Methodological opportunities in genomic data analysis to advance health equity

Brieuc Lehmann[1†], Leandra Bräuninger[1,5], Yoonsu Cho[3,4], Fabian Falck[2,5], Smera Jayadeva[5], Michael Katell[5], Thuy Nguyen[3], Antonella Perini[5], Sam Tallman[3], Maxine Mackintosh[3,5], Matt Silver[3,6], Karoline Kuchenbäcker[7,10], David Leslie[5, 1], Nilanjan Chatterjee[8, 10] & Chris Holmes[2,9, 10]

[1]Department of Statistical Science, University College London, London, United Kingdom.

[2]Department of Statistics, University of Oxford, Oxford, United Kingdom.

[3]Genomics England, London, United Kingdom.

[4]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom.

[5]The Alan Turing Institute, London, United Kingdom.

[6]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Fajara, Banjul, The Gambia.

[7]Division of Psychiatry, University College London, London, United Kingdom.

[8]Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA.

[9]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom.

[10]These authors contributed equally: Karoline Kuchenbäcker, David Leslie, Nilanjan Chatterjee, Chris Holmes

[†]e-mail: b.lehmann@ucl.ac.uk

Abstract | The roots and consequences of inequities in genomic research and medicine are complex and widespread. Efforts to improve diversity in the field are ongoing; however, an often overlooked source of inequity is the choice of analytical methods used to process, analyse and interpret genomic data. New statistical and machine learning techniques to understand, quantify and correct for the impact of biases in genomic data are emerging within the wider genomic research and genomic medicine ecosystem. At this crucial time point, it is important to clarify where improvements in methods and practices can, and cannot, play a role in improving equity in genomics. In this Perspective, we review existing approaches to promote equity and fairness in statistical analysis for genomics, and propose future methodological developments that are likely to yield the most impact for equity in genomics.

[H1] Introduction

For many years, concerns have been raised at the lack of diversity in genetic studies. For instance, genome-wide association studies (GWAS) are disproportionately comprised of individuals of European genetic ancestries[1–4], and large cancer genomics repositories, including The Cancer Genome Atlas (TCGA) exhibit similar stark disparities[5]. Despite repeated calls to improve the representativeness of genetic datasets, the proportion of GWAS conducted in individuals of European genetic ancestries has instead been increasing[2,4]. The lack of genomic data diversity is compounded by wider factors related to the sociopolitical system in which genomic research takes place, including the underrepresentation of genomic scientists from diverse backgrounds[6] and concerns from historically under-represented groups over data privacy and misuse[7]. Combined, these factors can limit genomics' role in achieving health equity, defined by the World Health Organisation as "the absence of unfair, avoidable or remediable differences among groups of people"[157].

Another, often-overlooked source of inequity is the choice of analytical methods used to process, analyse and interpret genomic data. Statistical models do not always adequately account for variation within the study sample, and often make assumptions of homogeneity across individuals. One consequence is that, to meet a model's assumptions, researchers elect to exclude minority groups of individuals from analyses entirely[8]. Another possibility is that important differences between the study sample and the target population are overlooked, threatening the validity and generalizability of scientific conclusions. For example, UK Biobank exhibits a 'healthy-volunteer' bias[9,10], whereby participants tend to be healthier than the general UK population. Such participation bias in UK Biobank has been shown to influence the results of GWAS across a range of traits[11].

As the use of genetic data for clinical decision-making expands in healthcare systems, the lack of diversity has the potential to exacerbate health inequalities - observable differences in health outcomes between population subgroups. For example, screening using polygenic scores (PGSs), estimates of disease risk based on genetic variation, can be less accurate for individuals of non-European genetic ancestries, especially for individuals of African genetic ancestries[12–14]. This may limit the ability to identify high-risk individuals in certain ancestry groups and deliver optimal care for them[15,16]. Meanwhile, disparities in variant effect misclassification rates for Mendelian disorders have been, in part, attributed to the lack of diversity in genetic studies[17–19]. This lack of generalizability of findings has the potential to amplify existing disparities for minoritised groups throughout the healthcare system.

Whilst there are many initiatives underway to address the urgent lack of diversity in genomic datasets[20–22], there remains a need to deal with imperfect data and models. There has been substantial interest in developing new statistical techniques to understand, quantify, and correct for the impact of existing biases. To date, this interest has largely been focused on statistical techniques for GWAS and PGS, which have been the

subject of several recent reviews[23–25]. While these efforts are directed to tackle technical challenges specific to a particular analysis task, they also reveal a variety of strategies to address biases in genomic data analysis more generally.

In this Perspective, we highlight the myriad ways in which statistical methods can serve health equity in genomic data analysis. Our focus is on the guiding principles applied to promote equity, rather than on the specific methods themselves. To place the role of analytical approaches within the wider genomic research and genomic medicine ecosystem, we propose a conceptual genomic data analysis framework that clarifies where improvements in analytical methods can (and cannot) improve equity in genomics. We also identify and synthesize existing strategies used to promote equity. Finally, we propose candidates for the most salient methodological gaps to spur future methodological developments likely to yield the most impact for equity.

[H1] A genomic data analysis framework

Recent efforts have sought to identify the potential impact of statistics and machine learning in equitable healthcare[26,27]. While genomics research can greatly benefit from these advancements, a tailored approach is essential due to the unique complexities of genomic data. We propose a bespoke conceptual framework to characterize the steps involved in the design, execution and deployment of genomic analysis models (**Fig. 1**). Here, we outline each stage of the framework and identify ways in which bias can enter each one. By bias, we refer to any process or context which can lead to results that differ systematically from the truth. The framework also outlines the key aspects of the sociopolitical ecosystem within which genomic research takes place (see Box 1). These aspects can directly or indirectly influence each stage of a genomic data analysis with substantial consequences for equity. Importantly, this highlights where methodological innovations do not have a role to play; better methods cannot address sociopolitical challenges at the ecosystem level. Recommendations on how to tackle such challenges have been discussed in detail elsewhere[6,45,46].

[H2] Research design and data acquisition

At a study's outset, researchers must make key design decisions and obtain the required data. Studies involving primary data collection must recruit and then collect data from participants, while secondary data analyses gather information from existing datasets.

[H3] Study design

Key considerations in study design include the sample size needed to offer sufficient statistical power, the characteristics of participants and which variables to measure, for example, the choice of genotyping array. Statistical theory and simulations can inform the design of a study to enhance its validity and reliability. The use

of appropriate population descriptors to characterize variation within the population is particularly important in ensuring the validity of scientific conclusions[28]. We use the term 'population' to refer to a group of individuals with a common attribute or perceived characteristic. Examples include, but are not limited to, geographical location, ethnicity and genetic ancestry (**Box 1**).

[H3] Study participation

For primary data collection, participants are recruited from a particular target population. Individuals in the target population, however, may not be equally likely to participate, resulting in an unrepresentative sample. For example, participants in volunteer-based studies such as the UK Biobank tend to be healthier and better educated than the general population[10,11]. Trust in genomics is another important factor (see Box 2). Statistical methods can have little direct influence in boosting study participation.

[H3] Data collection and availability

The data collection process may also introduce biases. Cultural preferences, for instance for saliva over blood samples[158], can potentially lead to uneven data quality between groups[159]. Unequal participant retention can lead to bias owing to higher dropout rates from individuals in minoritised groups[29]. Studies that rely on existing genomic datasets are susceptible to their lack of diversity[2,4], In particular, the lack of suitable reference datasets can inhibit genomic analyses in under-represented groups. This is compounded by challenges around data sharing, related to legitimate concerns around data privacy and data sovereignty, alongside technical issues associated with data storage and transfer[30], which methodological innovations may help to resolve[160].

[H2] Data preparation

Preparing data for analysis can involve a variety of steps depending on the type of study, but often include mapping, quality control (QC) and various forms of preprocessing. Although often considered value-neutral technical approaches, these steps can nevertheless introduce bias in ways that are often overlooked.

[H3] Mapping

Mapping involves aligning sequencing reads to a reference genome to determine their position within the genome. The current most widely-used human reference genome, GRCh38, is a mosaic of sub-sequences derived from a relatively small number of individuals, with a single individual contributing approximately 70% of the reference[31]. This reliance on a single, linear reference genome can lead to errors in identifying genetic variants, a form of reference bias, as the reads may not align correctly or may align to incorrect locations in the

genome[32]. Statistical methods can serve to identify and mitigate such errors to prevent them from biasing downstream analyses.

[H3] QC

QC aims to eliminate low-quality data and minimize technical artefacts that can lead to spurious statistical conclusions. QC procedures are typically highly specialized to the specific data collection technology, generally based on statistical or scientific theory that assumes a homogeneous sample population. For example, in GWAS, a variety of QC metrics are based on allele frequency estimates. In a diverse sample, applying these metrics without accounting for population structure may lead to variants and samples being incorrectly flagged as errors or outliers, particularly from under-represented groups[24].

