and genetic ancestry are closely related concepts, in ways that are non-trivial to understand (Coop, 2017a). All four of my grandparents came from Northern England. As a result, I share slightly more genetic variants in common with other people whose grandparents all came from Northern Europe than I shared with people whose grandparents came from elsewhere in the world. That's because, while everyone in the world share all of their genealogical common ancestors several thousand years ago, they are not weighted the same in different people's genetic family trees. A particular ancestor could appear tens or hundreds of times tracing back along different paths in my family tree if they are many generations back. That same ancestor many generations back could appear only once in someone else's family tree with only a single path to them. While they are an ancestor to both of us, I'm somewhat more likely to have inherited a (small) chunk of my genome from them than the other person is. Even though we all share many genealogical ancestors, I have many more paths back through my family tree to ancestors who are also ancestors many times over for other Northern Europeans than I do with someone from (say) Japan. As a result, I share slightly more genetic variants in common with another Northern European than I do with a person whose grandparents all came from Japan. My genetic resemblance to many Northern Europeans reflects the fact that we share somewhat more of our genealogical ancestry–but not in a way that maps simply onto statements that my ancestors are all European or that Europeans share some set of ancestors that people from elsewhere do not.

Our genetic ancestors are a well-defined set of people, who lived in particular places and times. Obviously, in practice, we don't know who they were, but we could hope to infer things about them. Population genomic data can inform us about our relatedness our shared genetic common ancestors, through summaries of genetic similarity among individuals, which reflects the sharing of genetic material transmitted through meiosis. A lot of the statistical

machinery of population genetics builds on these ideas to learn about evolutionary processes and history. However, much of our interpretation of these patterns comes from combining this information with geographic and sample descriptors of the analyzed genomic data. Thus, our interpretations and genetic sample labels always reflect, at least in part, the social context of how samples were chosen and described, and thus they are partially social constructs.

Through computational and statistical advances, the field of population genetics is getting much better at describing some properties of our vast number of shared genetic ancestors. A major recent breakthrough is the development of approaches to computationally reconstruct the so-called "ancestral recombination graph", or approximations of it, for large genomic samples (Speidel *et al.*, 2019; Kelleher *et al.*, 2019). The ancestral recombination graph describes the full set of genetic relationships among a set of samples in terms of their shared genetic ancestors (Hudson *et al.*, 1990; Marjoram and Griffiths, 1995). Along with this breakthrough has come the hope that approaches building on the ancestral recombination graph will allow a fuller description of "genetic ancestry". It is doubtlessly true that these advances will allow us a fuller picture of some of the properties of our vast clouds of genetic ancestors and underscore the point that we are all embedded in the same giant tree of humanity, something other methods can obscure. However, these representations will necessarily often be high dimensional and do not lend themselves easily to verbal summaries.

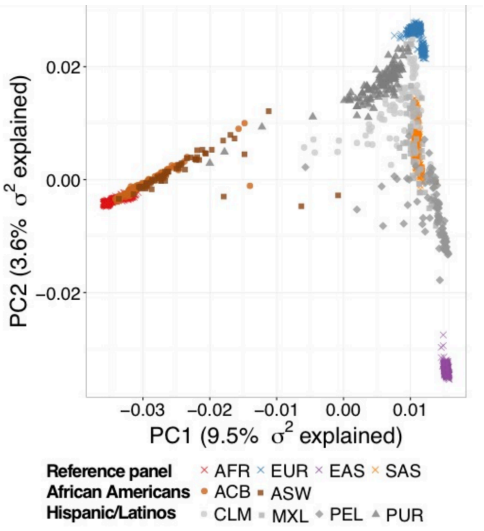

Figure 1: Figure from Martin *et al.* (2016) using 1000 genomes samples: "Principal components analysis of all samples showing the relative homogeneity of AFR, EUR, EAS, and SAS continental groups and continental mixture of admixed samples from the Americas (ACB, ASW, CLM, MXL, PEL, and PUR)." Cropped from Figure 1 in preprint (CC BY-NC 4.0), figure S1 in published paper. 1000 genome sample codes given at this link.

