Received: 8 March 2022 | Revised: 31 May 2022 | Accepted: 6 June 2022

DOI: 10.1002/gepi.22492

PERSPECTIVE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

# Including diverse and admixed populations in genetic epidemiology research

Amke Caliebe[1] | Fasil Tekola-Ayele[2] | Burcu F. Darst[3,4] | Xuexia Wang[5] | Yeunjoo E. Song[6] | Jiang Gui[7] | Ronnie A. Sebro[8] | David J. Balding[9] | Mohamad Saad[10,11] | Marie-Pierre Dubé[12,13] | On behalf of the IGES ELSI Committee

[1]Institute of Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Kiel, Germany

[2]Epidemiology Branch, Division of Population Health Research, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, USA

[3]Center for Genetic Epidemiology, University of Southern California, Los Angeles, California, USA

[4]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[5]Department of Mathematics, University of North Texas, Denton, Texas, USA

[6]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA

[7]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, One Medical Center Dr., Lebanon, New Hampshire, USA

[8]Department of Radiology, Mayo Clinic, Jacksonville, Florida, USA

[9]Melbourne Integrative Genomics, Schools of BioSciences and of Mathematics & Statistics, University of Melbourne, Melbourne, Australia

[10]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

[11]Neuroscience Research Center, Faculty of Medical Sciences, Lebanese University, Beirut, Lebanon

[12]Department of Medicine, and Social and Preventive Medicine, Université de Montréal, Montréal, Québec, Canada

[13]Beaulieu-Saucier Pharmacogenomcis Centre, Montreal Heart Institute, Montreal, Canada

**Correspondence**
Marie-Pierre Dubé, Montreal Heart Institute and Université de Montréal, Montréal, QC, H1T1C8, Canada,
Email: marie-pierre.dube@umontreal.ca

**Funding information**
A.C. is funded by the German Research Foundation, Grant/Award Number: FOR2488; F.T-A. is funded by the intramural program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health; B.F.D. is funded by the National Cancer Institute of the National Institutes of Health, Grant/Award Number: K99CA246063; B.F.D is funded by the

**Abstract**

The inclusion of ancestrally diverse participants in genetic studies can lead to new discoveries and is important to ensure equitable health care benefit from research advances. Here, members of the Ethical, Legal, Social, Implications (ELSI) committee of the International Genetic Epidemiology Society (IGES) offer perspectives on methods and analysis tools for the conduct of inclusive genetic epidemiology research, with a focus on admixed and ancestrally diverse populations in support of reproducible research practices. We emphasize the importance of distinguishing socially defined population categorizations from genetic ancestry in the design, analysis, reporting, and interpretation of genetic epidemiology research findings. Finally, we discuss the current state of

Amke Caliebe and Fasil Tekola-Ayele contributed equally to this study.

Achievement Rewards for College Scientists Foundation Los Angeles Founder Chapter; X. W. is supported by the University of North Texas Foundation and Dr. Linda Truitt Creagh; R.S was funded by NIMH, Grant/Award Number: R21MH120739; D.B. is supported by the Australian Research Council under, Grant/Award Number: DP190103188; M-P.D. is funded by the Canada Research Chairs Program and by the Health Collaboration Acceleration Fund from the Government of Quebec

genomic resources used in genetic association studies, functional interpretation, and clinical and public health translation of genomic findings with respect to diverse populations.

KEYWORDS

admixture, diversity, genetic association, genome-wide association study, inclusion

### BOX 1 DEFINITION OF TERMINOLOGIES USED IN THE MANUSCRIPT

**Additive genetic effect**: the total effect, either over alleles at a locus or over loci, equals the sum of the effects of the individual alleles or loci.

**Admixture mapping**: a gene mapping approach for phenotypes that exhibit ancestry differences used to identify associations between local ancestry and the phenotype in a sample of admixed individuals.

**Causal inference**: a process to study the causal effect of a particular factor on an outcome of interest based on data by relying on assumptions, study designs, and estimation strategies.

**Haplotype phasing**: the process of statistically inferring an individual's
genome into maternally and paternally inherited segments.

**Heritability**: a measure of the extent to which genetic variation explains phenotype differences.

**Mendelian randomization**: a causal inference approach that uses genetic variation as a natural experiment to infer causal relationship between modifiable risk factors and health outcomes.

**Meta analysis**: integrated analysis of data from different studies to answer a research question.

**Mixed models**: statistical models that include fixed effects and random effects for which the variance matrix is specified. In GWAS, the effect of a locus of interest is often modelled as fixed and is the target of inference, while the genome-wide additive genetic contribution to the phenotype is modelled as a random effect, with the variance specified by a measure of relatedness.

**Population bottleneck**: a short period of low population size during the history of a population that leads to reduced genetic variation.

**Principal component analysis**: a dimension-reduction technique that replaces a large number of correlated variables with uncorrelated linear combinations of the original variables. In genetic studies, the first few principal components typically reflect population structure.

**QTL mapping**: a method to identify genetic loci that are associated with a quantitative trait such as height or a measure of gene expression.

**Structural variants**: refers to a broad range of genome variations including insertion, deletion, inversion, duplication, translocation, and copy number variation.

## 1 | INTRODUCTION

The field of genetic epidemiology is relatively young, but it has grown rapidly alongside the accelerating technological advances in genomics. The number of genome-wide association studies (GWAS) published has grown 40-fold in the past decade, with over 5000 unique publications and now nearly 300,000 associations reported in the GWAS catalog (Buniello et al., 2019). However, the body of research in genetic epidemiology is limited by an under-representation of non-European populations (All of Us Research Program Investigators et al., 2019; Landry et al., 2018; Popejoy & Fullerton, 2016). From a social perspective, filling this gap is essential to ensure equitable health care worldwide (Popejoy et al., 2018) and from a health research perspective, embracing ancestral diversity in genetic studies offers more opportunities to improve our understanding of the etiology of diseases (Peterson et al., 2019; Rosenberg et al., 2010). The study of population differences in disease burden and disease-associated genetic variants can help identify functional genetic variants and any underlying evolutionary factors. One example is X-linked dystonia parkinsonism caused by a founder mutation originating from Panay Island in the Philippines (Lee et al., 2001). Another example identified from studying admixed populations is kidney disease due to *APOL1* variants which may have risen to high frequency in parts of Africa to provide protection against African sleeping sickness (Yusuf et al., 2021). Genetic analyses of admixed populations can provide novel evolutionary insights about demographic events giving rise to phenotypic diversity (Skoglund & Mathieson, 2018) and the origin of traits (Xu et al., 2017). Linkage-disequilibrium (LD) differences across populations are useful to refine the resolution of GWAS signals to a smaller number of potential causal variants (Helgason et al., 2007; Schaid et al., 2018; Wojcik et al., 2019) and can be leveraged to improve the accuracy of genotype imputation (Wojcik et al., 2018). Increasing ancestral diversity in genetic studies is intimately tied to advancing our knowledge of genetic and environmental factors and their potential interactions.

Addressing the existing under-representation of non-European populations in genetic epidemiology requires concerted efforts to conduct inclusive research, including diversifying recruitment, advancing data analysis approaches, and improving genomics resources necessary to support research based on diverse and admixed populations. Even the best analysis methods cannot outweigh a scarcity of data, and only sufficient and high-quality data can lead to meaningful results (Buniello et al., 2019). Increasing the participation of underrepresented communities in these efforts can increase diversity and help prioritize research topics that are most relevant to individual communities, while addressing the challenges of developing frameworks for research consent, data sharing mechanisms and the return of results that respond to the communities' needs (Hindorff et al., 2018). Lessons on these dynamics can be drawn from previous successful examples of recruiting underrepresented populations in low- and middle-income settings and minorities in the United States (Horowitz et al., 2017; International HapMap Consortium, 2004; Tekola et al., 2009a). Regarding analysis methods, it is crucial that they are sufficiently flexible to enable the inclusion of diverse genomes. This is an area of active development, improving upon existing genome-wide analytic approaches, such as mixed models, meta-analyses, and admixture mapping Box 1 (Peterson et al., 2019). To improve our genomics and bioinformatics infrastructure, better reference populations are necessary, as the accuracy of haplotype phasing and genotype imputation heavily relies on these resources (Kowalski et al., 2019). Working toward this goal, more diverse genomics resources have recently been developed, and researchers are encouraged to consider those that appropriately fit their study population.

In the past year, as members of the Ethical, Legal and Social Implications (ELSI) committee of the International Genetic Epidemiology Society (IGES), we have worked together to provide guidance and support to our community for the conduct of research that is more inclusive. We

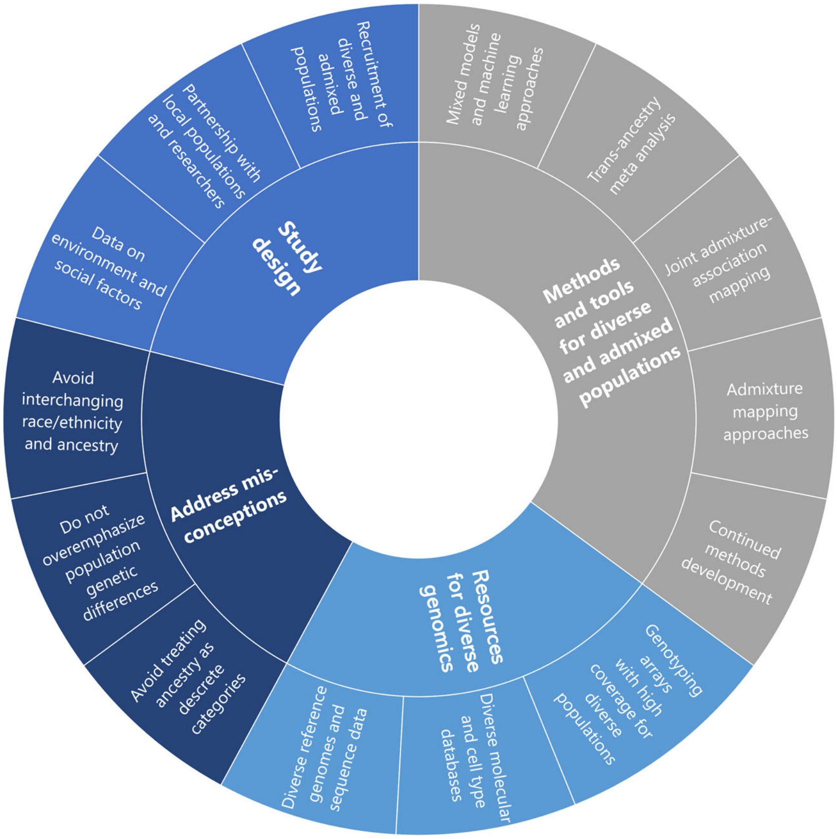
Study design
Recruitment of diverse and admixed populations
Partnership with local populations and researchers
Data on environment and social factors
Methods and tools for diverse and admixed populations
Mixed models and machine learning approaches
Trans-ancestry meta analysis
Joint admixture-association mapping
Admixture mapping approaches
Continued methods development
Resources for diverse genomics
Genotyping arrays with high coverage for diverse populations
Diverse molecular and cell type databases
Diverse reference genomes and sequence data
Address mis-conceptions
Avoid interchanging race/ethnicity and ancestry
Do not overemphasize population genetic differences
Avoid treating ancestry as descrete categories

**FIGURE 1** Key elements to address diversity in genetic epidemiology research

released a short guideline as a website statement on the use and reporting of race/ethnicity/ancestry in IGES abstracts and papers to promote inclusion diversity and equity (IGES ELSI Committee, 2021). Further building on that effort, we here offer perspectives on methods, analysis tools, and applications that enable the conduct of inclusive genetic epidemiology research (Figure 1), including admixed and ancestrally diverse populations.