[H3] Preprocessing

A variety of further preprocessing steps may be undertaken before analysis. A common example is imputation following array genotyping or low-pass sequencing, whereby missing genotypes are inferred based on the observed data. This process relies on a reference panel, a collection of known haplotypes in a particular population. Currently available reference panels lack diversity, which can lead to biased imputation for under-represented populations. For example, when a genetic variant is common in a group of non-European individuals but rare or absent in the reference panel, the imputation process might fail to accurately predict the presence of that variant[33]. This can result in incomplete or incorrect genetic data along axes of disparity, reducing study accuracy and potentially overlooking important disease associations[12].

[H2] Model development

Next, a model is developed that seeks to capture the particular characteristics of the dataset. This comprises several interrelated steps including feature engineering, model specification and model training.

[H3] Feature engineering

Most models do not operate directly on raw data, and rather take as input specific transformations, or 'features' of the data. Such transformation of data is known in the machine learning literature as feature engineering. This can overlap substantially with the preprocessing step described in the previous *Data preparation* stage; imputation, for example, can be seen as a form of feature engineering. A common form of feature engineering is dimensionality reduction, for instance the use of principal components analysis (PCA) to correct for population structure in GWAS and other settings. Other techniques, such as uniform manifold approximation and projection

(UMAP), are often used in single-cell genomics to cluster cell types[34] and have also been applied in tandem with PCA to summarise human genetic variation. UMAP excels at preserving local structures and identifying fine-scale patterns but distorts global distances, making geometric relationships between clusters non-meaningful in terms of genetic differentiation[161]. Another common but highly fraught application of dimensionality reduction occurs in the analysis of genetic ancestry groups (**Box 1** and discussed further under *More nuanced approaches to categorization*). Bias can occur if features are misinterpreted or derived with respect to unsuitable reference populations, which can negatively affect statistical performance and result in invalid scientific conclusions. For instance, overlaying self-reported ethnic groups onto UMAP to visualize genetic variation in the *All of Us* research programme[21] was criticized by researchers concerned that it risks reinforcing racist ideologies[35].

[H3] Model specification

Model specification refers to the choice of model type, the model's hyperparameters, and the selection of its functional form, that is, how features or variables interact within the model. Bias can enter here if the model's structure is based on data or biological knowledge derived predominantly from certain — typically European ancestry — populations, or on false conceptions of population differences (**Box 2**). For instance, gene-by-gene interactions or linkage disequilibrium (LD) structures that are more representative of individuals of European genetic ancestries might overshadow or inadequately account for genetic architectures more relevant to individuals of non-European genetic ancestrie[36]. Such mis-specified models produce less accurate results when applied to diverse populations[37].

[H3] Model fitting

Model fitting, or model training, is the process of estimating the model parameters that best align with the observed data. This typically includes obtaining measures of uncertainty around these estimates, used to draw statistical conclusions. Key choices include the training dataset, the loss function (that is, the measure of similarity between fitted model output and observed data) and often a procedure for model selection. Similarly to model mis-specification, bias can occur owing to false assumptions of homogeneity. For example, a model trained to optimize overall accuracy will favour groups that are well-represented in the training data over those that are under-represented[38].

[H2] Evaluation

Finally, model evaluation assesses the suitability, reliability and utility of a model. This is often performed iteratively with model specification as a way to improve the overall model. Different types of evaluation may be required depending on the purpose of the data analysis.

[H3] Validation

Validation typically involves verifying the assumptions of a model or estimating a model's performance on new data. For predictive tasks such as PGS, accuracy is assessed on a held-out test set, that is, a subset of the original dataset that was not used in model training. In these cases, bias can manifest if the test set is not representative of the target population. For example, PGS derived predominantly from individuals of European genetic ancestries have been shown to perform poorly for certain traits when applied to individuals of African or Asian genetic ancestries, underscoring the need for population-specific validation[12,13]. Importantly, model performance can vary significantly among populations traditionally homogenised using labels such as '*African*'. For instance, genetic risk scores derived from African American individuals exhibited variable predictive performance in South African Zulu and Ugandan cohorts, likely reflecting underlying genetic and environmental diversity[162].

[H3] Cost-benefit analysis

Models can be evaluated by weighing the potential health benefits of implementing a model in clinical practice against the costs associated with its deployment[39]. Such a cost-benefit analysis facilitates comparison to other potential uses of healthcare spending. It should account for the broader societal implications of deployment, including its impact on health inequities[40], and may also account for broader economic effects associated with inequities[41]. A focus solely on overall population health can conflict with efforts to reduce inequities, for instance, if delivering care effectively to disadvantaged groups is associated with extra costs[42].

[H3] Audit

The systematic review of performance after deployment serves to ensure that the model is continuing to function as intended and does not introduce unintended biases or inequities. Post-deployment audits are crucial to identify when models are used in more diverse populations or for different purposes than they were designed for, which can contribute to health disparities[43]. Audits should also consider the ethical implications of continued model use and whether adjustments or recalibrations are necessary to maintain equity in healthcare outcomes[44].

[H1] The role of statistical methods

The genomic data analysis framework illustrates the different stages at which biases can enter and their potential impact on equity. To highlight the ways in which statistical methods can promote equity, we now describe an ontology of methods for equitable genomic data analysis. This ontology bridges the gap between the statistical purpose of a method and its potential downstream benefit to health equity (**Fig. 2**). We also outline examples of specific methodological strategies and techniques that can be employed for different statistical purposes (Table 1). We largely focus on analyses of genetic data to illustrate these issues, although we stress that the principles are applicable to other types of genomic data. To illustrate the role methods can play throughout a single data analysis project, we also outline a case study on the development of an integrated risk tool for cardiovascular disease (**Box 3)**.

[H2] Reduce bias

Bias threatens the validity of a study's conclusions and can lead to disparities in accuracy. While there exist many different sources of bias (for examples, see the Catalog of Bias[51]), we focus here on those that are especially relevant to genomic equity.

[H3] Sampling bias

The lack of diversity in genomics can be seen as a sampling bias. Also known as ascertainment or selection bias, sampling bias occurs when the distribution of individuals in a study differs systematically from the population of interest. In machine learning, this difference between training data and target population is known as distribution shift.

Methods that address sampling bias and distribution shift can also enhance downstream statistical performance (see also *Increase statistical power for discovery and predictive accuracy* for further methods to mitigate sampling bias). One such method is transfer learning, also known as out-of-distribution generalization. This general machine learning strategy uses data from one context to boost performance in another context. For example, transfer learning can be used to adjust an existing PGS to a new target population[52,53]. Another approach relies on propensity scores, which quantify the probability that an individual will participate in a study given their personal characteristics. Propensity score regression can correct for population structure in GWAS, which unlike PCA-based methods also accounts for non-genetic factors[54]. Targeted study design can also be used to correct for sampling bias. For instance, the MULTIPOP framework guides the design of follow-up studies

using multi-population data to improve fine-mapping[55]. Data sampling strategies to construct training sets for PGS can also provide insight into where data collection efforts should be focused[56,57] .

[H3] Reference bias

Genomic references play a crucial role at various stages of genomic analyses, especially in mapping and imputation, but also in diagnostic analyses. Reference bias occurs due to the use of an inappropriate reference genome or panel. Naturally, the use of more representative references or panels reduces this bias, and has been found to improve imputation performance in diverse populations[58,] Meanwhile, meta-imputation provides a statistical approach to mitigate aggregate genotype data imputed with distinct reference panels into a single consensus imputed dataset[60]. Population-specific allele frequency estimates can reduce the search space of variants that may be more common in one group but rare in reference panel as a whole, but can mask within-group heterogeneity. Approaches to adjust allele frequencies accounting for this heterogeneity, such as Summix[96], can therefore provide an important tool to reduce variant effect misclassification. For mapping and variant detection, pangenomic methods offer a promising path towards reducing bias arising from reference genomes (see *Better genomic references, including pangenomic references*).

[H3] Confounding bias

Many genomic analyses seek to estimate the association between an 'exposure' or risk factor (for example a genetic variant) and an outcome of interest (for example, a disease). Confounding occurs when a third factor, a confounder, is associated with the exposure and influences the outcome. Consequently, an observed association between exposure and outcome may in fact be spurious. In GWAS, confounding due to population structure is well-recognised and a variety of robust methods are available to correct for this bias (see *Inclusion of more individuals)*. When conducting GWAS in admixed populations, further care must be taken to appropriately adjust for substantial intra-individual differences in allele frequencies, both at the global and local level. Importantly, which level to adjust for may depend on the analysis task, with global-ancestry adjustment preferred for screening and local-adjustment preferred for fine-mapping[62]. Approaches are also available for GWAS co-localization for expression quantitative trait loci (eQTL) mapping[63], and the analysis of population differences in gene regulation[64]. Note that such adjustments are typically achieved by simply adding estimates of population structure to the model, for instance the first ten components of a PCA. Such methods, however, are incomplete proxies for human genetic diversity (see *More nuanced approaches to categorization*) and can only ever partially mitigate confounding biases.