What then is a genetic ancestry group?

When human genetic researchers use ancestry group terms such as "European ancestry" or "East Asian ancestry" as a sample descriptor they are (nearly) always a description of genetic similarity to other present day individuals by some summary statistics (Mathieson and Scally, 2020). To illustrate this, we can work through some common ways in which ancestry groups are assigned in human genetics. In doing this, I note that I've certainly referred to these approaches in these terms in the recent past, and it can be a very convenient way to explain the concepts at play. My discussion of this topic is forward-looking rather than a judgment of past uses, including my own.

One of the most common ways that ancestry labels are assigned to samples is on the basis of how a person's genome clusters with other samples in a genetic PC plot. For example, Figure 1 shows the 1000 genomes samples positioned on their first two genetic principal components . If you projected my genome onto this plot, my genotype would doubtless cluster with the EUR samples.

On that basis, researchers might choose to label me as belonging to the "European ancestry group". However, the fact that I would fall close to samples labelled "European" is simply a statement that my genotype is similar to the genotypes of those people along the axes of variation captured by the top two PCs. It is also broadly a statement that I shared a higher degree of relatedness with these individuals along these axes (McVean, 2009) but it is not a statement that my genetic ancestors form a delimited group with other such individuals.

Conversely, imagine a person who comes with a sample descriptor "English", e.g., based on a self-identified label. If this person has 6 great-grandparents from England and 2 whose ancestors trace recently back to West Africa via the Caribbean, their genome might fall roughly 3/4s of the way between the African and European 1000 genomes panel samples on this plot. On that basis, researchers may choose to exclude them from the "European ancestry" group of individuals. Indeed, researchers usually predefine some set of cutoffs, e.g., if a person falls more than 3 standard deviations from the centroid of the cluster of individuals labelled "European" individuals then they are not retained in the European ancestry data subset. As discussed above, there can be good methodological reasons for wanting a relatively genetic homogeneous group for analysis. However, the decision of whom to include is necessarily a somewhat arbitrary exercise in applying discrete labels to continuous variation.

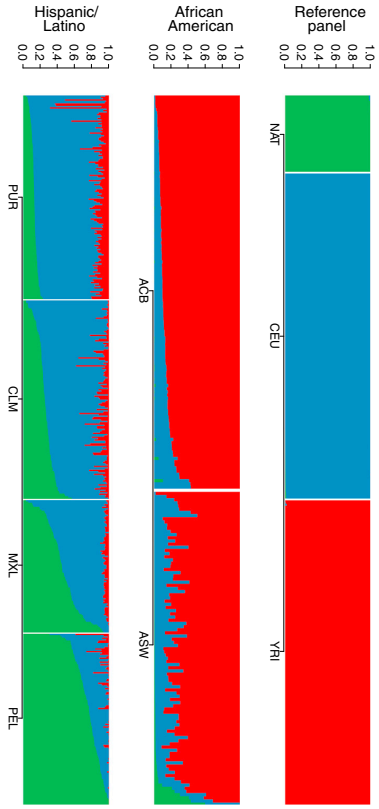

Figure 2: Figure from Martin *et al.* (2016) from their caption "ADMIXTURE analysis at K = 3 focusing on admixed Americas samples, with the NAT, CEU, and YRI as reference populations." They find that "[t]he six populations from the Americas demonstrate considerable continental admixture, with genetic ancestry primarily from Europe, Africa, and the Americas". NAT is a sample of Native Americans from Mao *et al.* (2007). Figure cropped from Figure 1 of preprint (CC BY-NC 4.0).