## 2 | BACKGROUND AND BASICS

### 2.1 | What is captured by genetic ancestry versus race/ethnicity

Here, we define *genetic ancestry* as referring to an individual's biological ancestors. Genetic ancestry can refer to ancestral groups based on the continent of residence from the majority of ancestors over many generations (e.g., African, Asian or European ancestry), or it can be more specific (e.g., West-Central African, East Asian, or Northern European ancestry). All individuals have ancestors originating from multiple different locations at different time depth; as such, it can be more accurate to refer to proportions of genetic ancestries (e.g., 20% African ancestry) (Baker et al., 2017; Shriner et al., 2014). Information on genetic ancestry is sometimes sought from questionnaires for example by asking for country of origin of ancestors. However, self-reported ancestry is likely to be imprecise, as study subjects may vary in the knowledge of their genetic ancestry and can choose to describe it at various levels and using differing terms (Risch et al., 2002). A variety of clustering techniques described below can be used to infer ancestry based on study participants' genetic data, which can be more informative than self-reported ancestry.

We will use the term *ethnicity* to refer to a social group identity that is based on shared characteristics, such as cultural traditions, ancestry, language, religion, or social experiences. Ethnicity may vary according to geographical regions and it can traverse different social classes (Peoples & Bailey, 2018). For example, French-Canadians represent a subgroup of the Canadian society who share a common language and culture. Individuals living in the United States with measured proportions of African ancestry as low as 1% identify themselves as African American (Bryc et al., 2015; Kodaman et al., 2013). Further, some individuals from Africa are more genetically different from many Africans than from many Europeans due to the high genetic diversity within Africa (Gurdasani et al., 2015; Tishkoff et al., 2009; Yu et al., 2002). The shared social experiences of an ethnic group can have health impacts over time. As a consequence of systemic racism and segregation, Hispanic/Latino and African American individuals may be exposed to higher levels of stressors that are of economic, psychosocial, physical and environmental origin (Sternthal et al., 2011; Williams et al., 2019). To the extent possible, specific social determinants should be directly measured and accounted for in scientific enquiries, while ethnicity can be used to loosely capture other unaccounted social factors impacting health. For example, atopic dermatitis is more common among African American children, however, no ancestry-related genetic effects could explain the increased disease prevalence in this population, suggesting that underlying social factors captured by self-reported ethnicity may be responsible for the difference (Dikilitas et al., 2020).

The term *race* has often been used to refer to biological differences between human groups, allegedly underpinned by genetic differences (Borrell et al., 2021; Yudell et al., 2016). But it is now understood that human variation is complex, with no generally accepted classification system, and that race is largely a social construct similar to ethnicity that may or may not be correlated with biological traits. Care must be taken when using the term race in genetic epidemiology, as it can lead to an overemphasis on biological differences between groups of individuals. In particular we should avoid using controversial terms such as “Caucasian,” originating from the 18th century in support of racist arguments (Birney et al., 2021; Race, Ethnicity, and Genetics Working Group, 2005). Genetic epidemiology investigations can be easily misinterpreted or co-opted to promote messages of biological determinism or inferiority, and it is our responsibility as scientists to ensure that we communicate findings in a way that avoids such misinterpretations. Genetic differences between groups may be of interest in population genetics research or in genetic epidemiology research of diverse and admixed populations. For many research questions, however, the reasons for considering ethnicity, race or ancestry are to adjust for non-genetic effects and ensure that biological interpretations and public health interventions are applicable across broad populations (Birney et al., 2021). The reasons and intent for the inclusion of ethnicity, race or ancestry must be stated clearly in scientific reports. It is also important to recognize that race/ethnicity may not adequately capture lived experiences of individuals that vary within a population group. Efforts should be strengthened to standardize psychological, demographic, and socioeconomic information, which can be integrated with genomics to ensure the conduct of more accurate and informative research. As we continue to progress toward the conduct of more inclusive and diverse studies in genetic epidemiology, we should strive to conduct health research based on the direct measurements of health factors that will take the science beyond notions of race (Borrell et al., 2021; Yudell et al., 2016).

## 2.2 | Forces that shape worldwide genetic diversity in human populations

The genetic diversity that exists worldwide today results from the interplay of population genetic forces, mainly mutation, recombination, genetic drift, selection, and migration. Underlying all genetic variability are *mutations* that create new alleles in individuals. Single-site substitutions occur at a rate of about $10^{-8}$ per base pair and per generation, creating single nucleotide variants (SNVs), which are commonly used in genetic epidemiology research. However, this mutation rate is highly variable and can depend on genetic and epigenetic context, sex, and age of the father at the time of conception, among other factors (Segurel et al., 2014).

*Recombination* occurs during meiosis, when crossing over between homologous chromosomes generates new chromosomal combinations that are passed on to the next generation, breaking down the LD between loci. The recombination rate varies across the genome and there appears to be subtle differences in recombination rates between populations (Graffelman et al., 2007; Penalba & Wolf, 2020; Wegmann et al., 2011).

*Genetic drift* is the variation in allele frequencies due to stochastic sampling of alleles across generations. This is relevant for genetic epidemiology studies in populations that are currently small and isolated, or larger populations that have grown rapidly from a small size, such as the out-of-Africa bottleneck experienced by the ancestors of non-African populations. Whereas genetic drift usually shifts allele distributions over a relatively long period of time, *selection* can do so more rapidly. Positive and negative selection can quickly cause fixation or removal of alleles, whereas balancing selection maintains the polymorphism. Selection can act differently in different populations due to varying environments, resulting in diverse allele frequencies between populations. Because selection typically acts through differential reproductive success, it has less of an impact on diseases with later age at onset, such as Parkinson's disease, or Alzheimer's disease.

Well-known examples of traits where selection played a major role include skin color, lactase persistence and sickle cell trait. It should be noted that such traits are not good indicators of the genetic ancestry of an individual. For example, an individual may fall into different categories of skin color in different geographic regions (Marcheco-Teruel et al., 2014; Parra et al., 2003). In Africa, skin color is diverse and genetic signatures of selection for both lighter and darker skin color have been reported (Crawford et al., 2017; Genomes Project Consortium et al., 2015). Similarly, lactase persistence has evolved via different genetic mechanisms and in different parts of the world where cattle-herding and milk consumption are culturally embedded (Heyer et al., 2011; Segurel & Bon, 2017; Tishkoff et al., 2007; Wiley, 2020).

Another factor that influences worldwide genetic diversity is *migration*. Recent large-scale migration leads to population mixing that is referred to as admixture. These scenarios can have different consequences for genetic studies and must be treated appropriately during data quality control analyses and interpretation. For example, recent migration can lead to long-range LD as recombination has not had time to do its mixing work, and recently admixed populations can present an excess of heterozygotes. While drift and selection can cause allele frequencies to diverge across populations, sharp genetic boundaries between neighboring populations are rare. Instead, migration causes allele frequencies to change smoothly with distance, which is referred to as a *cline*. This is exemplified by European populations where migration is known to have played a major role, and many allele frequencies vary smoothly across the continent, even across language boundaries and national borders (Lao et al., 2008).

# 3 | METHODS AND TOOLS FOR INCLUDING DIVERSE AND ADMIXED POPULATIONS

A variety of methods and tools are available for analyzing GWAS (M. H. Wang et al., 2019). The simplest ones are performed on a single-variant basis and use linear or logistic regression such as provided by PLINK (Chang et al., 2015). Linear or logistic mixed models are able to incorporate correlation structure between individuals arising, e.g., by relatedness or population structure (H. Chen et al., 2016; Kang et al., 2010; Zhou & Stephens, 2012). Recently, Bayesian approaches have become more common (Kaplan et al., 2020; Lock & Dunson, 2017). Utilizing the results of GWAS, polygenic scores (PGS) summarize the effects of multiple genetic variants in one variable. They are usually generated by selecting a number of most influential variants and adjusting for LD. For the calculation of individual PGS, the genotype dosage of variants is weighted by the variant effect sizes, which are either directly taken from GWAS or recalculated using refined methods (Hindorff et al., 2018).

Post GWAS investigations aim to detect the functional effects of genetic variants. Here, analyses of expression quantitative trait loci (eQTL) and of tissue- or cell-specific expression are frequently performed. Other approaches include colocalization, which aims to discover genes for which a single causal genetic variant underlies two phenotypes (Giambartolomei et al., 2014; Hormozdiari et al., 2016) and prediction of pathogenicity (Knecht & Krawczak, 2014). Populations with diverse and admixed ancestry due to recent migration require specific analytic considerations. Whereas causal mechanisms are assumed to be the same across populations, the between-population differences in LD between a causal variant and marker can lead to large differences in observed effect sizes. Because the LD structure can be considerably different between populations, methodological problems arise when jointly analyzing GWAS data from cohorts of different ancestry or of admixed populations. Additional complications can arise due to gene-environment interaction. Here, we highlight some of the methodological challenges in genetic epidemiology studies and present analytic strategies that may be used to mitigate these challenges. We also point out limitations, highlighting areas that require further methodological development.

## 3.1 | Population structure

Systematic allele frequency differences in subgroups of individuals within a study population, known as population structure, can confound genetic association studies, leading to false positive findings (Patterson et al., 2006; Price et al., 2006). Principal component analysis (PCA) is commonly used in GWAS to adjust for population structure with or without the addition of external reference samples. Plots of the first few principal components (PCs) can be used to identify subpopulations, and including PCs as covariates in regression models can adjust for confounding due to low levels of population structure (Lazaridis et al., 2014). While PCA is good at capturing global ancestry information across autosomal chromosomes, it is less effective at capturing different forms of population structure such as admixture which can inform association analysis. Another drawback is that determining the number of PCs to include as covariates is ad hoc and using too few or too many PCs can lead to spurious associations and loss of power (J. Zhang, 2010). Whereas autosomal variants are inherited from both mother and father, the Y chromosome is passed from father to son only and the mitochondrial DNA is passed from a mother to all her children. Depending on how admixture, intermixing and the asymmetry of intermixing occurred in previous generations, an individual's Y-chromosomal and mitochondrial ancestries may differ substantially from ancestries at autosomal loci.

Population structure can influence trait association, causal inference, heritability and trait prediction models. Statistical methods must be adapted to both the scientific question and the underlying population structure. For example, if ancestry is associated with an environmental factor that is causally associated with the trait, adjustment using PCs alone may insufficiently account for confounding. Ideally, the environmental factor should be included in the model, but this information is often unavailable (Lawson et al., 2020). Of note, the detection of population structure depends on sample size. Only broad population structure can be detected in small samples, whereas more fine-scale structure can be evident in larger studies.

Large genetic drift that can occur in small populations or strong bottlenecks can lead to allele frequency differences for which adjusting for PCs may not be effective in accounting for population structure (Lawson et al., 2018, 2020). Selection may also impact allele frequencies differently between populations, which could affect genetic associations. Especially susceptible to population structure are studies of rare genetic variants, where frequencies are highly variable across populations, and studies of haplotypes that are highly polymorphic. LD structure varies widely between populations as well. As such, investigations of rare variants or those dependent on LD, such as GWAS and PGS, can result in population-dependent results. In addition, ancestry-related assortative mating in a population can contribute to maintaining population stratification and can lead to long-range LD, which may further impact genetic association studies (Risch et al., 2009; Sebro et al., 2010).

## 3.2 | Ancestry estimation and analysis approaches for admixed populations

Demographic events such as migration and mating between isolated populations with distinct genetic structures lead to admixed populations with mosaic chromosomal segments from each isolated population (Shriner, 2017). Admixture is a common feature of human history and has shaped present-day human genetic variation (Hellenthal et al., 2014; Patterson et al., 2012; Shriner et al., 2016). One of the earliest known examples of archaic admixture in hominins is that of *Homo sapiens* and *Homo neanderthalensis* (Durvasula & Sankararaman, 2020; Sankararaman et al., 2014; Wall et al., 2013). Recent admixture triggered by forced and voluntary human migration and mixing has introduced major gene flow among geographically isolated populations (Micheletti et al., 2020). Examples of recently admixed populations include African Americans, an admixed population of African and European ancestry (Micheletti et al., 2020), populations in Central and South America, which are admixed populations of Amerindigenous, European and African ancestry (Bryc et al., 2010; Kehdy et al., 2015), and multi-ancestral admixed populations in South Africa (Kodaman et al., 2013). Koehl and Long investigated 17 admixed populations in the Americas and showed that their genetic diversity reflected varying contributions of admixture and genetic drift (Koehl & Long, 2018). In genetic epidemiology studies, admixed populations present unique challenges for association analyses between genotypes and phenotypes when the causal genetic variant has different frequencies across the ancestral populations. Individuals belonging to an admixed population group vary in their distributions of ancestry proportions. This genetic heterogeneity and complex substructure can confound associations and lead to spurious findings unless properly accounted for in the analyses (Landry et al., 2018; Thornton & Bermejo, 2014).