[H2] Increase statistical power for discovery and predictive accuracy

Methods to boost power for statistical inference or prediction provide benefits across the population, although given the current lack of diversity in genomic datasets, they can lead to outsized improvements in under-represented groups. We highlight three strategies to boost power and specific methodological techniques that have been employed for each: i) include more individuals, ii) include more traits, and iii) leverage non-genetic data.

[H3] Inclusion of more individuals

For any statistical task, two key determinants of performance are the sample size and the similarity of the training sample to the target population. Methods that facilitate the inclusion of more diverse individuals may improve statistical performance, especially for under-represented groups. General strategies include meta-analysis, joint or mixed modelling, and ensemble methods.

Meta-analysis methods combine summary statistics from multiple studies. Traditional meta-analytic approaches for GWAS typically use fixed-effects models, which assume that variant effect size is constant across studies. However, this assumption is often violated in meta-analyses involving diverse cohorts, owing to heterogeneity in LD structure, gene–gene interactions, gene-by-environment (GxE) interactions and variations in study design (for example, imputation artefacts). While cross-cohort heterogeneity in meta-analysis can be captured using random effects models[65] or trans-ancestral models[66,67], these approaches can suffer from lower discovery power.

Data from diverse cohorts can also be integrated through more sophisticated modelling. Such joint analysis is appealing because it includes all participants regardless of ancestry. A common example is the linear mixed model (LMM) for GWAS. LMMs can control for population stratification and relatedness (see *Confounding bias*) and benefit from increased power due to the larger sample size. Implementations are available for both continuous phenotypes[68,69] and for binary phenotypes[70,71]. The Tractor method also enables the inclusion of admixed individuals in GWAS[72] (see *Admixed populations*). As discussed in *Confounding bias,* LMMs may not fully control for population stratification, especially if there is further confounding from non-genetic factors that are correlated with genetic ancestry[163].

Beyond GWAS, joint modelling approaches can be used for fine-mapping. For example, extensions of the popular SuSIE model[73] to multiple populations are available in SuSIEx[74] and MeSusie[75]. Bayesian approaches that rely on priors, which allow for more heterogeneity in effect size estimates across populations, have also been proposed[76].

A wide array of PGS methods combine data from multiple populations to improve predictive accuracy. Bayesian approaches that jointly model data from distinct populations include PRS-CSx[77] and MUSSEL[78]. Ensemble methods such as CT-SLEB[79] and PROSPER[80] seek to optimally combine several distinct PGS trained on different populations. Other methods take a two-stage approach. For example, XP-BLUP leverages information from an independent (non-target) population by selecting SNPs that are strongly associated in the target population to inform the priors of the variance parameters in a LMM[81]. Methods specifically designed for admixed individuals remain in their infancy and generally rely on the use of local ancestry inference to inform effect size estimates (see *Admixed individuals).* These have been shown to improve performance in two-way recently admixed populations but have yet to be comprehensively evaluated against other PGS methods[164,165].

[H3] Use related traits

Data on related traits can be exploited to boost statistical power. Since common variants can influence different but related traits, information associated with these secondary traits can be relevant for the primary trait. The DeGAs method combines summary statistics from GWAS performed across many phenotypes, using a truncated singular value decomposition to identify shared components of genetic association and uncover novel variants and biological mechanisms across populations[82]. Relatedly, the adaptive sum of powered score (aSPU) statistical test is a form of variance components test that combines GWAS on related phenotypes to boost association power for rare variants[83]. XPXP improves PGS performance by incorporating population- and phenotype-specific effects within a LMM estimated using multiple traits and multiple populations[84].

[H3] Leverage non-genetic data

Non-genetic data, such as demographic information, environmental factors, clinical measures and other biological information, can enrich the contextual understanding of genetic information and further improve the generalizability of results. Functional annotations, information on the role of genetic variants across different cell types, can boost statistical power and predictive accuracy in cross-population settings for fine-mapping[85] and PGS[86,87]. Enrichment analyses, which combine information from variants implicated in common genes or functional pathways, can boost power to identify more genetic associations across ancestries[88]. Non-biological information can also be combined with genetic data. For example, incorporating family history boosts the accuracy of cross-population PGS[89]. Alternatively, existing clinical risk scores can be augmented with PGS to improve cross-population performance[90] (**Box 3**).

Although integrating such datasets can improve statistical performance, these additional data sources often exhibit the same lack of representativeness as genetic datasets and so can be subject to similar biases[91].

Functional annotations may be less comprehensive for biological processes associated with diseases that have been historically understudied[166]. Clinical information in databases of genetic variants such as ClinVar may be impacted unequal access to genomic healthcare. Moreover, the quality and completeness of such data may vary across the population[92]. For example, the lack of diversity has also been highlighted in molecular QTL datasets[93] and in epigenomic studies[94].

[H2] Assessing genetic variation

Statistical methods can be used to quantify variation at the population and individual levels. Assumptions of homogeneity can lead researchers to exclude certain subpopulations from an analysis, or if the assumptions are unfounded, bias the results of the analysis. The various techniques described above to increase statistical power can help in identifying variation; our focus here is instead on methods specifically designed to characterise heterogeneity in genetic structure within a population (Box 2).

[H3] Population structure

At the population level, quantifying genetic variation can provide insight into the transferability of models trained on genetic data. A generalization study is a form of hypothesis testing that aims to establish whether a particular association in a "discovery" GWAS on one population replicates in a "follow-up study" in another population[122,123]**.** The field of population genetics offers a wide array of approaches to characterize population structure. PCA, for instance, is widely used to correct for population structure in GWAS (see *Confounding bias*), by leveraging genetic relationship matrices derived from allele frequencies. Rare variants can further enhance resolution, uncovering fine-scale population structure that may remain hidden when relying solely on common variation[167]. In contexts where rare variants are unavailable—such as in cohorts genotyped exclusively using arrays—haplotype-based methods like identity-by-descent (IBD) and coalescent-informed approaches to construct genetic relationship matrices, can detect subtle patterns of heterogeneity that may not be apparent using allele frequency-based methods alone[126,168].

[H3] Trait-specific variation

Other measures assess genetic variation associated with a particular trait. For example, heritability quantifies the degree of variation of a trait that is due to genetics. This can vary with genetic architecture as well as the environment[97]. The popular LD score regression (LDSC) method estimates heritability for different populations using GWAS summary statistics[98]. LDSC, and related methods, typically rely on an accurate estimate of the LD matrix, which may not be available for heterogeneous or admixed populations. Approaches that account for

this heterogeneity, such as cov-LDSC[99], may enable more robust heritability estimation in such populations. The PESCA method employs ancestry group-specific LD estimates in combination with GWAS summary statistics to assess the distribution of shared versus population-specific causal variants for a trait[100]. Similarly, estimates of genetic correlation between two populations, as provided by methods such as Popcorn[101] and MAGIC[102], can be used in PGS development[103] and to provide insights into PGS transferability and the shared genetic architecture across populations[104].

[H3] Admixed populations

Admixed populations (Box 1) make up a significant part of global human genetic diversity, but are frequently overlooked in genomics research, limiting our scientific understanding of genetic variation and thus the evidence base for genomic medicine for such groups. The unique statistical challenges of admixed populations, due to their inheritance of genomic segments from several source populations, have been the main reason for their historical exclusion from genetic studies.

Methods to characterize the genetic structure of admixed populations can broadly be split into two categories: global and local ancestry inference (GAI and LAI, see Ref. [105] for a recent review)). GAI aims to estimate an individual's overall ancestry proportions from each source population, while LAI seeks to determine the source population for each portion of an individual's genome. Model-based approaches for GAI based on parental ancestry proportion an allele frequency include STRUCTURE[106] and ADMIXTURE[107], while fast algorithmic alternatives based on PCA-based heuristics are available for large datasets[108,109]. Many LAI methods are available[110]. These are often based on hidden Markov models, used to estimate the transitions along the genome from one source population to another[111,112]. The popular RFMix relies instead on conditional random fields to specify the relationship between genotype and local ancestry[113]. Both LAI and GAI estimates can be used to improve downstream analyses for admixed individuals, but further methodological work is needed (see *More nuanced approaches to categorisation*).

[H3] Sex differences

There is a growing recognition that sex differences in genetics should be studied to better understand inequalities across sex and gender[114,115]. The sex chromosomes have in the past often been excluded from GWAS, mainly due to the lack of availability of appropriate statistical approaches[116]. Specific methods and best practices are now available, providing tools needed to study sex differences and their role in health inequalities[117], including specific statistical tests for genetic association on the X chromosome[118,119], and

software to facilitate mapping, quality control, and imputation, and association testing for both sex chromosomes[120,121].