Other methods of assigning ancestry groups allow for a person's ancestry to be drawn from multiple "ancestral" groups (`STRUCTURE` and `ADMIXTURE` style approaches Pritchard *et al.*, 2000; Falush *et al.*, 2003; Alexander *et al.*, 2009); see Figure 2 for an example. These methods would allow a breakdown, for example, of someone's ancestry into 75% European and 25% African ancestry. In this case, alleles are modelled as being drawn from some hypothetical well-mixed populations; however, the "genetic ancestry" labels for those populations are being propagated by investigators from other labelled samples (eg a panel of references samples). Again, they are really statements about similarity: that 75% of the genome is

most similar to the genotypes of individuals from the CEU samples (CEPH European descent sample, Utah) and 25% of their genotype looks similar to the genotypes from the YRI sample (Yoruba in Ibadan, Nigeria).

Various other finer grain methods and representations also exist. For example, there are approaches where blocks of an individual's genome (haplotypes) are statistically assigned as coming from some limited set of (often) pre-specified ancestries (Falush *et al.*, 2003; Price *et al.*, 2009; Maples *et al.*, 2013), defined by representative samples (see Figure 3). Again, however, regions of the genome labelled as having African ancestry could simply be said to be more similar to YRI samples than to CEU samples.

Thus, all these statements about ancestry groups are really statements about the genetic similarity to other samples whose population descriptors we choose to propagate as our "ancestry group" labels. Samples are genetically similar because they share more ancestral paths, and so all of these approaches can be phrased in terms of hypotheses about shared genetic ancestors. Framed in that way they can quite be a reasonable set of analyses and hypotheses for future investigation to learn about population history. However, in most applications ancestry group labels are used as simple sample descriptors, which raises a large set of issues.

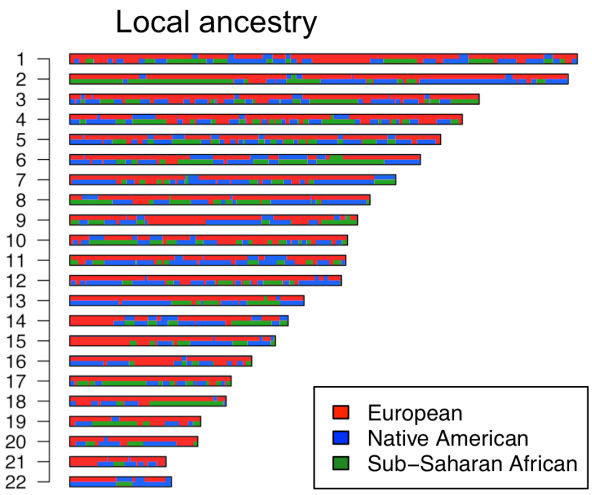

Figure 3: Figure from Moreno-Estrada *et al.* (2013) the genome of a Caribbean-descent individual recruited in South Florida, USA. The 22 automosomes are coloured by "Continental-level local ancestry calls" (Moreno-Estrada *et al.*, 2013) using an updated version of PCAdmix (Brisbin *et al.*, 2012). Reference samples used: YRI, CEU, and Native Americans from Mexico.

Some issues raised by the use of genetic ancestry group labels

A wide variety of issues have been repeatedly raised over the use of genetic ancestry labels. A common and fair criticism is that the genetic ancestry descriptors used are often based on continental geographic labels and so overlap with racial labels (Fujimura and Rajagopalan, 2011; Panofsky and Bliss, 2017; Lewis *et al.*, 2022). The ancestral populations alluded to by genetic ancestry labels are at most simple statistical modelling constructs, but they can easily become reified into a discrete view of ancestral populations. While this problem is most apparent in the deliberate misuse of genetic ancestry to reinforce racist narratives of human diversity, it also opens a number of pitfalls for researchers themselves. Primarily because such labels obscure the inhomogeneity within "ancestry groups" and the continuum of relatedness across them and in doing so can bias our thinking about genetic and environmental variation. For