Several statistical tools have been developed for ancestry inference and admixture analysis. Statistical tests such as the D test (Durand et al., 2011) and *f* test (Patterson et al., 2012) are useful to detect the existence of admixture and the direction and magnitude of gene flow between parental and admixed populations. An estimate of the genomic contributions of different ancestral populations in an admixed population can be inferred at global or local levels. The *local ancestry* is the ancestry at a specific genomic region, while the *global* ancestry of an individual is the genome-wide distribution of its *local* ancestry (Mersha, 2015). In Table 1, we summarize important characteristics of some of the many approaches that have been proposed to estimate ancestry proportions. Comparison of admixture mapping tools and detailed reviews have previously been published (Padhukasahasram, 2014; Seldin et al., 2011). Genetic ancestry inference techniques have different assumptions and limitations, and a careful selection of the appropriate method for a given study setting is necessary. The robustness of results can be assessed by comparing multiple methods. In general, differences between admixture inference tools can be due to the underlying model assumptions such as whether LD between SNPs is considered or not, applicability to multiway admixed populations, accuracy, computational speed, the need for dense genotypes and whether the input genotype data need to be phased.

The underlying models of fineSTRUCTURE (Lawson et al., 2012), ADMIXTURE (Alexander & Lange, 2011), and IPCAPS (Chaichoompu et al., 2019) estimate global ancestral proportions, while HAPMIX (Price et al., 2009), SEQMIX (Hu et al., 2013), LAMP-LD (Baran et al., 2012), RFMix (Maples et al., 2013), MULTI-MIX (Churchhouse & Marchini, 2013), and ALDsuite (Johnson et al., 2015) provide inference on local ancestry. To infer ancestry, SEQMIX (Hu et al., 2013) uses sequence reads from exome or targeted sequencing of selected regions of interest that have not mapped to the intended regions. HAPMIX models background LD and describes how the haplotypes of admixed individuals relate to those in the ancestral populations. Unlike HAPMIX and SEQMIX that are only applicable to admixtures of two ancestral populations, the other listed software packages in Table 1 are useful for estimating ancestries of populations that are admixtures of two or more ancestral populations. MULTI-MIX estimates local ancestry using a model on background LD and can handle either phased or unphased genotype data. A limitation of many current methods is that they require large LD reference panels, which remain under-representative of individuals from minority population groups. RFMix can utilize ancestry information contained within admixed samples themselves rather than an LD reference panel and is also fast enough to analyze tens of millions of SNPs (Maples et al., 2013; Uren et al., 2020).

In the analysis of genetic data from admixed populations, accounting for local ancestry in addition to the traditional approach of accounting for global ancestry improves the accuracy of result (Landry et al., 2018; Liu et al., 2013; Zhang & Stram, 2014). However, this comes at the cost of longer computation time. Therefore, a practical recommendation is to consider a two-stage approach in which the first model includes adjustment for global ancestry and the second model includes local ancestry adjustment on candidate loci (Gay et al., 2020). On the other hand, when causal genetic variants differ between ancestral populations, admixture mapping methods can leverage admixed populations to identify ancestry-specific loci associated with phenotypes (Patterson et al., 2004). Methods that jointly test genotype-phenotype associations and admixture mapping achieve higher power in genotype-phenotype tests and reduce the testing burden in ancestry-phenotype tests (W. Chen et al., 2015; Pasaniuc et al., 2011; Shriner et al., 2011).

### 3.3 | Multi-ancestry and meta-analysis GWAS approaches

Meta-analysis of individual studies grouping individuals by ancestral populations can be an effective way to circumvent issues related to population stratification and has similar power as analyses pooling individual-level data (Lin & Zeng, 2010; Willer et al., 2010). Traditional meta-analyses assume ancestral homogeneity between populations, which may not be appropriate for multi-ancestry studies where heterogeneity is common. Meta-analysis methods such as MR-MEGA, MANTRA and GWAMA aim to address this issue by modeling allele effect heterogeneity between populations (Table 2). However, stratification by ancestral populations can be problematic, as genetic ancestry is a continuum and individuals do not distinctly group into a single ancestral population, with the majority of individuals having

**TABLE 1** Software and methods for deriving population structure

| Software | Description | Software Platform | Input | Notes | References |
|---|---|---|---|---|---|
| *Local ancestry software*[a] | | | | | |
| HAPMIX | HMM<br>Two ancestral populations | C++ | Reference haplotypes<br>Unphased genotypes | Haplotypes of each admixed individual are viewed as being sampled from the reference populations. | Price et al. (2009) |
| SEQMIX | HMM<br>Two ancestral populations | C++ | Allele frequencies for reference populations Genetic distances<br>VCF for gene sequence variations | Using aligned bases of combined off-target and on-target sequence reads directly to infer whole-genome ancestry in admixed samples | Hu et al. (2013) |
| LAMP-LD | HMM<br>Multiple (≥2) ancestral populations | C++ | Reference haplotypes<br>Unphased genotypes | Using nonoverlapping windows<br>No transitions between ancestries within each window<br>An order of magnitude faster than HAPMIX | Baran et al. (2012) |
| RFMix | Linear-chain conditional random field (CRF) model + random forests<br>Multiple (≥2) ancestral populations | C++ | Reference haplotypes VCF for gene sequence variations | Dividing each chromosome into windows and inferring local ancestry within each window by using a CRF parameterized by random forests trained on reference panels | Maples et al. (2013) |
| MULTI-MIX | HMM<br>Multiple (≥2) ancestral populations | C++ | Either phased or unphased data | A multivariate normal model on haplotype probabilities given ancestry and an HMM on how ancestry changes along a chromosome | Churchhouse and Marchini (2013) |
| ALDsuite | HMM + PCA<br>Multiple (≥2) ancestral populations | R | Phased haplotypes | Using an HMM to model switches between ancestral states across each individual's\ chromosomes; using PCs to approximate admixture LD; offering both local and global ancestry inference using dense marker data | Johnson et al. (2015) |
| *Global ancestry software* | | | | | |
| fineSTRUCTURE | HMM, MCMC<br>Multiple (≥2) ancestral populations | C++/R | Phased haplotypes<br>Recombination rates | A haplotype segment viewed as the 'recipient' of genetic material from nearest neighbor 'donor' haplotype segments; such donor-recipient relationships created for every haplotype in the data | Lawson et al. (2012) |

TABLE 1 (Continued)

| Software | Description | Software Platform | Input | Notes | References |
|---|---|---|---|---|---|
| ADMIXTURE | Bayesian clustering approach<br>Multiple (≥2) ancestral populations | C++ | Unphased genotypes | Estimating individual ancestries by computing maximum likelihood estimates in a parametric model | Alexander and Lange (2011) |
| IPCAPS | PCA<br>Multiple (≥2) ancestral populations | R | Unphased genotypes | Building on the iterative pruning PCA framework that systematically assigns individuals to genetically similar subgroups. It can be used with transcriptome or epigenome data. | Chaichoompu et al. (2019) |

Abbreviations: HMM, hidden Markov model; MCMC, Markov chain Monte Carlo; PCA, principal component analysis; VCF, variant call format.

[a] Global ancestry proportion can be calculated as the genome-wide average of local ancestry proportions.

mixed ancestry. Because of this, the practice of grouping individuals into single homogeneous populations, which typically entails excluding those who do not group into a distinct population may not be optimal. An alternative is to conduct a joint analysis across ancestries while allowing for differences in trait variance across groupings based on population characteristics as a random variable in a mixed model (Wojcik et al., 2019).

### 3.4 | Mixed models and relatedness between individuals

Unaccounted relatedness can lead to spurious association signals due to false assumptions of independence between individuals (William & David, 2009). To address relatedness in association analysis, there are analysis approaches that can pool together individuals of diverse ancestry by using mixed models to incorporate a measure of genetic similarity, such as the popular genetic relationship matrix (GRM), as a random effects variance-covariance matrix (Table 2). Genetic relatedness is a fundamental concept in human genetics, and approaches for defining and estimating relatedness have been extensively described previously (Speed & Balding, 2015; Thompson, 2013; J. Wang, 2016). Speed and Balding (2015) prefer the term *genetic similarity matrices* from the perspective that genetic similarity is what is captured from such approaches rather than relatedness itself. Ramstetter et al. (2017) evaluated 12 pairwise relatedness inference methods using genome-wide data, including for example KING (Manichaikul et al., 2010) and PC-Relate (Conomos et al., 2016), which provide an excellent reference. In Table 3, we highlight some additional GRM computation tools recently developed, focusing on those that accommodate samples of heterogeneous ancestry and sequence data, excluding IBD segment inference tools. Many pooled GWAS analysis tools using mixed models include a GRM computation feature. Pooled GWAS analysis methods that also include the estimation of a GRM include BOLT-LMM (Loh et al., 2015, 2018), SAIGE (Mogil et al., 2018), SUGEN (Lin et al., 2014), GENESIS (Naseri et al., 2019), and GEMMA (Zhou & Stephens, 2012) (Table 2).

Association analyses based on mixed models have the advantage of enabling the inclusion of individuals with recent and distant relatedness and circumvent the need to arbitrarily keep unrelated individuals only. Because pooled analyses use individual-level data, covariates need to be consistently included across studies, which also allows the investigation of potential interactions. In such studies, it is common to additionally adjust for PCs capturing genetic ancestry, and adjustment for

**TABLE 2** GWAS methods for inclusion of multi-ancestry populations

| Method | Description | Outcome (binary/ quantitative) | Software platform | Suitable for $N > 100$ k scale analyses | Notes | Reference |
|---|---|---|---|---|---|---|
| *Pooled approaches accounting for relatedness and ancestry* | | | | | | |
| REGENIE | Ridge regression<br>Generalized linear model | Binary<br>Quantitative | C++ | Yes | Reduces computational burden of calculating GRM by instead accounting for local LD within blocks with "local polygenic scores" calculated using ridge regression<br>Supports Firth and SPA for binary traits<br>Can analyze multiple phenotypes in parallel | Mbatchou et al. (2021) |
| fastGWA | Linear mixed model[a],[b] | Binary<br>Quantitative | C++ | Yes | Uses grid-search-based REML algorithm to estimate a sparse GRM<br>Implemented in GCTA | Jiang et al. (2019) |
| BOLT-LMM | Linear mixed model[a],[b] | Binary<br>Quantitative | C++ | Yes | Models non-infinitesimal genetic architecture via a Bayesian mixture prior on variant effect sizes | Loh et al. (2015)<br>Loh et al. (2018) |
| SAIGE | Logistic mixed model[a] | Binary<br>Quantitative | R | Yes | Controls for unbalanced case-control ratios using SPA | Zhou et al. (2018) |
| SUGEN | Generalized estimating equations | Binary<br>Quantitative | C++ | Yes (?) | Accounts for unequal inclusion probabilities using weighted version of GEEs | Lin et al. (2014) |
| GENESIS | Logistic mixed model[a]<br>Linear mixed model[a] | Binary<br>Quantitative | R | Yes (?) | Allows for multiple variance components<br>Part of a suite of tools utilizing GDS format | Gogarten et al. (2019) |
| GMMAT | Logistic mixed model[a] | Binary<br>Quantitative | R | No | Utilizes GDS format<br>Beta coefficients not provided for GWAS | H. Chen et al. (2016) |
| GEMMA | Linear mixed model[a] | Quantitative | C++ | No | Exact method | Zhou and Stephens (2012) |
| EMMAX | Linear mixed model[a] | Quantitative | C++ | No | Approximation method | Kang et al. (2010) |
| Tractor | Local ancestry aware logistic regression model | Binary | Python, Hail | Yes | Facilitates the inclusion of admixed individuals by leveraging local ancestry | Atkinson et al. (2021) |
| *Meta-analysis approaches enabling multiple ancestries* | | | | | | |
| MANTRA | Bayesian partition model | Binary<br>Quantitative | Fortran (?) | Yes | Models allelic heterogeneity between populations using a hybrid fixed and random effects approach | Morris (2011) |
| MR-MEGA | Multi-dimensional scaling | Binary<br>Quantitative | C++ | Yes | Models heterogeneity in allelic effects as a function of pairwise mean allele frequency differences between populations | Magi et al. (2017,2020) |