[H2] Identifying disparities in existing analyses

Finally, statistical methods can be used to detect inequitable group-level differences in the results of genomic studies. In doing so, these methods highlight when results are and are not applicable for different populations. A key challenge consists in establishing the source of these differences; methods outlined in the previous section can help to disentangle biological variation from true biases. This can in turn improve clinical or scientific decisions made based on the evidence from these studies. Moreover, it can inform the wider research agenda and thus influence the design of future studies.

A clear example of this lies in the development of PGSs. A series of papers highlighted the differential performance of PGSs across genetic ancestries[12,13,47], using population-aware modelling, simulation studies, and hypothesis testing. This instigated an avalanche of methodological developments targeted at reducing this gap and spurred further investigations into the factors underlying this differential performance[14,104].

Hypothesis testing can be used to confirm whether observed disparities are statistically significant. Hypothesis tests have also been used to highlight differences in variant prioritization between genetic ancestry groups[19], to identify disparities in diagnostic performance between ethnic groups for hypertrophic myopathy[17] and cystic fibrosis[18], and to disentangle the role of ethnicity and genetic ancestry in disease outcomes[124].

We caution that investigations into disparities must be carried out with great care. Attempts to identify group-level disparities carry the risk of supporting the idea that there exists an innate biological hierarchy between groups of humans. In particular, conflating sociopolitical groupings as biological categories can feed into racist ideologies (**Box 1**, and *More nuanced approaches to categorization*).

[H1] Future outlook

Methods development in genomics and genomic medicine is increasingly addressing equity, but there remain many challenges. Here, we outline several priorities for methodological innovation that we believe would yield the greatest positive impact to equity.

[H2] More nuanced approaches to categorization

A key challenge in genomics is the appropriate use of population descriptors. Researchers often categorise individuals into discrete groups, often by ethnicity, geography, or genetic similarity. Which categorization to use — or whether to use discrete categories at all — should be driven by the scientific question, and researchers

should articulate and justify their choice[125]. Whilst data-driven approaches for assigning group labels have been proposed[126,127], others have argued for a shift away from discretizations, which tend to exclude admixed indiviuals from analyses[128,129], risk ignoring within-population heterogeneity, and may over-emphasise apparent differences between groups. A recent National Academies report instead recommends the use of genetic similarity measures, without additional labelling, for many genetic analyses[28], with further methodological work needed to build such measures into standard genetics analyses. Local ancestry methods may provide a more fine-grained characterization of variation which is more inclusive of admixed individuals, though these hinge on the choice of source populations and availability of suitable reference data from these populations. Continuous notions of diversity may also facilitate a move away from discrete groupings that fail to reflect biological realities. For example, a recent approach that utilised a continuous metric of genetic distance from the centre of the training cohort to assess PGS performance represents an important step toward more accurate and inclusive frameworks[14]. Methods that approximate ancestral recombination graphs (ARGs), such as Relate or tsinfer, offer another promising avenue[169,170]. By capturing the complex branching patterns of genetic ancestry, ARGs provide a detailed representation of genomic variation without relying on categorisation, unlocking opportunities for high-resolution and innovative applications in statistical genomics[171]. New ARG-based methods are needed to perform essential tasks such as association analyses, variant prioritization, and PGS construction Given the varied practical, ethical and statistical considerations in these approaches, an interdisciplinary collaboration is paramount in developing better ways of characterizing diversity.

[H2] Harnessing advances in genomic references

Oversimplified categorization occurs implicitly in the use of reference genomes and panels. For instance, the current most commonly used human reference genome GRCh38 is not representative of the global population or indeed any population[130]. The potential for reference bias has prompted calls for alternative references[130]. A promising alternative is the draft human pangenome reference[131]. A pangenome is a collection of genome sequences, typically represented by a graph-based data structure. New statistical methods are needed to reap the rewards of this new data structure. While some bioinformatics techniques are available to perform read mapping and variant calling[132–134], further work is needed to improve computational scalability, and to extend the use of pangenomes to common analysis tasks such as genotype imputation. Potential biases in reference datasets can have direct clinical implications in the misclassification of variants for disease diagnoses[17]. Mitigating these biases is challenging as the datasets only publish summary statistics rather than individual-level data. Methodological advancements in federated learning may provide a path forward to safely overcoming these challenges (see ***Facilitating data sharing and federated learning safely).***

[H2] Understanding the role of social and environmental effects

Although the multifaceted nature of health inequalities has been well-documented[136,137], its role in the context of genetic analyses has been less studied. Understanding this interplay is crucial to developing effective interventions to reduce inequalities, which exist along numerous axes, including gender, ethnicity and other social determinants of health. At the population level, these social determinants may be correlated with genetic structure, which can introduce confounding bias. Advances in causal inference techniques, including Mendelian randomization[142], are required to handle more complex forms of interactions between these factors, particularly in the context of highly polygenic traits[173]. While there has been recent interest in the use of GxE methods to investigate health inequalities[139,140].these are subject to similar biases as genetic studies[141] with new methods required to mitigate these effects. Moreover, analysis of social and environmental data introduces a range of new challenges including noisy measurements, self-reporting biases, and inconsistent metrics. Effect sizes are often small and so achieving sufficient statistical power can be difficult. Recent efforts to use metabolomics to enhance environmental exposure studies may provide a useful path forward to address some of these challenges[174].

[H2] Facilitating data sharing and federated learning safely

Widening the pool of existing genomic data available for statistical analysis can boost statistical power. The widespread publication of GWAS summary statistics, for example, has significantly advanced our understanding of the genetic architecture of complex traits and diseases. Nevertheless, important technical, statistical, and ethical issues must be addressed to harness distributed datasets safely and effectively[30]. Beyond GWAS, a fundamental statistical challenge is to determine which summary statistics are needed to make inference on a given statistical model. Combining genetic and non-genetic information (such as electronic health records or socioeconomic data) provides a further challenge. Data privacy and sovereignty is a crucial consideration (see Box 2). Statistical methods that reliably preserve privacy can help pave the way towards more inclusive and representative genomic databases[143]. New methods are needed to efficiently combine information across datasets while reducing the risk of identification. Privacy-preserving synthetic data generation (SDG) could be used to fit models or train clinical algorithms[144]. However, most existing SDG methods introduce statistical bias into downstream analyses, although approaches are available to mitigate this[145]. Moreover, while privacy-preserving methods may help to build trust, interdisciplinary efforts are nonetheless necessary to ensure appropriate compliance with ethical and legal issues around data sharing[175].

[H2] Methods for multi-omics and pharmacogenomic studies

Many of the methods we have outlined were developed for GWAS or their downstream applications. By contrast, relatively few methods are available in emerging areas of genomics, such as tumour sequencing, single-cell sequencing, spatial transcriptomics, organoids, pharmacogenomic and multi-omics studies. However, the data disparities in GWAS are also present in other omics fields, with individuals of European genetic ancestries comprising 93% of protein QTL (pQTL) studies, and 82% of the GTEx project, a commonly used eQTL study[146,147]. New 'multi-modal' methods to integrate and analyse data from multi-omics studies are required to understand the biological mechanisms underlying diseases and provide insight in how to best prevent and treat diseases[148]. Importantly, mutation rates, DNA methylation and mRNA expression have been shown to vary by genetic ancestry[149] . Given their role in identifying cancer pathways and driver genes, such genetic ancestry-related differences must be properly accounted for. Further methodological developments are required to detect, characterize and appropriately handle population heterogeneity within these studies.

[H1] Conclusions

In this Perspective, we have highlighted how methodological developments can play a crucial part in promoting equity. In conjunction with the ongoing drive to build more representative genomic datasets, we encourage researchers to take equity into account at all stages of a genomic study, from inception to data acquisition and analysis. Emerging fields such as single-cell genomics can learn lessons from GWAS by considering equity from the beginning, before data disparities and biases are 'baked in'. Meaningful patient and public involvement from study design to analysis can provide an additional human perspective into more statistical notions of equity. Such an approach will also help to build trust in genomic research and medicine[7]. We also urge researchers to improve the reporting of methods, both in terms of performance across key demographic strata and the representativeness of the training data on which they were developed[176]. Realistic measures of uncertainty in conclusions or estimates that accommodate non-representative data will improve the robustness of conclusions and highlight potential biases. Finally, we ask researchers to look further afield than genomics for innovations with the potential to improve equity. The push towards better, more equitable analytical tools has been initiated across several disciplines including clinical trials, computational medicine, epidemiology and public health, as well as in purely methodological areas of the statistical and machine learning literature[150]. As well as technical solutions, these disciplines provide a wealth of useful frameworks[27], standards[176], and toolkits[177] from which, suitably adapted, genomics could stand to benefit in its efforts towards advancing health equity.

Acknowledgements

The authors gratefully acknowledge the speakers and attendees of the joint Data Science for Health Equity workshops on ‘Challenges to statistical approaches for fairness in genomics’ and ‘Challenges to statistical approaches for health equity’ held in January 2022. They also thank C. Harbron, G. McVean, S. Walker, D. Deen, A. Shalek and H. Martin for comments on an earlier version of this manuscript.