example, in GWAS individuals of "European genetic ancestry" are often grouped together for analyses across multiple countries. While there can be good methodological justifications for grouping samples, there are also implicit decisions about who shares enough similarity in genetics and environments to be grouped together. A number of technical issues can also be raised about the use of genetic ancestry labels. For example, even if we accept the idea of using ancestry groups based on genetic similarity, the resolution and naming of ancestry groups is a function of reference samples used. For example, by some methods discussed above a person living in the middle east might find themselves assigned as having "European" ancestry if middle Eastern samples are not included in the reference set. As a result of this, changes to the panels used to assess ancestry groups can also lead to confusing changes in ancestry labels, the most obvious setting where this occurs is in personal genomics but in practice it can also occur as datasets are reused across scientific papers. Nor will these issues be resolved by including finer scale sampling in reference panels, as in the limit, there will often be a fairly continuous spread of peoples' genotypes, and so there is no natural place to carve human diversity to assign "ancestries".

Another important aspect is the time frame implicitly assumed by descriptions of ancestry groups. Indeed, although statements of ancestry usually bracket genetic ancestors at a specific time period, this aspect is usually missing from the descriptions in papers. For example, in analyses of samples in the Americas, many people will have generations of recent ancestors who also lived in the Americas. However, through the choice of ancestry reference the analysis of genetic ancestry is often implicitly targeted at describing the locations of the genetic ancestors of American people >600 years in the past. Moreover, it is now clear that there have been large-scale movements of people over the past ten thousand years, which make labels based on current-day sampling locations complicated. For example, my genetic ancestors likely lived in Europe, the middle east, and the Russian Steppe ten thousand years ago (Lazaridis *et al.*, 2014; Haak *et al.*, 2015). So these statements of genetic ancestry are best thought of as descriptions of genetic ancestors in a bracketed time period of 600 years ago to a few thousand years ago. Quite why the field has settled on this time period as a basis for comparison is often unclear.

Responses

Partially in response to some of these criticisms, a number of alternative approaches have been put forward for population descriptors based on ancestry. One idea might simply be to switch from using terms like "European ancestry" to "European ancestries" to avoid implying that there is one homogeneous ancestral population underlying patterns of similarity. This proposal seems reasonable if we are to keep using ancestry group labels but does not move away from using broad geographic labels. A second, more substantial proposal is to move towards fine-grained ancestry labels, partially motivated by the reasonable objective of moving away from continental labels. For example, 23&me now breaks down my ancestry into Northwestern European, British and Irish ancestry, and to even fine-grainer "ancestries" like "Greater London". This is based on the similarity of my haplotypes to other present-day people whose grandparents come from these regions (Durand *et al.*, 2021). These fine-scale approaches are also being used to examine ancestry in medical settings (Belbin *et al.*, 2021). However, if we have trouble defining what we mean precisely by ancestry at broad geographic

scales, we will run into even more difficulties being precise about what we mean by phrases like "British and Irish ancestry," let alone "Greater London" ancestry.

Other approaches to ancestry have been laid out where a person's ancestry could be reported for various different time epochs, which would remove the need for the implicit choice of time-period. Or we could imagine tracing genetic ancestors back across geographical space over the generations, which would allow a more continuous view of ancestry (Osmond and Coop, 2021; Wohns *et al.*, 2022). Such approaches may well be useful for population geneticists and genetic anthropologists interested in human history. Indeed, advances in population genetics and genetic anthropology combined with ancient DNA are making large inroads into describing human history, and all these fields should work to integrate a better set of descriptors of genetic ancestors. However, the vast amount of research in human genetics (notably that funded by the NIH) is not about selling personal genomics or studying human history.