TABLE 2 (Continued)

| Method | Description | Outcome (binary/ quantitative) | Software platform | Suitable for $N > 100$ k scale analyses | Notes | Reference |
|---|---|---|---|---|---|---|
| TransMeta | Kernel-based random-effects model | Binary<br>Quantitative | R | Yes | Models heterogeneity based on the correlation structure of allelic effects, which are treated as random variables<br>Beta coefficients not provided | Shi and Lee (2016) |
| GWAMA | Random- or fixed-effects meta-analysis | Binary<br>Quantitative | C++ | Yes | Models allelic heterogeneity using random-effects in the presence of heterogeneity and fixed effects otherwise | Magi and Morris (2010) |
| METAL | Fixed effects meta-analysis | Binary<br>Quantitative | C++ | Yes | Meta-analyzes either 1) P-values and effect directions or 2) effect size estimates weighted by standard errors. | Willer et al.(2010) |
| METASOFT | Random- or fixed-effects meta-analysis | Binary<br>Quantitative | Java | Yes | Includes conventional models and a random effects model that increases power under heterogeneity. | Han and Eskin (2011) |

Abbreviations: GDS, genomic data storage; GEE, generalized estimating equations; GRM, genetic relationship matrix; GWAS, genome-wide association studies; LD, linkage disequilibrium; REML, restricted maximum likelihood; SPA, saddlepoint approximation.

[a] Mixed models require either an externally pre-calculated or internally calculated GRM.

[b] Note that linear mixed models are not designed to analyze binary traits and can have inflated type I error rates.

self-reported race/ethnicity has been applied in previous studies to account for potential social or cultural confounders (Wojcik et al., 2019). Studies have shown that pooled approaches using mixed models control for genomic inflation (estimated using the lambda factor; an indicator of the impact of population structure and other confounders on association results [Devlin & Roeder, 1999]) better than either adjusting for PCs or excluding related individuals in generalized linear models (Y. Zhang & Pan, 2015). However, the need for individual-level data in mixed models has the limitation of being more computationally intensive and more subject to data sharing and privacy issues than studies that rely on the meta-analysis of GWAS summary statistics alone (Pasaniuc & Price, 2017). The use of mixed models for multi-ancestry GWAS do leave some open questions with respect to how to best model effect size differences that can occur between populations due to the different LD structures and differences in social and environmental factors (Peterson et al., 2019). More research from a methodological perspective, either theoretical or by simulations, is warranted. More recently, a statistical method and software named *Tractor* (Atkinson et al., 2021) facilitates the inclusion of diverse and minority populations in genomic research (Table 2) by accounting for local ancestry and population differences in minor allele frequencies at each tested variant. Peterson et al., provide a useful review of quality control considerations and analysis strategies in diverse populations (Peterson et al., 2019).

## 3.5 | PGS in the context of diverse populations

PGS, including polygenic risk scores (for case-control traits), aggregate many genetic variants investigated in GWAS (Hindorff et al., 2018; Janssens, 2019). PGS have the advantage of summarizing the effects of multiple genetic variants, often with small effects, into a single variable under the assumption of additive genetic effects. PGS have been applied in biomedical and social science research (Knowles & Ashley, 2018; Pasaniuc & Price, 2017) and by direct-to-consumer genetic testing companies. It is expected that PGS will improve health outcomes by optimizing diagnostic screening practices and patient-tailored treatments (Duncan et al., 2019; Torkamani et al., 2018), potentially bridging the gap between discovery of susceptibility genetic variants and clinical implementation (Wand et al., 2021), particularly when combined with conventional risk predictors (Elliott et al., 2020).

**TABLE 3** Tools for genetic relatedness matrix (GRM) estimation for inclusion of multi-ancestry populations.

| Tool | Method type | Description | Software language | Input file type and external info needed | Output | Notes | Reference |
|---|---|---|---|---|---|---|---|
| *For called genotype (SNP) data only* | | | | | | | |
| SNPRelate | Allele frequency-based IBD estimate | For multi-core symmetric multiprocessing computer architectures For principal component analysis (PCA) and relatedness analysis | R with C/C++ kernel | GDS/VCF/NetCDF | IBD 0,1,2 GRM matrix | Can be used for population structure with KING-robust method | Zheng et al. (2012) |
| RaPID | IBD-segment based IBD estimate | Based on the positional Burrows-Wheeler transform<br>Linear time search for shared segments in an arbitrarily large cohort | C++ | VCF<br>Genetic map | IBD 1,2 | Assumes randomly distributed genotyping errors and correct genetic map | Naseri et al. (2019) |
| PedKin | Pedigree-based | Set of algorithms for computing the kinship coefficient of a set of individuals with a known pedigree | C++ | PLINK ped<br>Inbreeding values for founders | GRM matrix | Considers the possibility that the founders of the known pedigree may themselves be inbred | Kirkpatrick et al. (2019) |
| IBIS | IBD-segment based IBD estimate | Locates long regions of allele sharing between unphased individuals, detects IBD segments and infers degrees of relatedness | C++ | PLINK binary<br>Genetic map | Kinship, IBD 2 | For both admixed and unadmixed samples<br>Assumes universal allele sharing for missing data sites | Seidman et al. (2020) |
| popkin | Allele frequency-based IBD estimate | Method of moments estimation for kinship and FST<br>Practically unbiased for any population structure | R/Rcpp | Genotype matrix or PLINK BED<br>Vector of baseline kinship values | GRM matrix | Accounts for heterogeneous samples with population structure and admixture | Ochoa and Storey (2021) |
| *For called genotype (SNP) data/genotype likelihood (dosage) data/sequence data* | | | | | | | |
| MAPGD | Maximum Likelihood-based IBD estimate | Estimates of inbred-relatedness coefficients from population genomic data | C++ | BAM<br>mfileup file from samtools | 7 genotypic correlation coefficients | | Ackerman et al. (2017) |
| SEEKIN | Allele frequency-based IBD estimate | Estimates kinship for both homogeneous samples and heterogeneous samples using sparse sequence reads | C++ | VCF<br>Ancestry coordinates generated by LASER | GRM matrix | Accounts for heterogeneous samples with population structure and admixture | Dou et al. (2017) |

TABLE 3 (Continued)

| Tool | Method type | Description | Software language | Input file type and external info needed | Output | Notes | Reference |
|---|---|---|---|---|---|---|---|
| TRUFFLE | IBD-segment based IBD estimate | Integrates computational techniques and statistical principles for the identification and visualization of IBD segments | C++ | VCF | IBD 0,1,2 | Uses un-phased data by skipping the haplotype phasing step and, instead, relying on a simpler region-based approach | Dimitromanolakis et al. (2019) |
| NgsRelate v2 | Allele frequency-based IBD estimate | Estimates the 9 condensed Jacquard coefficients along with various other relatedness statistics from high-throughput sequencing data | C++ | BAM/CRAM, VCF/BCF, PLINK BED | IBD 0,1,2 | Accounts for arbitrary inbreeding patterns within homogeneous population | Hanghoj et al. (2019) |
| NGSremix | Maximum Likelihood-based IBD estimate | Maximum likelihood estimation tool of relatedness between pairs of admixed individuals from low-depth NGS data | C/C++ | PLINK binary or BEAGLE format for genotype likelihood<br>Individual ancestry proportions and population specific allele frequencies | IBD 0,1,2 | Accounts for heterogeneous samples with population structure and admixture<br>Takes the uncertainty of the genotypes into account via genotype likelihoods and handles admixture by estimating paired ancestry proportions and including these in the model | Nohr et al. (2021) |

Abbreviations: BAM, binary alignment map; BCF, BIM collaboration format; GRM, genetic relationship matrix; IBD, identity by descent (IBD 0, 1, 2 denotes the probability that two individuals have 0, 1, or 2 alleles at a locus identical by descent); VCF, variant call format.

Interest in developing and applying PGS to predict genetic liability to complex traits has been fueled by new methodologies (e.g., clumping and *p* value thresholding [Choi & O'Reilly, 2019; Wray et al., 2014] penalized regression (lassosum [Mak et al., 2017]) and accounting for LD between predictors (LDpred [Vilhjalmsson et al., 2015]) and the publication of thousands of GWAS with large sample sizes (Hindorff et al., 2018; Khera et al., 2018). However, because PGS are largely derived from studies of European ancestry participants, they often have limited performance in non-European ancestry populations (Dikilitas et al., 2020; Duncan et al., 2019; Martin et al., 2019; Popejoy & Fullerton, 2016) with few studies or traits providing evidence of trans-ancestry portability (Conti et al., 2021; M. Wang et al., 2020). Population differences in allele frequencies of susceptibility variants, their effect sizes and LD patterns contribute to the non-generalizability of PGS across populations (Martin et al., 2017).

Recently, approaches have been proposed to develop PGS for diverse populations. Simulations and real data applications have shown that including in PGS variants discovered in diverse ancestral populations led to reduced bias and higher genetic risk accuracy in admixed individuals (Cavazos & Witte, 2021). Another approach, XP-BLUP, proposes a multiple-component linear mixed model to incorporate ancestry-specific weights to address the need for efficient prediction in minority populations (Coram et al., 2017). An approach called partial PGS, a method that estimates local genetic ancestry and applies a combination of ancestry-specific PGS, found that when GWAS data are available for more than one ancestry, the combination of multiple partial PGS improves trait predictability in admixed individuals (Marnetto et al., 2020). The formation of PGS consortia such as The Polygenic Risk Scores Diversity Consortium (https://www.genome.gov/Funded-Programs-Projects/Polygenic-Risk-Score-Diversity-Consortium) will likely lead to further improvements in PGS performance across populations.

## 4 | RESOURCES FOR INCLUDING DIVERSE AND ADMIXED POPULATIONS

Genetic resources from diverse populations, such as sequencing and genome-wide genotyping array datasets, are important in population genetics and genetic epidemiology research, particularly as reference panels for estimating ancestry or LD. The LD structure is fundamental for many analyses in genetic epidemiology and required for, e.g., accurate imputation, colocalization and fine-mapping. Although a diverse reference panel might be preferred for better imputation across various ancestry groups, colocalization and fine-mapping can be performed meaningfully only if the LD structure can be inferred based on samples from an ancestrally similar population. However, existing resources have small sample sizes for non-European ancestry populations and do not reflect global diversity. To reduce health disparities in genomic medicine, it is necessary to continue to enrich those resources with samples from diverse population groups. Key to this is to increase sample sizes for underrepresented populations by building research consortia. Highly populated regions of the world consist of diverse populations that remain underrepresented in genomics (Martin et al., 2019; Wojcik et al., 2019). Previous and ongoing global projects that have built sequence resources on diverse genomes include the 1000 Genomes Project (Byrska-Bishop et al., 2021; Genomes Project Consortium et al., 2015), the Human Genome Diversity Project (Bergstrom et al., 2020), and the Human Genome Structural Variation Consortium (Ebert et al., 2021). Regional initiatives such as the African Genome Variation Project (Gurdasani et al., 2015), the Human Heredity and Health in Africa (H3Africa) (H3Africa Consortium et al., 2014), the Non-Communicable Diseases Genetic Heritage Study (NCD-GHS) consortium in Nigeria (Fatumo, Yakubu, et al., 2022), the Qatar Genome Program (Thareja et al., 2021), Japan Biobank (Sakaue et al., 2021), and the GenomeAsia100K Consortium (Wall et al., 2019) have made important contributions to genomics research for African and Asian populations. In the United States, recent large-scale programs to accelerate the representation of minority and underrepresented populations in genomics and beyond include the NHLBI TOPMed Program (Taliun et al., 2021) and the All of US research program (All of Us Research Program Investigators et al., 2019). Admixed populations pose additional analysis challenges because neither of the population-specific reference databases can be applied directly. For such purposes, more diverse reference panels are needed. The composition of these is non-trivial and further research has to be launched to investigate what constitutes a meaningful reference database and under which analysis scenario for admixed populations.