Competing interests

This manuscript was informed by a project commissioned by the Diverse Data (DD) initiative at Genomics England (GEL) in December 2022 to explore the use of statistical and machine learning methods to improve fairness and equity in genomics. K.K. is the Scientific Lead for DD. S.T, T.N., and Y.C are Genomic Data Scientists at GEL. M.S. was the Lead Genomic Data Scientist for DD, and M.M. was the Programme Lead for DD. B.L. and L.B .were paid consultants to GEL for the project. M.M. is Director of One HealthTech, which provides the secretariat for the Data Science for Health Equity community, which B.L. is also the co-founder of.

549 **Table 1**. Categorization of methods for genomic data analysis and their potential benefit to health equity

Purpose of method		**Benefit(s) to equity**	**Analysis stage**	**Methodological strategy or technique**	**Example applications**
Reduce bias	*Sampling bias* Account and correct for data representativeness	Improve generalizability of results; Reduce disparities in statistical performance and clinical utility	Research design & data acquisition; Model development	Tailored study design Transfer learning	PGS[56,57]
				Transfer learning	PGS[52,53]
	Confounding bias Establish genuine causal effects	Improve generalizability of results	Model development	*Population-aware modelling:*	
				Mixed models	GWAS[68–71]
				PC regression	GWAS[62], eQTL analysis[63]
	Reference bias Reduce errors due to unrepresentative reference genomes or panels	Improve generalizability of results; Reduce disparities in statistical performance and clinical utility	Data preparation; Model development	Meta-imputation	Genotype imputation[59,60]
				Pangenome methods	Variant calling[131,132]
Increase statistical power (e.g. fine-mapping; variant detection) and **Boost predictive accuracy** (e.g. polygenic scoring)		Improve generalizability of results; Reduce disparities in statistical performance and clinical utility	Research design & data acquisition; Model development	*Inclusion of more individuals:*	
				Meta-analysis	GWAS[65–67]
				Mixed/joint models	GWAS[68-72,164] Fine-mapping[74-76]

Purpose of method	Benefit(s) to equity	Analysis stage	Methodological strategy or technique	Example applications
				PGS[77,78]
			Ensemble methods	PGS[79,80]
			Multi-trait analysis	GWAS[82,83] , PGS[84]
			Leverage non-genetic data	GWAS[85,87,88] , PGS[86,89,90]
Assess genetic variation	Inform research agenda; Improve future downstream analyses	Model development; Evaluation	Hypothesis testing	GWAS[122,123]
			Trait-specific variation	Heritability[99] GWAS[100] Genetic correlation[101,102]
			Population-aware modelling:	
			Admixed populations	Global ancestry inference[106–109] Local ancestry inference[110–112]
			Sex chromosomes	GWAS[117–121]
Identify disparities	Inform research agenda	Evaluation	Hypothesis testing	Variant classification[17-19]
			Population-aware modelling	PGS[14,104]

550 'Population-aware models' are those that take into account particular characteristics, such as genetic
551 architecture, of the study population (**Box 2**).
552
553 **Table 2**. Challenges and opportunities for methodological innovation to advance equity in genomics

	Key challenges	Methodological opportunities

More nuanced approaches to categorization	- Overreliance on discrete population groupings that do not reflect human genetic diversity - Exclusion of admixed individuals from analyses	- Approaches that rely solely on genetic similarity measures without group labels - Continuous characterisations of genetic ancestry - Tools based on ancestral recombination graphs
Harnessing advances in genomic references	- Standard reference genomes and panels are unrepresentative of the global population - Methods for alternative references are computationally expensive	- Scalable implementations of pangenome methods and extensions to standard genetic analyses - Federated learning combining reference-based summary statistics and individual-level data
Understanding the role of social and environmental effects	- Confounding biases from correlations between social determinants and genetic structure. - Challenges in analyzing noisy, self-reported, or inconsistent data.	- Causal inference techniques, including Mendelian randomization, for high-dimensional data - Careful use of proxies, e.g. metabolomics for environmental exposures
Facilitating data sharing and federated learning safely	- Combining genetic and non-genetic information - Compliance with ethical and legal considerations	- Development of summary statistics beyond GWAS - Privacy-preserving synthetic data generation and federated learning
Methods for multi-omics and pharmacogenomic studies	- European ancestry bias in omics datasets - Limited availability of population-aware tools for emerging fields	- Multi-modal tools to integrate different data types - Adaptations of GWAS approaches that account for population heterogeneity

Figure 1. A conceptual framework for a general genomic data analysis task. Genomic research operates within a broader sociopolitical ecosystem (see Box 2). Critical factors for equity at the ecosystem-level include workforce diversity, partnerships, public and patient involvement, and funding. Each of these stages are interwoven and have an impact on data analysis but in particular influence the Research agenda and prioritization, which in turn shapes the subsequent design of a research study. At the outset of a study, the research design & data acquisition stage determines what data is both required and available to investigate a particular scientific question. Who gets included in the study data is determined via study design and participation as well as the limitations of data collection/availability. This is followed by a data preparation stage which consists of tasks such as read mapping and alignment to a reference genome, quality control, and preprocessing steps such as imputation. The model development stage typically involves feature engineering, whereby data is transformed into 'features' that can be used in a particular model; determining the model specification to characterize the functional form of the data; and model training in which the parameters of a model are tuned to best align with the observed data. In many machine learning models, features or representations are learnt from the data iteratively through model training. Once a model has been specified and trained, it undergoes an evaluation stage. In the research context, model validation assesses the accuracy and reliability of the results. Meanwhile, in the clinical context, a cost-benefit analysis determines its practical implementation value. Additionally, once a model is deployed, an audit process aims to monitor its

performance. There may be several iterations of model development and evaluation steps to optimize performance. The outcomes of a genomic data analysis feed back into the sociopolitical ecosystem by advancing our scientific understanding through knowledge generation and the formulation of clinical policy, both of which loop back to influence the research agenda and prioritization, highlighting the cyclical nature of genomic research and its impact on genomic medicine. Analytical tools have variable influence throughout this framework, comprising key aspects of a data analysis but having little impact on the sociopolitical backdrop.

Figure 2. Pathways to equity. Statistical methods play a crucial role throughout a genomic data analysis. This schematic illustrates where and how statistical methods can benefit equity in genomics research, highlighting the different stages within the genomic data analysis framework (Figure 1), their goal or purpose at a statistical level, and the subsequent benefits to equity. At the model evaluation stage, statistical methods can serve to identify disparities in the results of genetic studies. The appropriate selection of evaluation cohorts and outcome measures can play a crucial role in identifying such disparities, which can serve to influence the research agenda and inform the design of future studies. Methods to assess genetic variation within the population can be used at both the model development and evaluation stages. Data preparation can also have an impact here; important population-specific variation may be overlooked as noise or smoothed over. Characterisations of genetic variation can inform both the specification of downstream analyses, reducing bias and improving power, as well as the design of future studies. Techniques to reduce statistical bias and increase statistical power can be used at various stages of a data analysis. These serve the twin goals of improving the generalizability and reducing disparities in accuracy of model outputs. In turn, these enhance the validity of scientific conclusions and the evidence-base for clinical decision making, particularly for those currently under-represented in genomic studies.

Box 1. Population descriptors

Genomics research studies patterns in genomic data to gain insights into the biological mechanisms that underpin human life, and how they interact with the environmental context. Researchers often use discrete labels to group individuals as a convenient way to describe the continuous and intricate patterns of human genetic variation shaped by history, migration, and evolution. In practice, no single categorization can adequately represent the complexity of genetic variation, and the use of population descriptors should be carefully tailored to the scientific question at hand to avoid further exacerbating disparities[28]. Here we provide key definitions and terminology for population descriptors commonly used in genomics.

Population

A population refers to a group of individuals with a common attribute or perceived characteristic. Examples include geographical location, race, ethnicity, genetic ancestry, sex, and gender.

Race, ethnicity and ancestry

These terms are often - and mistakenly - conflated. Race is a sociopolitically constructed system, often used to classify and rank people based on supposed innate biological characteristics. Ethnicity is another sociopolitically constructed system for classifying people based on shared heritage or cultural similarities, such as language or religion. Both of these systems vary globally. Ancestry is a context-dependent term that generally refers to a person's descent over (a varying amount of) generations, and can encompass both social and biological factors. As such, the term "ancestry" alone should be avoided in genetic analyses. Instead, it should be qualified to clarify its intended meaning—such as "genetic ancestry" (see *Genetic ancestry* below), or "genealogical ancestry" for descent through family lineages, to prevent misconceptions and ensure precision. Similarly, neither race nor ethnicity should be confused with genetic ancestry, or used as proxies for population genetic variation.