An additional justification often offered for the use of genetic ancestry is that in addition to genetics it also captures socio-environmental factors that can covary with ancestry. On this basis in many cases ancestry group labels are used to subset data or as a covariate in analyses to capture non-genetic effects. However, in doing so, there is an issue of mixing social, geographic, and genetic labels. For example, does a set of "East Asian ancestry" polygenic scores refer to people of East Asian ancestry in the US or from the Japanese biobank and why is East Asian the correct ancestry label for what may be a narrower subset of people? Such distinctions appear already to be important in a number of cases (Giannakopoulou *et al.*, 2021). Yet too often, studies center the genetic ancestry label and so confound together relatedness, social environment, and sampling location. This conflation sets up a situation where it is all too easy to slip into viewing differences in genetic ancestry as a genetic cause of differences in phenotypic and health outcomes between groups.

Along similar lines, within racial or ethnic groupings the proportion of an individual's genome from particular "ancestries" can be correlated with phenotypic outcomes due to environmental factors as well as generations of racism and discrimination. For example, in African Americans in the US the proportion of African ancestry is correlated with geography, socioeconomic outcomes, and patterns of migration out of the American south (Micheletti *et al.*, 2020; Baharian *et al.*, 2016). But all too often papers reporting correlations with genetic ancestry in recently admixed populations do not acknowledge the potential for socio-environmental confounding and can slip straight to discussions of genetic causes.

Human genetics should move away from the concept of "genetic ancestry groups" and towards "genetic similarity" descriptors.

Arguably, much of human genetics cares about matching for genetic similarity, not ancestry, when making a set of comparisons or assembling a set of controls. Consider the following commonly posed questions: "What is the disease risk of someone of this specific genotype with this broader genetic background?", "Are my biological replicates and controls appropriately matched for this gene expression study?", or "What set of polygenic score weights

should I use for prediction for this person or haplotype?". In all of these cases, the genetics questions we are asking about are based on matching people based on genetic similarity, not a vague sense of who a person's ancestors were.

Researchers are also often interested in controlling for the environment, and given that in many cases the relevant environmental variables may be unknown or unmeasured, it may be reasonable to use non-genetic sample descriptors in these analyses as a control for unmeasured environmental factors. In some cases, a researcher may lack socially relevant sample descriptors and so want to use a genetic label as a proxy for a social label to account for unmeasured environmental factors. Indeed, whenever we include a genetic similarity variable (e.g., a principal component) in an analysis of traits, it may become a proxy for both genetic and correlated environments and social variables. In all such cases, we should be clear that correlates between trait outcomes and genetic similarity could be the result of both genetic and environmental causes rather than relying on the idea of "genetic ancestry" to telegraph that idea.

As a field we should move away from genetic ancestry labels and towards simple statements of genetic similarity: "This sample/haplotype is genetically similar to the XX sample set (in comparisons to YYY samples using ZZZ metric)" is much closer to how population genetic methods can be used to provide genetic sample descriptors. For example, "Graham is genetically similar to the GBR 1000 genome samples (on the first 10 principal components)" rather than "Graham has Northwestern European genetic ancestry". The former sounds a little more awkward, but that awkwardness reflects the truth of how these labels work and comes with many fewer built-in assumptions and pitfalls.

From the technical standpoint, moving toward using genetic similarity labels puts a focus on how similar we need to make the match and by what measure we judge similarity. It also directs attention to the panels used to judge similarity and forces us to ask ourselves whether our panels are representative and fine-grained enough for the comparison we wish to make.

Importantly, in my view, the term "genetically similar to" also helps to avoid the assumption of homogeneity within labels; "similar to" does not imply "same as". Similarity-based sample descriptors also move us some way to acknowledging the continuous nature of genetic variation across human groups in our sample descriptions. I am more genetically similar to some samples than I am to others but that does not imply that there are natural groupings. Nor do similarity-based labels imply how I, as an individual, might choose to identify or what distribution of environments I might experience. For example, a person may be genetically similar to Southern Asian 1000 genomes samples, yet in itself this similarity does not identify them as Southern Asian, whereas stating that a person has South Asian genetic ancestry comes much closer to making that linkage in people's minds.