A key resource for the identification and functional interpretation of disease-associated variants in whole genome or exome sequencing analysis is the human reference genome (Lappalainen et al., 2019; Wong et al., 2020). However, the most widely used haploid reference genomes (release GRCh37/GRCh38) do not yet adequately represent the genomic diversity of human populations. This means that population-specific rare variants, haplotypes and structural variants cannot be captured well for certain populations. This can have implications for the development of individualized therapies based on those markers (Chrisman et al., 2022). The recent Human Pangenome Reference Consortium (HPRC) is working to establish a human genome reference that reflects the existing worldwide human diversity (https://humanpangenome.org/). Future work is needed to investigate optimal usage of ancestry and admixture in genome references.

Genome function databases from diverse populations are valuable post-GWAS tools to support the interpretation of genetic associations. Integrating GWAS loci with molecular traits, such as gene expression and methylation in trait-relevant tissues, enables the prioritization of potentially causal genes, which can be targeted for developing molecular diagnostics and therapeutics. One approach in functional genomics uses genotype data to predict gene expression by identifying eQTL (Barbeira et al., 2018; Gamazon et al., 2015; Shriner et al., 2016), and associations between the predicted gene expression and a given trait of interest are evaluated using Mendelian randomization for causal inferences (Hauberg et al., 2017; Pavlides et al., 2016; Zhu et al., 2016). Another approach is colocalization to identify variants underpinning multiple molecular traits, such as gene expression and methylation (Giambartolomei et al., 2014; Hormozdiari et al., 2016). However, the most widely used database for integrated functional analyses of tissue-specific gene expression targets, the Genotype Tissue Expression Portal (GTEx), consists of more than 80% European-ancestry individuals (GTEx Consortium, 2020). The accuracy and power of gene expression prediction models improve when models are built in individuals whose ancestry is representative of the phenotyped individuals (Keys et al., 2020; Mikhaylova & Thornton, 2019; Mogil et al., 2018). Diverse samples also enable the identification of potential differences in eQTL effects between populations (GTEx Consortium, 2020) and detecting additional colocalization signals that can advance functional interpretation of GWAS loci (Mogil et al., 2018). Therefore, concerted efforts are warranted to increase the representation of diverse groups not only in GWAS, but also in tissue- and cell-specific molecular databases to advance the interpretability of GWAS discoveries. Another area which can benefit significantly from data from diverse populations is fine-mapping (Tehranchi et al., 2019). By applying different statistical methods, fine-mapping utilizes LD structure and/or functional annotation to prioritize variants that are potentially causal which can be followed up in functional studies. Because diverse populations present divergent LD structures, multi-ethnic fine-mapping is superior in power to single-ethnic analyses (Mahajan et al., 2022). There are several methods and tools available such as trans-ethnic PAINTOR (Kichaev & Pasaniuc, 2015) and MsCAVIAR (LaPierre et al., 2021) that both use a multivariate normal distribution for modelling. MsCAVIAR additionally integrates functional annotation, MR-MEGA applies trans-ethnic mega regression (Magi et al., 2017).

To build more diverse and representative resources and databases, concerted efforts are needed. Methods adapted from the social science field can be used to explore recruitment strategies into genomics research for underrepresented research participants (Bull et al., 2012). A practical example in an African setting is an implementation of community-level rapid assessment before a genomics research on podoconiosis (Tekola et al., 2009a, 2009b). The assessment revealed locally sensitive notions and concerns, which have been incorporated in recruitment and consent process strategies to minimize stigmatization of research participants, address cultural norms, which led to successful recruitment (Tekola Ayele et al., 2012). In their roadmap to increase diversity in genomic studies, Fatumo, Chikowore, et al. (2022) propose that building trust can promote the engagement of underrepresented participants in genomic research. Researchers can develop genuine partnerships with local communities and minority groups such as through community advisory boards that are sustained by community members. Funders can also play a pivotal role in promoting research in underrepresented populations. Strategic funding from the National Institutes of Health (NIH) and Wellcome Trust was successful in creating the H3Africa initiative. Strategic programs that can account for the needs of local research communities that struggle with lower competitive edges should be considered. The funding opportunities from research intensive countries should be more inclusive to applications from underrepresented populations and consider developing funding schemes that promote scientific networking and sustainability of the research capacity in those populations (Fatumo, Chikowore, et al., 2022).

## 5 | BIOMEDICAL, CLINICAL, AND PUBLIC HEALTH PERSPECTIVES

The possibility of adapting PGS for clinical use is under consideration for some traits. However, the use of PGS in both research and clinical applications raise potential ethical concerns, such as the exacerbation of health inequities (see Section 3.5 for further discussion) (Martin et al., 2019; Palk et al., 2019; Polygenic Risk Score Task Force of the International Common Disease Alliance, 2021).

Scientific knowledge of the clinical significance of genetic variants, particularly rare variants, is not yet based on data from diverse populations. Most samples in the Genome Aggregation Database (gnomAD) (Karczewski et al., 2020) come from populations of European ancestry, and many of the variants listed as "pathogenic" or "likely pathogenic" in ClinVar (Popejoy et al., 2018) were identified in European ancestry populations. This has consequences for the translation to health care because clinically actionable genetic variants are often initially identified using in silico prediction tools that rely on population-based datasets. When genomic data from a limited range of populations are used to develop in silico prediction pipelines, the accuracy of variant classification outside these populations may be poor, which can lead to incorrect molecular diagnostics. For example, genetic variants that were misclassified as pathogenic for hypertrophic cardiomyopathy and subsequently used in clinical decisions were later found to be common among African Americans, which led to a reversal of some clinical decisions (Manrai et al., 2016). The problem is of particular relevance for African populations because of the high genetic diversity of the African continent and for which rare pathogenic variants have not yet been included in databases. One example includes monogenic variants for hearing loss where a comparatively low number of variants have been identified for African compared to European populations (Wonkam & de Vries, 2020). Also, environment-related selection gives rise to variants that are more relevant in specific groups such as African-ancestry enriched variants related to sickle cell disease (Wonkam & de Vries, 2020), and *APOL1* variants that are thought to have undergone selection rising to high frequency in sub-Saharan Africa to protect against a deadly form of sleeping sickness, but are associated with higher risk for chronic kidney disease (Genovese et al., 2010). Rare mutations that contribute to about half of congenital non-syndromic hearing impairment in European populations are clustered in the *GJB2* gene, but these mutations have limited predictive relevance in sub-Saharan African populations (Lebeko et al., 2016).

Pharmacogenomics has been successfully applied in clinical genetics. However, underrepresented populations have lower potential to benefit from predicted drug responses. In this regard, a recent GWAS on schizophrenia patients of African ancestry undergoing clozapine treatment showed a strong association of a regulatory variant in the *ACKR1* gene with the development of neutropenia, a serious adverse event that results in stopping the clozapine treatment (Bentley et al., 2020; M. H. Wang et al., 2019). This variant is rare in European or Asian ancestry individuals, and was only discovered because the GWAS was performed in African ancestry individuals.

Increasing the diversity of the biomedical research workforce can also contribute to improving scientific innovation derived from having different perspectives that promote creativity. Concerted efforts are being made to promote the training and support of scientists from underrepresented backgrounds working in the genomic field (Bonham & Green, 2021). Because groups who do not take part in genomic studies are less likely to benefit from medical advances based on genomics, it is imperative that the scientific community strives to be inclusive at all stages of study design, both in terms of study participants and researchers. Working to promote transparency, authors should describe clearly how diverse participants were included or explain why they were excluded from a study (Hindorff et al., 2018).

### 5.1 | Limitations and interpretation issues

Interpretation of admixture mapping findings in the context of trait differences between populations warrants careful consideration of its assumptions and limitations. Ancestry estimation in admixed populations is error-prone because the true historic ancestral parent populations are unknown and the availability of representative ancestral reference genetic data is limited for some populations (Baran et al., 2012). Therefore, replication in independent cohorts remains critical to validate associations identified by admixture mapping (Mersha, 2015). Regarding interpretation, it is important to bear in mind that allele frequency differences between geographically distant ancestral populations may be driven by confounding environmental factors unrelated to ancestry or the phenotype under consideration. An association of global ancestry with a phenotype may be partly or fully due to a confounding correlation of global ancestry with lifestyle and sociodemographic factors associated with the phenotype. Therefore, it is important to avoid generalizations that invoke "genetic" explanations for group differences based on observed associations of global ancestry with trait differences between population groups.

Genetic differences between individuals within the same population are often higher than differences between individuals from different populations (Rosenberg, 2011; Witherspoon et al., 2007). In addition, population differences are sometimes artificially enhanced by study design or analysis methods, for instance when studies investigating differences between populations exclude admixed participants, leading to exaggeration of population differences. Software designed to genetically differentiate populations may depend on the use of discrete populations, ignoring the possibility of clines (Lawson et al., 2018). Furthermore, the popular practice of removing "outliers," for example, based on a PCA plot, can generate artificial population groupings and lead to the mistaken impression that populations are more genetically distinct than they actually are. Genetic population data should be reviewed to decide whether clusters or continuous clines adequately describe the population structure, taking external population information into account. This not only has implications for the statistical methods applied but also for the interpretation and communication of results. Indeed, scientists are encouraged to use precise terminology for population-specific reference samples and avoid broad generalizations (Hunt & Megyesi, 2008).

## 6 | CONCLUSIONS

The field of genetic epidemiology is undergoing a transformation with the collection of data with more diverse ancestral and social origins and with research staff and participants that are gradually becoming more inclusive. This shift is fuelled in part by concerted efforts toward the collection of diverse data at the study design level and as part of large reference data sets. We must be prepared to continue to support these efforts by engaging financially and scientifically in the years ahead to achieve broad data collections that will have meaningful impacts. The benefits of this shift toward more diversified genetic research include the potential for providing better treatment options to currently underrepresented communities, more equitable power dynamics with respect to health decisions and the potential for novel genetic discoveries that will be valuable to all communities, not just those where the discoveries are made.

To fully benefit from the growing admixed and diversified data collections, methodological tools should follow suit. We need to move beyond the traditional stratified datasets based on arbitrary population boundaries. We should support existing and future data analysts with documentation, training, and guidelines. As new models and approaches are being developed, our community should proceed with transparency following our hard-earned standards of methodological rigor. Concerted efforts to evaluate and compare the performance and limitations of the various methods being developed will help to document, improve and set standards for the conduct of analyses with diverse and admixed genetic data sets.

## ACKNOWLEDGMENTS

We are thankful for the support of the International Genetic Epidemiology Society for encouraging us to put this study together, with particular thanks to Heike Bickeböller and Peter Kraft. A. C. is funded by the German Research Foundation (FOR2488). F. T.-A. is funded by the intramural program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health. B. F. D. is funded by the National Cancer Institute of the National Institutes of Health (K99CA246063) and an award from the Achievement Rewards for College Scientists Foundation Los Angeles Founder Chapter. X. Wang is supported by the University of North Texas Foundation and Dr. Linda Truitt Creagh. R. S was funded by NIMH R21MH120739. D. B. is supported by the Australian Research Council under grant DP190103188. M.-P.D. funded by the Canada Research Chairs Program and by the Health Collaboration Acceleration Fund from the Government of Quebec.

## DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no data sets were generated or analysed during the current study.