Genetic ancestry

Genetic ancestry, defined by the complex mosaic of stretches of genomes inherited from different genealogical ancestors [151], is formalized in a structure called an ancestral recombination graph (ARG), which traces how genetic variation is inherited by individuals over time. Whilst researchers often define genetic ancestry groups by clustering together genetically similar individuals and assigning a geographic label to their members, such groups are modelling constructs with that vastly reduce the full complexity of genetic ancestry. Due to the risk of conflating these group labels with race or ethnicity and thus feeding into racist ideologies, this process should be undertaken carefully and with a comprehensive understanding of its limitations (see *More nuanced approaches to categorization)*.

Admixture

Statistical models in genomics often rely on an assumption of the existence of discrete ancestral populations from which individuals today inherit their genetic material. In this context, admixture occurs when individuals from different ancestral groups mix. It is important to note that this concept is timescale-dependent and

makes the simplifying assumption of the existence of 'pure', homogeneous populations in the past. By definition, all humans are admixed, but not everyone is recently admixed. In practice, the term 'admixed' is typically used to refer to individuals with recent admixture (<100 generations).

Sex and gender

These are different concepts that are sometimes mistakenly used interchangeably. Sex is a biological characteristic that is generally defined according to reproductive organs, chromosomes, or hormones. Sex is typically treated as a binary trait, although there are cases of intersex individuals who do not fit within this binary characterization. Gender, on the other hand, refers to a sociocultural identity held by an individual and refers to a spectrum.

595

Box 2. The sociopolitical backdrop to genomic data analysis

Each stage of the genomic data analysis framework (Figure 1) is shaped by the broader sociopolitical context of genomic research and medicine. Better methods have limited direct influence here; tackling these challenges will instead require significant ecosystem-level efforts[6,45,46]. Here, we highlight key ecosystem-level considerations with cascading effects on genomic data analysis.

Knowledge generation and clinical policy

A study can influence clinical policy and lead to advances in scientific knowledge. Biases at any stage can lead to spurious conclusions or limit the relevance of evidence to support clinical decision-making. Moreover, the interpretation of findings may also reflect existing social or cultural prejudices. Such biases can result in less accurate diagnoses and suboptimal treatment recommendations, ultimately resulting in poorer health outcomes for those historically under-represented and marginalized in genomics research[4,47]. Clinical policy may also be impacted by funding constraints and strategic priorities, particularly in the face of other public health challenges such as infectious diseases.

Workforce diversity, funding and partnerships

Across biomedical research, scientists from low-income countries and diverse ethnic groups are under-represented[6]. This lack of workforce diversity diminishes important perspectives and research questions that experts from different backgrounds contribute to genomics. Similar diversity gaps in funding agencies and journal editorial boards may also inadvertently create barriers to opportunities at different points in the

research cycle. Conversely, funders can promote equity in genomics by prioritizing greater inclusion of underserved populations in study designs and review criteria[48]. Equitable partnerships between high-income and lower-income countries also have a key role to play in improving diversity in genomics[45].

Patient and public engagement

Exploitative research in genomics can be harmful to study participants specifically and damage trust in genomics medicine more generally[178,179]. Concerns over data privacy and misuse, particularly those from historically under-represented groups, can act as a barrier towards equitable participation in genomic studies[7]. Issues of data ownership and consent have significant implications for data sharing and secondary analyses[181]. Engagement of study populations throughout the research process also has a crucial role to play in enhancing the quality of research and ultimately reducing health disparities[180]. What constitutes proper engagement is rapidly evolving, moving away from viewing participants as mere "research subjects" toward involving them as active collaborators, for example, by providing input into study design[49].

Research agenda and prioritization

Because of these ecosystem-level factors, the research questions that are posed and the focus of efforts in genomic medicine do not always represent diverse priorities. Funding incentives alongside the personal interests of a homogeneous workforce move the focus of genomic data analysis to diseases that affect privileged demographic groups, often further marginalizing minoritized populations[49,50]. For example, intense research efforts have been invested to study cardiovascular disease, which disproportionately affects older men, but comparatively little research has been undertaken into sickle cell disease, which affects mostly people of African descent. The influence on the research agenda in turn affects study design and data availability, bringing us back to the beginning of our framework. Thus, the framework is cyclical in nature: biases can enter at various stages of analysis, cascading negative effects to downstream stages and future studies, and amassing inequity in each cycle.

596

Box 3. Population-aware modelling

The appropriate choice of model for any statistical analysis is essential to drawing valid scientific insights from observed data. False assumptions of homogeneity across populations can result in invalid conclusions and

suboptimal statistical performance arising from genomic data analysis. A well-studied example of this is the impact of population stratification in GWAS: differences in population structure that are left uncorrected can lead to spurious associations. In some cases, these assumptions of homogeneity can be a direct consequence of lack of available data. The misclassification of genetic variants for hypertrophic cardiomyopathy in Black Americans rested both on a lack of control data in diverse cohorts and an implicit assumption of similar allele frequencies across populations[17]. In other circumstances, researchers may instead elect to exclude minority groups of individuals from analyses entirely, in order to meet the assumptions of the model, thus limiting the scientific evidence base for these groups[8].

Population-aware strategies are those that acknowledge and leverage the diverse genetic landscapes present in human populations to mitigate against false assumptions of homogeneity. The term 'population-aware' encompasses 'ancestry-aware' methods that account for differences in genetic architecture that occur between populations as well as approaches that address other forms of population variation, including sex differences and social disparities. The availability of ancestry-aware methods, especially for GWAS and PGS, has increased markedly in recent years, carrying the potential to promote equity in genomics research (see *The role of statistical methods* for several examples). Further methodological development towards approaches that account for other forms of population variation is required to further our understanding of the interplay between the genetic and non-genetic factors underlying health inequalities (see *Understanding the role of social and environmental effects)*.

Box 4. Case study: an integrated risk score for cardiovascular disease

Statistical methods can promote health equity in genomics in numerous ways, at several stages of analysis, even for a single study. To illustrate this, we place a recently developed integrated risk tool for atherosclerotic cardiovascular disease (ASCVD-IRT)[90] in the context of the genomic data analysis framework (**Fig. 1**).

Research agenda and prioritization

Clinical risk scores for cardiovascular diseases (CVD) such as the QRISK and atherosclerotic cardiovascular disease pooled cohort equations (ASCVD-PCE) scores are being used in clinical practice to identify and offer preventative treatment to those at increased risk of CVD. Combining these risk scores with a polygenic score (PGS) to create an integrated risk tool (IRT) can boost predictive performance, although this has yet to be

rigorously evaluated across diverse population groups. ASCVD-IRT was expressly developed "to predict 10-year risk of ASCVD across diverse ethnicity and genetic ancestry groups"[90].

Research design and data acquisition

To construct the PGS, ten GWAS datasets for ASCVD were obtained. These represented individuals from multiple ancestry groups and geographies. ASCVD-IRT was then developed by combining ASCVD-PCE with the PGS, with the relative contribution of the two risk scores tuned using data from four further, genetically diverse cohorts. The use of diverse cohorts can improve transferability beyond individuals of European genetic ancestries (see *Inclusion of more individuals*).

Data preparation

The PGS was constructed using summary statistics from existing GWAS and did not carry out new sequencing. These GWAS may be subject to bias arising from the use of a single reference genome (see *Better genomic references, including pangenomic references*), although we note that the impact on PGS performance of such reference bias has not yet been evaluated. Details to perform each GWAS were not all available, but we note that the use of GWAS in diverse cohorts typically requires tailored techniques for preprocessing, including quality control (QC), adjustment for population structure, and genotype imputation[24].

Model development

The PGS was developed using LDPred[152] following a fixed-effects meta-analysis of the ten GWAS datasets. This step also used trait-specific functional information to inform the effect size estimates for each SNP[153], which may serve to enhance the generalizability of the PGS (see *Leverage non-genetic data*). We note that only a single PGS was developed; the use of ancestry-specific PGS may yield improved predictive performance for CVD across genetic ancestries[154], although this would carry further difficulties in implementation to determine which PGS an individual should receive.

ASCVD-IRT was trained by combining ASCVD-PCE with a rescaled PGS, where the scaling factor was tuned using four diverse cohorts in which both the required genetic and clinical data was available. The scaling factor was estimated separately by sex and by African v non-African ancestry, reflecting empirical patterns observed in the four cohorts (see *Assessing genetic variation*). This is an example of *Population-aware modelling* (**Box 2**), whereby heterogeneity is accounted for in order to improve statistical performance across the population.

Evaluation

The performance of ASCVD-IRT was carefully evaluated across three diverse cohorts, two in the US and one in the UK[90]. Various performance metrics were used, including sensitivity, specificity, and net reclassification improvement over ASCVD-PCE. The reporting of these metrics was stratified by self-reported ethnicity, genetically inferred ancestry, age group, and sex. Hypothesis testing was used to confirm the statistical significance of the performance improvements of ASCVD-IRT over ASCVD-PCE for different subgroups. Performance for admixed individuals, however, was not evaluated.