Working out the genealogical history of individuals and of groups of people from around the world is a fascinating area of research, but it should not be the day job of the majority of researchers in the field of human genetics, who simply need appropriate sample descriptors.

Acknowledgments

Thanks to Vince Buffalo, Doc Edge, Jeff Groh, Emily Josephs, James Kitchens, Peter Ralph, Alexis Simon, and Silu Wang for comments on a earlier draft.

References

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome research *19*(9): 1655–1664.

Baharian, S., M. Barakatt, C. R. Gignoux, S. Shringarpure, J. Errington, W. J. Blot, C. D. Bustamante, E. E. Kenny, S. M. Williams, M. C. Aldrich, and others, 2016 The great migration and African-American genomic diversity. PLoS genetics *12*(5): e1006059.

Belbin, G. M., S. Cullina, S. Wenric, E. R. Soper, B. S. Glicksberg, D. Torre, A. Moscati, G. L. Wojcik, R. Shemirani, N. D. Beckmann, and others, 2021 Toward a fine-scale population health monitoring system. Cell *184*(8): 2068–2083.

Biddanda, A., D. P. Rice, and J. Novembre, 2020 A variant-centric perspective on geographic patterns of human allele frequency variation. Elife **9:** e60107.

Brisbin, A., K. Bryc, J. Byrnes, F. Zakharia, L. Omberg, J. Degenhardt, A. Reynolds, H. Ostrer, J. G. Mezey, and C. D. Bustamante, 2012 PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Human biology *84*(4): 343.

Coop, G., 2013 How many genetic ancestors do I have? `https://gcbias.org/2013/11/11/how-does-your-number-of-genetic-ancestors-grow-back-over-time/`.

Coop, G., 2017a Where did your genetic ancestors come from? `https://gcbias.org/2017/12/19/1628/`.

Coop, G., 2017b Your ancestors lived all over the world. `https://gcbias.org/2017/11/28/your-ancestors-lived-all-over-the-world/`.

Donnelly, K. P., 1983 The probability that related individuals share some section of genome identical by descent. Theoretical population biology *23*(1): 34–63.

Durand, E. Y., C. B. Do, P. R. Wilton, J. L. Mountain, A. Auton, G. D. Poznik, and J. M. Macpherson, 2021 A scalable pipeline for local ancestry inference using tens of thousands of reference haplotypes. bioRxiv.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics *164*(4): 1567–1587.

Fujimura, J. H. and R. Rajagopalan, 2011 Different differences: The use of 'genetic ancestry' versus race in biomedical human genetic research. Social Studies of Science *41*(1): 5–30.

Giannakopoulou, O., K. Lin, X. Meng, M.-H. Su, P.-H. Kuo, R. E. Peterson, S. Awasthi, A. Moscati, J. R. Coleman, N. Bass, and others, 2021 The genetic architecture of depression in individuals of East Asian ancestry: a genome-wide association study. JAMA psychiatry *78*(11): 1258–1269.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, and others, 2015 Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*(7555): 207–211.

Hudson, R. R. and others, 1990 Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology *7*(1): 44.

Kelleher, J., Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean, 2019 Inferring whole-genome histories in large population datasets. Nature Genetics *51*(9): 1330–1338.

Lazaridis, I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, and others, 2014 Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*(7518): 409–413.

Leslie, S., B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E. C. Royrvik, B. Cunliffe, D. J. Lawson, and others, 2015 The fine-scale genetic structure of the British population. Nature *519*(7543): 309–314.

Lewis, A. C., S. J. Molina, P. S. Appelbaum, B. Dauda, A. Di Rienzo, A. Fuentes, S. M. Fullerton, N. Garrison, N. Ghosh, E. M. Hammonds, and others, 2022 Getting genetic ancestry right for science and society. Science *376*(6590): 250–252.