## ORCID

*Burcu F. Darst* http://orcid.org/0000-0002-6205-4632
*Xuexia Wang* http://orcid.org/0000-0002-0613-2608
*Marie-Pierre Dubé* http://orcid.org/0000-0001-8442-4393

## REFERENCES

Abuabara, K., You, Y., Margolis, D. J., Hoffmann, T. J., Risch, N., & Jorgenson, E. (2020). Genetic ancestry does not explain increased atopic dermatitis susceptibility or worse disease control among African American subjects in 2 large US cohorts. *Journal of Allergy and Clinical Immunology*, *145*(1), 192–8.e11.

Ackerman, M. S., Johri, P., Spitze, K., Xu, S., Doak, T. G., Young, K., & Lynch, M. (2017). Estimating seven coefficients of pairwise relatedness using population-genomic data. *Genetics*, *206*(1), 105–118.

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246.

All of Us Research Program Investigaors, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., & Dishman, E. (2019). The "All of Us" Research Program. *New England Journal of Medicine*, *381*(7), 668–676.

Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., Ulirsch, J. C., Kamatani, Y., Okada, Y., Finucane, H. K., Koenen, K. C., Nievergelt, C. M., Daly, M. J., & Neale, B. M. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nature Genetics*, *53*(2), 195–204.

Baker, J. L., Rotimi, C. N., & Shriner, D. (2017). Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Scientific Reports*, *7*(1), 1572.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., & Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, *28*(10), 1359–1367.

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., Stahl, E. A., Huckins, L. M., Consortium, G. T., Nicolae, D. L., Cox, N. J., & Im, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, *9*(1), 1825.

Bentley, A. R., Callier, S. L., & Rotimi, C. N. (2020). Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ Genomic Medicine*, *5*, 5.

Bergstrom, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanche, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484), eaay5012.

Birney, E., Inouye, M., Raff, J., Rutherford, A., & Scally, A. (2021). The language of race, ethnicity, and ancestry in human genetic research. arXiv*;2106.10041*.

Bonham, V. L., & Green, E. D. (2021). The genomics workforce must become more diverse: A strategic imperative. *The American Journal of Human Genetics*, *108*(1), 3–7.

Borrell, L. N., Elhawary, J. R., Fuentes-Afflick, E., Witonsky, J., Bhakta, N., Wu, A. H. B., Bibbins-Domingo, K., Rodriguez-Santana, J. R., Lenoir, M. A., Gavin, J. R., 3rd, Kittles, R. A., Zaitlen, N. A., Wilkes, D. S., Powe, N. R., Ziv, E., & Burchard, E. G. (2021). Race and genetic ancestry in Medicine—A time for reckoning with racism. *New England Journal of Medicine*, *384*(5), 474–480.

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *American Journal of Human Genetics*, *96*(1), 37–53.

Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C. D., & Ostrer, H. (2010). Colloquium paper: Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(Suppl. 2), 8954–8961.

Bull, S., Farsides, B., & Tekola Ayele, F. (2012). Tailoring information provision and consent processes to research contexts: The value of rapid assessments. *Journal of Empirical Research on Human Research Ethics: JERHRE*, *7*(1), 37–52.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy-Gallego, E., The Human Genome Structural Variation Consortium, Flicek, P., Germer, S., Brand, H., Hall, I. M., … Zody, M. C. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv. 2021*:2021.02.06.430068*.

Caliebe, A., Leverkus, F., Antes, G., & Krawczak, M. (2019). Does big data require a methodological change in medical research? *BMC Medical Research Methodology*, *19*(1), 125.

Cavazos, T. B., & Witte, J. S. (2021). Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *HGG Advances*, *2*(1).

Chaichoompu, K., Abegaz, F., Tongsima, S., Shaw, P. J., Sakuntabhai, A., Pereira, L., & Van Steen, K. (2019). IPCAPS: An R package for iterative pruning to capture population structure. *Source Code for Biology and Medicine*, *14*, 2.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7.

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedon, J. C., Redline, S., Papanicolaou, G. J., Thornton, T. A., Laurie, C. C., Rice, K., & Lin, X. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *American Journal of Human Genetics*, *98*(4), 653–666.

Chen, W., Ren, C., Qin, H., Archer, K. J., Ouyang, W., Liu, N., Chen, X., Luo, X., Zhu, X., Sun, S., & Gao, G. (2015). A generalized sequential Bonferroni procedure for GWAS in admixed populations incorporating admixture mapping information into association tests. *Human Heredity*, *79*(2), 80–92.

Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, *8*(7), giz082.

Chrisman, B. S., Paskov, K. M., He, C., Jung, J. Y., Stockham, N., Washington, P. Y., & Wall, D. P. (2022). A method for localizing non-reference sequences to the human genome. *Pacific Symposium on Biocomputing*, *27*, 313–324.

Churchhouse, C., & Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology*, *37*(1), 1–12.

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics*, *98*(1), 127–148.

Conti, D. V., Darst, B. F., Moss, L. C., Saunders, E. J., Sheng, X., Chou, A., Schumacher, F. R., Olama, A. A. A., Benlloch, S., Dadaev, T., Brook, M. N., Sahimi, A., Hoffmann, T. J., Takahashi, A., Matsuda, K., Momozawa, Y., Fujita, M., Muir, K., Lophatananon, A., … Haiman, C. A. (2021). Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature Genetics*, *53*(1), 65–75.

Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L., & Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *American Journal of Human Genetics*, *101*(2), 218–226.

Crawford, N. G., Kelly, D. E., Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., Pfeifer, S. P., Jensen, J. D., Campbell, M. C., Beggs, W., Hormozdiari, F., Mpoloka, S. W., Mokone, G. G., Nyambo, T., Meskel, D. W., … Tishkoff, S. A. (2017). Loci associated with skin pigmentation identified in African populations. *Science*, *358*(6365).

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004.

De La Vega, F. M., & Bustamante, C. D. (2018). Polygenic risk scores: A biased prediction? *Genome Medicine*, *10*(1), 100.

Dikilitas, O., Schaid, D. J., Kosel, M. L., Carroll, R. J., Chute, C. G., Denny, J. A., Fedotov, A., Feng, Q., Hakonarson, H., Jarvik, G. P., Lee, M. T. M., Pacheco, J. A., Rowley, R., Sleiman, P. M., Stein, C. M., Sturm, A. C., Wei, W. Q., Wiesner, G. L., Williams, M. S., … Kullo, I. J. (2020). Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. *American Journal of Human Genetics*, *106*(5), 707–716.

Dimitromanolakis, A., Paterson, A. D., & Sun, L. (2019). Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data via TRUFFLE. *American Journal of Human Genetics*, *105*(1), 78–88.

Dou, J., Dou, H., Mu, C., Zhang, L., Li, Y., Wang, J., Li, T., Li, Y., Hu, X., Wang, S., & Bao, Z. (2017). Whole-genome restriction mapping by "Subhaploid"-based RAD sequencing: An efficient and flexible approach for physical mapping and genome scaffolding. *Genetics*, *206*(3), 1237–1250.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, *10*(1), 3328.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252.

Durvasula, A., & Sankararaman, S. (2020). Recovering signals of ghost archaic introgression in African populations. *Science Advances*, *6*(7), eaax5097.

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537).

Elliott, J., Bodinier, B., Bond, T. A., Chadeau-Hyam, M., Evangelou, E., Moons, K. G. M., Dehghan, A., Muller, D. C., Elliott, P., & Tzoulaki, I. (2020). Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. *Journal of the American Medical Association*, *323*(7), 636–645.

Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A. R., & Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. *Nature Medicine*, *28*(2), 243–250.

Fatumo, S., Yakubu, A., Oyedele, O., Popoola, J., Attipoe, D. A., Eze -Echesi, G., Modibbo, F. Z., Ado-Wanka, N., Osakwe, Y., Braimah, O., Julius-Enigimi, E., Akindigh, T. M., Kusimo, B., Akpulu, C., Nwuba, C., Ebong, O., Anyika, C., Adewunmi, O., Ibrahim, Y., … NCD-GHS Consortium. (2022). Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. *Nature Genetics*, *54*, 531–536.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Consortium, G. T., Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098.

Gay, N. R., Gloudemans, M., Antonio, M. L., Abell, N. S., Balliu, B., Park, Y., Martin, A. R., Musharoff, S., Rao, A. S., Aguet, F., Barbeira, A. N., Bonazzola, R., Hormozdiari, F., Consortium GT, Ardlie, K. G., Brown, C. D., Im, H. K., Lappalainen, T., Wen, X., & Montgomery, S. B. (2020). Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biology*, *21*(1), 233.

Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., Bowden, D. W., Langefeld, C. D., Oleksyk, T. K., Uscinski Knob, A. L., Bernhardy, A. J., Hicks, P. J., Nelson, G. W., Vanhollebeke, B., Winkler, C. A., Kopp, J. B., Pays, E., & Pollak, M. R. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, *329*(5993), 841–845.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383.

Gogarten, S. M., Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., Rice, K. M., & Conomos, M. P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, *35*(24), 5346–5348.

Graffelman, J., Balding, D. J., Gonzalez-Neira, A., & Bertranpetit, J. (2007). Variation in estimated recombination rates across human populations. *Human Genetics*, *122*(3-4), 301–310.

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., … Sandhu, M. S. (2015). The African genome variation project shapes medical genetics in Africa. *Nature*, *517*(7534), 327–332.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimaki, T., Raitoharju, E., Kahonen, M., Seppala, I., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), 245–252.

H3Africa, Consortium, Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V. M., Adebamowo, C., Adebiyi, E., Ademola, A. D., Adeyemo, A., Adu, D., Affolabi, D., Agongo, G., Ajayi, S., Akarolo-Anthony, S., Akinyemi, R., Akpalu, A., Alberts, M., Alonso Betancourt, O., Alzohairy, A. M., … Zar, H. (2014). Research capacity. Enabling the genomic revolution in Africa. *Science*, *344*(6190), 1346–1348.

Han, B., & Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American Journal of Human Genetics*, *88*(5), 586–598.

Hanghoj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, *8*(5).

Hauberg, M. E., Zhang, W., Giambartolomei, C., Franzen, O., Morris, D. L., Vyse, T. J., Ruusalepp, A., CommonMind, C., Sklar, P., Schadt, E. E., Bjorkegren, J. L. M., & Roussos, P. (2017). Large-scale identification of common trait and disease variants affecting gene expression. *American Journal of Human Genetics*, *100*(6), 885–894.

Helgason, A., Palsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I., Benediktsson, R., Hinney, A., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Schafer, H., Faruque, M., Doumatey, A., … Stefansson, K. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nature Genetics*, *39*(2), 218–225.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751.

Heyer, E., Brazier, L., Segurel, L., Hegay, T., Austerlitz, F., Quintana-Murci, L., Georges, M., Pasquet, P., & Veuille, M. (2011). Lactase persistence in central Asia: phenotype, genotype, and evolution. *Human Biology*, *83*(3), 379–392.

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., & Green, E. D. (2018). Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, *19*(3), 175–185.

Hoffmann, T. J., Ehret, G. B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P. Y., Iribarren, C., Chakravarti, A., & Risch, N. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature Genetics*, *49*(1), 54–64.

Hormozdiari, F., van de Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *American Journal of Human Genetics*, *99*(6), 1245–1260.

Horowitz, C. R., Ferryman, K., Negron, R., Sabin, T., Rodriguez, M., Zinberg, R. F., Bottinger, E., & Robinson, M. (2017). Race, genomics and chronic disease: What patients with African ancestry have to say. *Journal of Health Care for the Poor and Underserved*, *28*(1), 248–260.

Hu, Y., Willer, C., Zhan, X., Kang, H. M., & Abecasis, G. R. (2013). Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *American Journal of Human Genetics*, *93*(5), 891–899.

Hunt, L. M., & Megyesi, M. S. (2008). Genes, race and research ethics: Who's minding the store? *Journal of Medical Ethics*, *34*(6), 495–500.

IGES_ELSI_Committee. (2021). *Short guideline on the use and reporting of race/ethnicity/ancestry*. International Genetic Epidemiology Society (IGES). https://iges.memberclicks.net/reporting-race-ethnicity-ancestry

International HapMap Consortium. (2004). Integrating ethics and science in the international HapMap project. *Nature Reviews Genetics*, *5*(6), 467–475.