The potential clinical benefit of a closely related IRT was also evaluated in two follow-up studies investigating the clinical utility, feasibility, and acceptability to both participants and healthcare providers in implementing the IRT into clinical practice in the UK National Health Service[155,156].

References

1. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. Genomics for the world. *Nature* **475**, 163–165 (2011).
2. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat Med* **28**, 243–250 (2022).
3. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
4. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
5. Spratt, D. E. *et al.* Racial/Ethnic Disparities in Genomic Sequencing. *JAMA Oncol.* **2**, 1070–1074 (2016).
6. Bentley, A. R., Callier, S. & Rotimi, C. N. Diversity and inclusion in genomic research: why the uneven progress? *J. Community Genet.* **8**, 255–266 (2017).
7. Atutornu, J., Milne, R., Costa, A., Patch, C. & Middleton, A. Towards equitable and trustworthy genomics research. *eBioMedicine* **76**, (2022).
8. Manolio, T. A. Using the Data We Have: Improving Diversity in Genomic Research. *Am. J. Hum. Genet.* **105**, 233–236 (2019).
9. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
10. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

11. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* 1–12 (2023) doi:10.1038/s41562-023-01579-9.
12. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635–649 (2017).
13. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
14. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* 1–8 (2023) doi:10.1038/s41586-023-06079-4.
15. Int Common Dis Alliance *et al.* Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.* **27**, 1876–1884 (2021).
16. Kullo, I. *et al.* Polygenic scores in biomedical research. *Nat. Rev. Genet.* **23**, 524–532 (2022).
17. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
18. Schrijver, I. *et al.* The Spectrum of CFTR Variants in Nonwhite Cystic Fibrosis Patients: Implications for Molecular Diagnostic Testing. *J. Mol. Diagn.* **18**, 39–50 (2016).
19. Tallman, S. *et al.* Missing genetic diversity impacts variant prioritisation for rare disorders. 2024.08.12.24311664 Preprint at https://doi.org/10.1101/2024.08.12.24311664 (2024).
20. THE H3AFRICA CONSORTIUM *et al.* Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
21. The “All of Us” Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
22. Our Future Health. *Our Future Health* https://ourfuturehealth.org.uk/.
23. Wang, Y., Tsuo, K., Kanai, M., Neale, B. & Martin, A. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu. Rev. Biomed. DATA Sci.* **5**, 293–320 (2022).
24. Peterson RE *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).
25. Kachuri, L. *et al.* Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* 1–18 (2023) doi:10.1038/s41576-023-00637-2.
26. Chen, I. Y. *et al.* Ethical Machine Learning in Health Care. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021).
27. Burr, C. & Leslie, D. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI Ethics* **3**, 73–98 (2023).
28. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. (National Academies Press, Washington, D.C., 2023). doi:10.17226/26902.

29. Yancey, A. K., Ortega, A. N. & Kumanyika, S. K. EFFECTIVE RECRUITMENT AND RETENTION OF MINORITY RESEARCH PARTICIPANTS. *Annu. Rev. Public Health* **27**, 1–28 (2006).
30. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* **21**, 615–629 (2020).
31. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
32. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* **22**, 8 (2021).
33. Huang, L. *et al.* Genotype-Imputation Accuracy across Worldwide Human Populations. *Am. J. Hum. Genet.* **84**, 235–250 (2009).
34. Xiang, R. *et al.* A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Front. Genet.* **12**, 646936 (2021).
35. Kozlov, M. 'All of Us' genetics chart stirs unease over controversial depiction of race. *Nature* (2024) doi:10.1038/d41586-024-00568-w.
36. Lin, P.-I., Vance, J. M., Pericak-Vance, M. A. & Martin, E. R. No Gene Is an Island: The Flip-Flop Phenomenon. *Am. J. Hum. Genet.* **80**, 531–538 (2007).
37. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biol.* **19**, 179 (2018).
38. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 77–91 (PMLR, 2018).
39. Payne, K., Gavan, S. P., Wright, S. J. & Thompson, A. J. Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nat. Rev. Genet.* **19**, 235–246 (2018).
40. Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision Public Health for the Era of Precision Medicine. *Am. J. Prev. Med.* **50**, 398–401 (2016).
41. LaVeist, T. A. *et al.* The Economic Burden of Racial, Ethnic, and Educational Health Inequities in the US. *JAMA* **329**, 1682 (2023).
42. Cookson, R. *et al.* Using Cost-Effectiveness Analysis to Address Health Equity Concerns. *Value Health* **20**, 206–212 (2017).
43. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
44. Liu, X. *et al.* The medical algorithmic audit. *Lancet Digit. Health* **4**, e384–e397 (2022).

45. Martin, A. R. *et al.* Increasing diversity in genomics requires investment in equitable partnerships and capacity building. *Nat. Genet.* **54**, 740–745 (2022).
46. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
47. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).
48. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
49. Rebbeck, T. R. *et al.* A Framework for Promoting Diversity, Equity, and Inclusion in Genetics and Genomics Research. *JAMA Health Forum* **3**, (2022).
50. Pereira, L., Mutesa, L., Tindana, P. & Ramsay, M. African genetic diversity and adaptation inform a precision medicine agenda. *Nat. Rev. Genet.* **22**, 284–306 (2021).
51. Catalogue of Bias Collaboration. Catalog of Bias. https://catalogofbias.org/.
52. Tian, P. *et al.* Multiethnic polygenic risk prediction in diverse populations through transfer learning. *Front. Genet.* **13**, null (2022).
53. Zhao, Z., Fritsche, L. G., Smith, J. A., Mukherjee, B. & Lee, S. The construction of cross-population polygenic risk scores using transfer learning. *Am. J. Hum. Genet.* **109**, 1998–2008 (2022).
54. Zhao H, Rebbeck TR, & Mitra N. A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors. *Genet Epidemiol* **33**, 679–90 (2009).
55. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging Genetic Variability across Populations for the Identification of Causal Variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).
56. Cai, W. *et al.* Adaptive Sampling Strategies to Construct Equitable Training Datasets. in *2022 ACM Conference on Fairness, Accountability, and Transparency* 1467–1478 (Association for Computing Machinery, New York, NY, USA, 2022). doi:10.1145/3531146.3533203.
57. Lehmann, B., Mackintosh, M., McVean, G. & Holmes, C. Optimal strategies for learning multi-ancestry polygenic scores vary across traits. *Nat. Commun.* **14**, 4023 (2023).
58. Jimenez-Kaufmann, A. *et al.* Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes. *Front. Genet.* **12**, (2022).
59. Sengupta, D. *et al.* Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics* 100332 (2023) doi:10.1016/j.xgen.2023.100332.

60. Yu, K. *et al.* Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet.* **109**, 1007–1015 (2022).
61. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).
62. Martin, E. R. *et al.* Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet. Epidemiol.* **42**, 214–229 (2018).
63. Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* **21**, 233 (2020).
64. Natri, H. M. *et al.* Genetic architecture of gene regulation in Indonesian populations identifies QTLs associated with global and local ancestries. *Am. J. Hum. Genet.* **109**, 50–65 (2022).
65. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
66. Mägi, R. *et al.* Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).
67. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).
68. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
69. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genet.* **9**, e1003264 (2013).
70. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
71. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
72. Atkinson, E. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195-+ (2021).
73. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
74. Yuan, K. *et al.* Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. *Nat. Genet.* **56**, 1841–1850 (2024).
75. Gao, B. & Zhou, X. MESuSiE enables scalable and powerful multi-ancestry fine-mapping of causal variants in genome-wide association studies. *Nat. Genet.* **56**, 170–179 (2024).

76. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).
77. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
78. Jin, J. *et al.* MUSSEL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups. *Cell Genomics* **4**, (2024).
79. Zhang, H. *et al.* A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nat. Genet.* **55**, 1757–1768 (2023).
80. Zhang, J. *et al.* An ensemble penalized regression method for multi-ancestry polygenic risk prediction. *Nat. Commun.* **15**, 3238 (2024).
81. Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. & Tang, H. Leveraging Multi-ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. *Am. J. Hum. Genet.* **101**, 218–226 (2017).
82. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).
83. Kim, J., Bai, Y. & Pan, W. An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. *Genet. Epidemiol.* **39**, 651–663 (2015).
84. Xiao, J. *et al.* XPXP: improving polygenic prediction by cross-population and cross-phenotype analysis. *Bioinformatics* **38**, 1947–1955 (2022).
85. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
86. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
87. Amariuta, T. *et al.* Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat. Genet.* **52**, 1346–1354 (2020).
88. Smith, S. P. *et al.* Enrichment analyses identify shared associations for 25 quantitative traits in over 600,000 individuals from seven diverse ancestries. *Am J Hum Genet* **109**, 871–884 (2022).
89. Hujoel MLA, Loh PR, Neale BM, & Price AL. Incorporating family history of disease improves polygenic risk scores in diverse populations. *Cell Genom* **2**, (2022).
90. Weale, M. E. *et al.* Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. *Am. J. Cardiol.* **148**, 157–164 (2021).
91. Committee on Improving the Representation of Women and Underrepresented Minorities in Clinical Trials and Research, Committee on Women in Science, Engineering, and Medicine,

Policy and Global Affairs, & National Academies of Sciences, Engineering, and Medicine. *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups*. 26479 (National Academies Press, Washington, D.C., 2022). doi:10.17226/26479.