Lewontin, R. C., 1972 The Apportionment of Human Diversity. In T. Dobzhansky, M. K. Hecht, and W. C. Steere (Eds.), *Evolutionary Biology*, Chapter 14, pp. 381–398. New York: Appleton-Century-Crofts.

Manrubia, S. C., B. H. Derrida, and D. H. Zanette, 2003 Genealogy in the Era of Genomics. American Scientist *91*(2): 158–165.

Mao, X., A. W. Bigham, R. Mei, G. Gutierrez, K. M. Weiss, T. D. Brutsaert, F. Leon-Velarde, L. G. Moore, E. Vargas, P. M. McKeigue, and others, 2007 A genomewide admixture mapping panel for Hispanic/Latino populations. The American Journal of Human Genetics *80*(6): 1171–1178.

Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics *93*(2): 278–288.

Marjoram, P. and R. Griffiths, 1995 *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications.*, Chapter An ancestral recombination graph, pp. 257–270.

Martin, A. R., C. R. Gignoux, R. K. Walters, G. L. Wojcik, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny, 2016 Population genetic history and polygenic risk biases in 1000 Genomes populations. bioRxiv: 070797.

Mathieson, I. and A. Scally, 2020 What is ancestry? PLoS Genetics *16*(3): e1008624.

McVean, G., 2009 A genealogical interpretation of principal components analysis. PLoS genetics *5*(10): e1000686.

Micheletti, S. J., K. Bryc, S. G. A. Esselmann, W. A. Freyman, M. E. Moreno, G. D. Poznik, A. J. Shastri, M. Agee, S. Aslibekyan, A. Auton, and others, 2020 Genetic consequences of the transatlantic slave trade in the Americas. The American Journal of Human Genetics *107*(2): 265–277.

Moreno-Estrada, A., S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes, C. R. Gignoux, P. A. Ortiz-Tello, R. J. Martínez, D. J. Hedges, R. W. Morris, and others, 2013 Reconstructing the population genetic history of the Caribbean. PLoS genetics *9*(11): e1003925.

National Academies of Sciences, Engineering, and Medicine, 2022 Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. `https://www.nationalacademies.org/our-work/use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research`.

Osmond, M. M. and G. Coop, 2021 Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. bioRxiv.

Panofsky, A. and C. Bliss, 2017 Ambiguity and scientific authority: population classification in genomic science. American Sociological Review *82*(1): 59–87.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS genetics *5*(6): e1000519.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics *155*(2): 945–959.

Privé, F., H. Aschard, S. Carmi, L. Folkersen, C. Hoggart, P. F. O'Reilly, and B. J. Vilhjálmsson, 2022 Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. The American Journal of Human Genetics *109*(1): 12–23.

Rohde, D. L., S. Olson, and J. T. Chang, 2004 Modelling the recent common ancestry of all living humans. Nature *431*(7008): 562–566.

Skoglund, P. and I. Mathieson, 2018 Ancient genomics of modern humans: the first decade. Annu. Rev. Genomics Hum. Genet **19:** 381–404.

Speidel, L., M. Forest, S. Shi, and S. R. Myers, 2019 A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics *51*(9): 1321–1329.

Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research Meetings, 2022a Public Workshop 1. https://www.nationalacademies.org/event/02-14-2022/committee-on-use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research-meeting-1.

Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research Meetings, 2022b Public Workshop 2. https://www.nationalacademies.org/event/04-04-2022/committee-on-use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research-meeting-2-and-public-workshop.

Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research Meetings, 2022c Public Workshop 3. https://www.nationalacademies.org/event/06-14-2022/use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research-meeting-3-and-public-workshop.

Wohns, A. W., Y. Wong, B. Jeffery, A. Akbari, S. Mallick, R. Pinhasi, N. Patterson, D. Reich, J. Kelleher, and G. McVean, 2022 A unified genealogy of modern and ancient genomes. Science *375*(6583)**:** eabi8264.