Janssens, A. (2019). Validity of polygenic risk scores: Are we measuring what we think we are? *Human Molecular Genetics*, *28*(R2), R143–R150.

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), 1749–1755.

Johnson, R. C., Nelson, G. W., Zagury, J. F., & Winkler, C. A. (2015). ALDsuite: Dense marker MALD using principal components of ancestral linkage disequilibrium. *BMC Genetics*, *16*, 23.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354.

Kaplan, A., Lock, E. F., & Fiecas, M., Alzheimer's Disease Neuroimaging Initiative. (2020). Bayesian GWAS with structured and non-local priors. *Bioinformatics*, *36*(1), 17–25.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443.

Kehdy, F. S., Gouveia, M. H., Machado, M., Magalhaes, W. C., Horimoto, A. R., Horta, B. L., Moreira, R. G., Leal, T. P., Scliar, M. O., Soares-Souza, G. B., Rodrigues-Soares, F., Araujo, G. S., Zamudio, R., Sant Anna, H. P., Santos, H. C., Duarte, N. E., Fiaccone, R. L., Figueiredo, C. A., Silva, T. M., … Brazilian, E. P. C. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(28), 8696–8701.

Keys, K. L., Mak, A. C. Y., White, M. J., Eckalbar, W. L., Dahl, A. W., Mefford, J., Mikhaylova, A. V., Contreras, M. G., Elhawary, J. R., Eng, C., Hu, D., Huntsman, S., Oh, S. S., Salazar, S., Lenoir, M. A., Ye, J. C., Thornton, T. A., Zaitlen, N., Burchard, E. G., & Gignoux, C. R. (2020). On the cross-population generalizability of gene expression prediction models. *PLoS Genetics*, *16*(8), e1008927.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224.

Kichaev, G., & Pasaniuc, B. (2015). Leveraging Functional-annotation data in trans-ethnic fine-mapping studies. *American Journal of Human Genetics*, *97*(2), 260–271.

Kirkpatrick, B., Ge, S., & Wang, L. (2019). Efficient computation of the kinship coefficients. *Bioinformatics*, *35*(6), 1002–1008.

Knecht, C., & Krawczak, M. (2014). Molecular genetic epidemiology of human diseases: From patterns to predictions. *Human Genetics*, *133*(4), 425–430.

Knowles, J. W., & Ashley, E. A. (2018). Cardiovascular disease: The rise of the genetic risk score. *PLoS Medicine*, *15*(3), e1002546.

Kodaman, N., Aldrich, M. C., Smith, J. R., Signorello, L. B., Bradley, K., Breyer, J., Cohen, S. S., Long, J., Cai, Q., Giles, J., Bush, W. S., Blot, W. J., Matthews, C. E., & Williams, S. M. (2013). A small number of candidate gene SNPs reveal continental ancestry in African Americans. *Annals of Human Genetics*, *77*(1), 56–66.

Koehl, A. J., & Long, J. C. (2018). The contributions of admixture and genetic drift to diversity among post-contact populations in the Americas. *American Journal of Physical Anthropology*, *165*(2), 256–268.

Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., Jain, D., Argos, M., Arnett, D. K., Avery, C., Barnes, K. C., Becker, L. C., Bien, S. A., Bis, J. C., Blangero, J., Boerwinkle, E., Bowden, D. W., Buyske, S., Cai, J., … Li, Y. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genetics*, *15*(12), e1008500.

Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack of diversity in genomic databases is A barrier to translating precision Medicine research into practice. *Health Affairs*, *37*(5), 780–785.

Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., … Kayser, M. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology*, *18*(16), 1241–1248.

LaPierre, N., Taraszka, K., Huang, H., He, R., Hormozdiari, F., & Eskin, E. (2021). Identifying causal variants by fine mapping across multiple studies. *PLoS Genetics*, *17*(9), e1009733.

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. *Cell*, *177*(1), 70–84.

Lawson, D. J., Davies, N. M., Haworth, S., Ashraf, B., Howe, L., Crawford, A., Hemani, G., Davey Smith, G., & Timpson, N. J. (2020). Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics*, *139*(1), 23–41.

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 3258.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, *8*(1), e1002453.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prufer, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409–413.

Lebeko, K., Sloan-Heggen, C. M., Noubiap, J. J., Dandara, C., Kolbe, D. L., Ephraim, S. S., Booth, K. T., Azaiez, H., Santos-Cortez, R. L., Leal, S. M., Smith, R. J., & Wonkam, A. (2016). Targeted genomic enrichment and massively parallel sequencing identifies novel nonsyndromic hearing impairment pathogenic variants in Cameroonian families. *Clinical Genetics*, *90*(3), 288–290.

Lee, L. V., Munoz, E. L., Tan, K. T., & Reyes, M. T. (2001). Sex linked recessive dystonia parkinsonism of Panay, Philippines (XDP). *Molecular Pathology*, *54*(6), 362–368.

Legge, S. E., Pardinas, A. F., Helthuis, M., Jansen, J. A., Jollie, K., Knapper, S., MacCabe, J. H., Rujescu, D., Collier, D. A., O'Donovan, M. C., Owen, M. J., & Walters, J. T. R. (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. *Molecular Psychiatry*, *24*(3), 328–337.

Lin, D. Y., Tao, R., Kalsbeek, W. D., Zeng, D., Gonzalez, F., 2nd, Fernandez-Rhodes, L., Graff, M., Koch, G. G., North, K. E., & Heiss, G. (2014). Genetic association analysis under complex survey sampling: The Hispanic Community Health Study/Study of Latinos. *American Journal of Human Genetics*, *95*(6), 675–688.

Lin, D. Y., & Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology*, *34*(1), 60–66.

Liu, J., Lewinger, J. P., Gilliland, F. D., Gauderman, W. J., & Conti, D. V. (2013). Confounding and heterogeneity in genetic association studies with admixed populations. *American Journal of Epidemiology*, *177*(4), 351–360.

Lock, E. F., & Dunson, D. B. (2017). Bayesian genome- and epigenome-wide association studies with gene level dependence. *Biometrics*, *73*(3), 1018–1028.

Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P., & Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature Genetics*, *50*(7), 906–908.

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290.

Magi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., & McCarthy, M. I. (2017). Cogent-Kidney Consortium TDGC, Morris AP. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Human Molecular Genetics*, *26*(18), 3639–3650.

Magi, R., & Morris, A. P. (2010). GWAMA: Software for genome-wide association meta-analysis. *BMC Bioinformatics*, *11*, 288.

Mahajan, A., Spracklen, C. N., Zhang, W., Ng, M. C. Y., Petty, L. E., Kitajima, H., Yu, G. Z., Rüeger, S., Speidel, L., Kim, Y. J., Horikoshi, M., Mercader, J. M., Taliun, D., Moon, S., Kwak, S.-H., Robertson, N. R., Rayner, N. W., Loh, M., Kim, B.-J., … FinnGen, e M. C. (2022). Multi-ancestry genetic study of type

2 diabetes highlights the power of diverse populations for discovery and translation. *Nature Genetics*, *54*(5), 560–572.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., & Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, *41*(6), 469–480.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873.

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., & Kohane, I. S. (2016). Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, *375*(7), 655–665.

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

Marcheco-Teruel, B., Parra, E. J., Fuentes-Smith, E., Salas, A., Buttenschon, H. N., Demontis, D., Torres-Espanol, M., Marin-Padron, L. C., Gomez-Cabezas, E. J., Alvarez-Iglesias, V., Mosquera-Miguel, A., Martinez-Fuentes, A., Carracedo, A., Borglum, A. D., & Mors, O. (2014). Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. *PLoS Genetics*, *10*(7), e1004488.

Marnetto, D., Parna, K., Lall, K., Molinaro, L., Montinaro, F., Haller, T., Metspalu, M., Magi, R., Fischer, K., & Pagani, L. (2020). Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nature Communications*, *11*(1), 1628.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *American Journal of Human Genetics*, *100*(4), 635–649.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*(4), 584–591.

Martin, E. R., Tunc, I., Liu, Z., Slifer, S. H., Beecham, A. H., & Beecham, G. W. (2018). Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genetic Epidemiology*, *42*(2), 214–229.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), 1097–1103.

Mersha, T. B. (2015). Mapping asthma-associated variants in admixed populations. *Frontiers in Genetics*, *6*, 292.

Micheletti, S. J., Bryc, K., Ancona Esselmann, S. G., Freyman, W. A., Moreno, M. E., Poznik, G. D., Shastri, A. J., 23andMe Research Team, Beleza, S., & Mountain, J. L. (2020). Genetic consequences of the transatlantic slave trade in the mericas. *American Journal of Human Genetics*, *107*(2), 265–277.

Mikhaylova, A. V., & Thornton, T. A. (2019). Accuracy of gene expression prediction from genotype data with PrediXcan varies across and within continental populations. *Frontiers in Genetics*, *10*, 261.

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., Johnson, W. C., Im, H. K., Liu, Y., & Wheeler, H. E. (2018). Genetic architecture of gene expression traits across diverse populations. *PLoS Genetics*, *14*(8), e1007586.

Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology*, *35*(8), 809–822.

Naseri, A., Liu, X., Tang, K., Zhang, S., & Zhi, D. (2019). RaPID: Ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome Biology*, *20*(1), 143.

Nohr, A. K., Hanghoj, K., Erill, G. G., Li, Z., Moltke, I., & Albrechtsen, A. (2021). NGSremix: A software tool for estimating pairwise relatedness between admixed individuals from next-generation sequencing data. *G3*, *11*, jkab174.

Ochoa, A., & Storey, J. D. (2021). Estimating FST and kinship for arbitrary population structures. *PLoS Genetics*, *17*(1), e1009241.

Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics*, *5*, 204.

Palk, A. C., Dalvie, S., de Vries, J., Martin, A. R., & Stein, D. J. (2019). Potential use of clinical polygenic risk scores in psychiatry—Ethical implications and communicating high polygenic risk. *Philosophy, Ethics, and Humanities in Medicine: PEHM*, *14*(1), 4.

Parra, F. C., Amado, R. C., Lambertucci, J. R., Rocha, J., Antunes, C. M., & Pena, S. D. (2003). Color and genomic ancestry in Brazilians. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(1), 177–182.

Pasaniuc, B., & Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, *18*(2), 117–127.

Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. H., Ruczinski, I., Fornage, M., Siscovick, D. S., Zhu, X., Larkin, E., Lange, L. A., Cupples, L. A., Yang, Q., Akylbekova, E. L., Musani, S. K., Divers, J., Mychaleckyj, J., Li, M., … Price, A. L. (2011). Enhanced statistical tests for GWAS in admixed populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genetics*, *7*(4), e1001371.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J., & Reich, D. (2004). Methods for high-density admixture mapping of disease genes. *American Journal of Human Genetics*, *74*(5), 979–1000.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190.

Pavlides, J. M., Zhu, Z., Gratten, J., McRae, A. F., Wray, N. R., & Yang, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Medicine*, *8*(1), 84.

Penalba, J. V., & Wolf, J. B. W. (2020). From molecules to populations: Appreciating and estimating recombination rate variation. *Nature Reviews Genetics*, *21*(8), 476–492.

Peoples, J. G., & Bailey, G. (2018). *Humanity: An introduction to cultural anthropology*. Wadsworth Publishing.

Petersen, D. C., Libiger, O., Tindall, E. A., Hardie, R. A., Hannick, L. I., Glashoff, R. H., Mukerji, M., Indian Genome Variation Consortium, Fernandez, P., Haacke, W., Schork, N. J., & Hayes, V. M. (2013). Complex patterns of genomic admixture within southern Africa. *PLoS Genetics*, *9*(3), e1003309.

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C. Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., … Duncan, L. E. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell*, *179*(3), 589–603.

Polygenic Risk Score Task Force of the International Common Disease Alliance. (2021). Responsible use of polygenic risk scores in the clinic: Potential benefits, risks and gaps. *Nature Medicine*, *27*(11), 1876–1884.

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164.

Popejoy, A. B., Ritter, D. I., Crooks, K., Currey, E., Fullerton, S. M., Hindorff, L. A., Koenig, B., Ramos, E. M., Sorokin, E. P., Wand, H., Wright, M. W., Zou, J., Gignoux, C. R., Bonham, V. L., Plon, S. E., Bustamante, C. D., & Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG). (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Human Mutation*, *39*(11), 1713–1720.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

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

Race, Ethnicity, and Genetics Working Group. (2005). The use of racial, ethnic, and ancestral categories in human genetics research. *American Journal of Human Genetics*, *77*(4), 519–532.

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Mezey, J. G., & Williams, A. L. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, *207*(1), 75–82.

Risch, N., Burchard, E., Ziv, E., & Tang, H. (2002). Categorization of humans in biomedical research: Genes, race and disease. *Genome Biology*, *3*(7), comment2007.

Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela, R., Rodriguez-Santana, J. R., Rodriguez-Cintron, W., Avila, P. C., Ziv, E., & Gonzalez Burchard, E. (2009). Ancestry-related assortative mating in Latino populations. *Genome Biology*, *10*(11), R132.

Rosenberg, N. A. (2011). A population-genetic perspective on the similarities and differences among worldwide human populations. *Human Biology*, *83*(6), 659–684.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, *11*(5), 356–366.

Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., Ishigaki, K., Suzuki, A., Suzuki, K., Obara, W., Yamaji, K., Takahashi, K., Asai, S., Takahashi, Y., Suzuki, T., … Okada, Y. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics*, *53*(10), 1415–1424.

Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., Patterson, N., & Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, *507*(7492), 354–357.

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491–504.

Sebro, R., Hoffman, T. J., Lange, C., Rogus, J. J., & Risch, N. J. (2010). Testing for non-random mating: Evidence for ancestry-related assortative mating in the Framingham heart study. *Genetic Epidemiology*, *34*(7), 674–679.

Segurel, L., & Bon, C. (2017). On the evolution of lactase persistence in humans. *Annual Review of Genomics and Human Genetics*, *18*, 297–319.

Segurel, L., Wyman, M. J., & Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. *Annual Review of Genomics and Human Genetics*, *15*, 47–70.

Seidman, D. N., Shenoy, S. A., Kim, M., Babu, R., Woods, I. G., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., & Williams, A. L. (2020). Rapid, phase-free detection of long Identity-by-descent segments enables effective relationship classification. *American Journal of Human Genetics*, *106*(4), 453–466.

Seldin, M. F., Pasaniuc, B., & Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, *12*(8), 523–528.

Shi, J., & Lee, S. (2016). A novel random effect model for GWAS meta-analysis and its application to trans-ethnic meta-analysis. *Biometrics*, *72*(3), 945–954.

Shriner, D. (2017). Overview of admixture mapping. *Current Protocols in Human Genetics/Editorial Board, Jonathan L. Haines. [et al.]*, *94*, 1.23.1–1.23.8.

Shriner, D., Adeyemo, A., & Rotimi, C. N. (2011). Joint ancestry and association testing in admixed individuals. *PLoS Computational Biology*, *7*(12), e1002325.

Shriner, D., Tekola-Ayele, F., Adeyemo, A., & Rotimi, C. N. (2014). Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. *Scientific Reports*, *4*, 6055.

Shriner, D., Tekola-Ayele, F., Adeyemo, A., & Rotimi, C. N. (2016). Ancient human migration after out-of-Africa. *Scientific Reports*, *6*, 26565.

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(4), 1080.

Skoglund, P., & Mathieson, I. (2018). Ancient genomics of modern humans: The first decade. *Annual Review of Genomics and Human Genetics*, *19*, 381–404.

Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nature Reviews Genetics*, *16*(1), 33–44.

Sternthal, M. J., Slopen, N., & Williams, D. R. (2011). RACIAL DISPARITIES IN HEALTH: How much does stress really matter? *Du Bois Review*, *8*(1), 95–113.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S. B., Tian, X., Browning, B. L., Das, S., Emde, A. K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299.

Tehranchi, A., Hie, B., Dacre, M., Kaplow, I., Pettie, K., Combs, P., & Fraser, H. B. (2019). Fine-mapping cis-regulatory variants in diverse human populations. *eLife*, *8*.

Tekola, F., Bull, S., Farsides, B., Newport, M. J., Adeyemo, A., Rotimi, C. N., & Davey, G. (2009a). Impact of social stigma on the process of obtaining informed consent for genetic research on podoconiosis: A qualitative study. *BMC Medical Ethics*, *10*, 13.

Tekola, F., Bull, S. J., Farsides, B., Newport, M. J., Adeyemo, A., Rotimi, C. N., & Davey, G. (2009b). Tailoring consent to context: Designing an appropriate consent process for a biomedical study in a low income setting. *PLoS Neglected Tropical Diseases*, *3*(7), e482.

Tekola-Ayele, F., Adeyemo, A., Finan, C., Hailu, E., Sinnott, P., Burlinson, N. D., Aseffa, A., Rotimi, C. N., Newport, M. J., & Davey, G. (2012). HLA class II locus and susceptibility to podoconiosis. *New England Journal of Medicine*, *366*(13), 1200–1208.

Tekola-Ayele, F., Adeyemo, A., Chen, G., Hailu, E., Aseffa, A., Davey, G., Newport, M. J., & Rotimi, C. N. (2015). Novel genomic signals of recent selection in an Ethiopian population. *European Journal of Human Genetics*, *23*(8), 1085–1092.

Thareja, G., Al-Sarraj, Y., Belkadi, A., Almotawa, M., Qatar Genome Program Research Consortium, Suhre, K., & Albagha, O. M. E. (2021). Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. *Nature Communications*, *12*(1), 1250.

The 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*(2), 301–326.

Thornton, T. A., & Bermejo, J. L. (2014). Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genetic Epidemiology*, *38*(Suppl. 1), S5–S12.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, *324*(5930), 1035–1044.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., & Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, *39*(1), 31–40.

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, *19*(9), 581–590.

Uren, C., Hoal, E. G., & Moller, M. (2020). Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genetics*, *21*(1), 40.

Vilhjalmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindstrom, S., Ripke, S., Genovese, G., Loh, P. R., Bhatia, G., Do, R., Hayeck, T., Won, H. H., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) Study, Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., … Price, A. L. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics*, *97*(4), 576–592.

Wall, J. D., Stawiski, E. W., Ratan, A., Kim, H. L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E. S., Purbojati, R. W., Bhangale, T., Stepanov, V., Kharkov, V., Schröder, M. S., Ramprasad, V., Tom, J., Durinck, S., Bei, Q., Li, J., Guillory, J., … GenomeAsia KC (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*, *576*(7785), 106–111.

Wall, J. D., Yang, M. A., Jay, F., Kim, S. K., Durand, E. Y., Stevison, L. S., Gignoux, C., Woerner, A., Hammer, M. F., & Slatkin, M. (2013). Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics*, *194*(1), 199–209.

Wand, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., Kullo, I. J., Rowley, R., Dron, J. S., Brockman, D., Venner, E., McCarthy, M. I., Antoniou, A. C., Easton, D. F., Hegele, R. A., Khera, A. V., Chatterjee, N., Kooperberg, C., Edwards, K., … Wojcik, G. L. (2021). Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, *591*(7849), 211–219.

Wang, J. (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, *107*, 4–13.

Wang, M., Menon, R., Mishra, S., Patel, A. P., Chaffin, M., Tanneeru, D., Deshmukh, M., Mathew, O., Apte, S., Devanboo, C. S., Sundaram, S., Lakshmipathy, P., Murugan, S., Sharma, K. K., Rajendran, K., Santhosh, S., Thachathodiyl, R., Ahamed, H., Balegadde, A. V., … Khera, A. V. (2020). Validation of a genome-wide polygenic score for coronary artery disease in South Asians. *Journal of the American College of Cardiology*, *76*(6), 703–714.

Wang, M. H., Cordell, H. J., & Van Steen, K. (2019). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*, *55*, 53–60.

Wegmann, D., Kessner, D. E., Veeramah, K. R., Mathias, R. A., Nicolae, D. L., Yanek, L. R., Sun, Y. V., Torgerson, D. G., Rafaels, N., Mosley, T., Becker, L. C., Ruczinski, I., Beaty, T. H., Kardia, S. L. R., Meyers, D. A., Barnes, K. C., Becker, D. M., Freimer, N. B., & Novembre, J. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics*, *43*(9), 847–853.

Wiley, A. S. (2020). Lactose intolerance. *Evolution, Medicine, and Public Health*, *2020*(1), 47–48.

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190–2191.

William, A., & David, J. B. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, *24*(4), 451–471.

Williams, D. R., Lawrence, J. A., Davis, B. A., & Vu, C. (2019). Understanding how discrimination can affect health. *Health Services Research*, *54*(Suppl. 2), 1374–1388.

Witherspoon, D. J., Wooding, S., Rogers, A. R., Marchani, E. E., Watkins, W. S., Batzer, M. A., & Jorde, L. B. (2007). Genetic similarities within and between human populations. *Genetics*, *176*(1), 351–359.

Wojcik, G. L., Fuchsberger, C., Taliun, D., Welch, R., Martin, A. R., Shringarpure, S., Carlson, C. S., Abecasis, G., Kang, H. M., Boehnke, M., Bustamante, C. D., Gignoux, C. R., & Kenny, E. E. (2018). Imputation-Aware Tag SNP selection to improve power for large-scale, multi-ethnic association studies. *G3*, *8*(10), 3255–G67.

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., … Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, *570*(7762), 514–518.

Wong, K. H. Y., Ma, W., Wei, C. Y., Yeh, E. C., Lin, W. J., Wang, E. H. F., Su, J. P., Hsieh, F. J., Kao, H. J., Chen, H. H., Chow, S. K., Young, E., Chu, C., Poon, A., Yang, C. F., Lin, D. S., Hu, Y. F., Wu, J. Y., Lee, N. C., … Kwok, P. Y. (2020). Towards a reference genome that captures global genetic diversity. *Nature Communications*, *11*(1), 5482.

Wonkam, A., & de Vries, J. (2020). Returning incidental findings in African genomics research. *Nature Genetics*, *52*(1), 17–20.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *55*(10), 1068–1087.

Xu, D., Pavlidis, P., Taskent, R. O., Alachiotis, N., Flanagan, C., DeGiorgio, M., Blekhman, R., Ruhl, S., & Gokcumen, O. (2017). Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation. *Molecular Biology and Evolution*, *34*(10), 2704–2715.

Yu, N., Chen, F. C., Ota, S., Jorde, L. B., Pamilo, P., Patthy, L., Ramsay, M., Jenkins, T., Shyue, S. K., & Li, W. H. (2002). Larger genetic differences within Africans than between Africans and Eurasians. *Genetics*, *161*(1), 269–274.

Yudell, M., Roberts, D., DeSalle, R., Tishkoff, S., & SCIENCE AND SOCIETY. (2016). Taking race out of human genetics. *Science*, *351*(6273), 564–565.

Yusuf, A. A., Govender, M. A., Brandenburg, J. T., & Winkler, C. A. (2021). Kidney disease and APOL1. *Human Molecular Genetics*, *30*(R1), R129–R137.

Zhang, J. (2010). Ancestral informative marker selection and population structure visualization using sparse Laplacian eigenfunctions. *PLoS One*, *5*(11), e13734.

Zhang, J., & Stram, D. O. (2014). The role of local ancestry adjustment in association studies using admixed populations. *Genetic Epidemiology*, *38*(6), 502–515.

Zhang, Y., & Pan, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples: Competitors or complements? *Genetic Epidemiology*, *39*(3), 149–155.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W. Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), 1335–1341.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487.

**How to cite this article:** Caliebe, A., Tekola-Ayele, F., Darst, B. F., Wang, X., Song, Y. E., Gui, J., Sebro, R. A., Balding, D. J., Saad, M., Dubé, M.-P., & On behalf of the IGES ELSI Committee. (2022). Including diverse and admixed populations in genetic epidemiology research. *Genetic Epidemiology*, 46, 347–371. https://doi.org/10.1002/gepi.22492