92. Mitra, R. *et al.* Learning from data with structured missingness. *Nat. Mach. Intell.* **5**, 13–23 (2023).
93. Long, E. *et al.* The case for increasing diversity in tissue-based functional genomics datasets to understand human disease susceptibility. *Nat. Commun.* **13**, 2907 (2022).
94. Breeze, C. E., Beck, S., Berndt, S. I. & Franceschini, N. The missing diversity in human epigenomic studies. *Nat. Genet.* **54**, 737–739 (2022).
95. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).
96. Arriaga-MacKenzie, I. *et al.* Summix: A method for detecting and adjusting for population structure in genetic summary data. *Am. J. Hum. Genet.* **108**, 1270–1282 (2021).
97. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
98. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
99. Luo, Y. *et al.* Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Hum. Mol. Genet.* **30**, 1521–1534 (2021).
100. Shi, H. *et al.* Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. *Am. J. Hum. Genet.* **106**, 805–817 (2020).
101. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
102. Lu, H. *et al.* Evaluating marginal genetic correlation of associated loci for complex diseases and traits between European and East Asian populations. *Hum Genet* **140**, 1285–1297 (2021).
103. Cai, M. *et al.* A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am. J. Hum. Genet.* **null**, null (2021).
104. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
105. Tan, T. & Atkinson, E. G. Strategies for the Genomic Analysis of Admixed Populations. *Annu. Rev. Biomed. Data Sci.* **6**, 105–127 (2023).
106. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945–959 (2000).

107. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
108. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genet.* **2**, e190 (2006).
109. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genet. Epidemiol.* **39**, 276–293 (2015).
110. Wu, J., Liu, Y. & Zhao, Y. Systematic Review on Local Ancestor Inference From a Mathematical and Algorithmic Perspective. *Front. Genet.* **12**, (2021).
111. Salter-Townshend, M. & Myers, S. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics* **212**, 869–889 (2019).
112. Li, N. & Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**, 2213–2233 (2003).
113. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
114. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).
115. Accounting for sex in the genome. *Nat. Med.* **23**, 1243–1243 (2017).
116. Sun, L., Wang, Z., Lu, T., Manolio, T. A. & Paterson, A. D. eXclusionarY: 10 years later, where are the sex chromosomes in GWASs? *Am. J. Hum. Genet.* **110**, 903–912 (2023).
117. Khramtsova, E. A. *et al.* Quality control and analytic best practices for testing genetic models of sex differences in large populations. *Cell* **186**, 2044–2061 (2023).
118. Clayton, D. Testing for association on the X chromosome. *Biostatistics* **9**, 593–600 (2008).
119. Loley, C., Ziegler, A. & König, I. R. Association Tests for X-Chromosomal Markers – A Comparison of Different Test Statistics. *Hum. Hered.* **71**, 23–36 (2011).
120. Gao, F. *et al.* XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. *J. Hered.* **106**, 666–671 (2015).
121. Webster, T. H. *et al.* Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience* **8**, giz074 (2019).
122. Sofer, T. *et al.* A powerful statistical framework for generalization testing in GWAS, with application to the HCHS/SOL. *Genet. Epidemiol.* **41**, 251–258 (2017).
123. Huang, Q. Q. *et al.* Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. *Nat. Commun.* **13**, 4664 (2022).

124. Kaseniit, K. E., Haque, I. S., Goldberg, J. D., Shulman, L. P. & Muzzey, D. Genetic ancestry analysis on >93,000 individuals undergoing expanded carrier screening reveals limitations of ethnicity-based medical guidelines. *Genet. Med.* **22**, 1694–1702 (2020).
125. Khan, A. T. *et al.* Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: Experiences from the NHLBI TOPMed program. *Cell Genomics* **2**, 100155 (2022).
126. Belbin, G. M. *et al.* Toward a fine-scale population health monitoring system. *Cell* **184**, 2068-2083.e11 (2021).
127. Peterson RE *et al.* The utility of empirically assigning ancestry groups in cross-population genetic studies of addiction. *Am J Addict* **26**, 494–501 (2017).
128. Martschenko, D. O., Wand, H., Young, J. L. & Wojcik, G. L. Including multiracial individuals is crucial for race, ethnicity and ancestry frameworks in genetics and genomics. *Nat. Genet.* 1–6 (2023) doi:10.1038/s41588-023-01394-y.
129. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science* **376**, 250–252 (2022).
130. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biol.* **20**, 159 (2019).
131. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
132. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
133. The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018).
134. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
135. Kelleher, J. *et al.* Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
136. Marmot, M. Social determinants of health inequalities. *The Lancet* **365**, 1099–1104 (2005).
137. Marmot, M. & Allen, J. J. Social Determinants of Health Equity. *Am. J. Public Health* **104**, S517–S519 (2014).
138. Hunter, D. J. Gene–environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
139. Salas, L. A. *et al.* A transdisciplinary approach to understand the epigenetic basis of race/ethnicity health disparities. *Epigenomics* **13**, 1761–1770 (2021).
140. Cerutti, J., Lussier, A. A., Zhu, Y., Liu, J. & Dunn, E. C. Associations between indicators of

socioeconomic position and DNA methylation: a scoping review. *Clin. Epigenetics* **13**, 221 (2021).
141. Yousefi, P. D. *et al.* DNA methylation-based predictors of health: applications and statistical considerations. *Nat. Rev. Genet.* **23**, 369–383 (2022).
142. Sanderson, E. *et al.* Mendelian randomization. *Nat. Rev. Methods Primer* **2**, 1–21 (2022).
143. Bonomi, L., Huang, Y. & Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* **52**, 646–654 (2020).
144. Arora, A. Synthetic data: the future of open-access health-care datasets? *The Lancet* **401**, 997 (2023).
145. Ghalebikesabi, S. *et al.* Mitigating statistical bias within differentially private synthetic data. in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence* 696–705 (PMLR, 2022).
146. Yang, G., Mishra, M. & Perera, M. A. Multi-Omics Studies in Historically Excluded Populations: The Road to Equity. *Clin. Pharmacol. Ther.* **113**, 541–556 (2023).
147. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
148. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
149. Carrot-Zhang, J. *et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell* **37**, 639-654.e6 (2020).
150. Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annu. Rev. Stat. Its Appl.* **8**, 141–163 (2021).
151. Mathieson, I. & Scally, A. What is ancestry? *PLOS Genet.* **16**, e1008624 (2020).
152. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
153. Márquez-Luna, C. *et al.* Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nat. Commun.* **12**, 6052 (2021).
154. Busby, G. B. *et al.* Ancestry-specific polygenic risk scores are risk enhancers for clinical cardiovascular disease assessments. *Nat. Commun.* **14**, 7105 (2023).
155. Fuat, A. *et al.* A polygenic risk score added to a QRISK®2 cardiovascular disease risk calculator demonstrated robust clinical acceptance and clinical utility in the primary care setting. *Eur. J. Prev. Cardiol.* **31**, 716–722 (2024).
156. Samani, N. J. *et al.* Polygenic risk score adds to a clinical risk score in the prediction of cardiovascular disease in a clinical setting. *Eur. Heart J.* **45**, 3152–3160 (2024).

157. World Health Organization. A conceptual framework for action on the social determinants of health. World Health Organization (2010).

158. Kowal E, Greenwood A, McWhirter RE. All in the Blood: A Review of Aboriginal Australians' Cultural Beliefs About Blood and Implications for Biospecimen Research. *Journal of Empirical Research on Human Research Ethics*. 10(4):347-359 (2015).

159. Yao, R.A., Akinrinade, O., Chaix, M. *et al.* Quality of whole genome sequencing from blood versus saliva derived DNA in cardiac patients. *BMC Med Genomics* **13**, 11 (2020).

160. Boscarino, N., Cartwright, R.A., Fox, K. *et al.* Federated learning and Indigenous genomic data sovereignty. *Nat Mach Intell* **4**, 909–911 (2022). https://doi.org/10.1038/s42256-022-00551-y

161. Diaz-Papkovich, A., Anderson-Trocmé, L. & Gravel, S. A review of UMAP in population genetics. *J Hum Genet* **66**, 85–91 (2021). https://doi.org/10.1038/s10038-020-00851-4

162. Kamiza, A.B., Toure, S.M., Vujkovic, M. *et al.* Transferability of genetic risk scores in African populations. *Nat Med* **28**, 1163–1166 (2022). https://doi.org/10.1038/s41591-022-01835-x

163.D. Heckerman, D. Gurdasani, C. Kadie, C. Pomilla, T. Carstensen, H. Martin, K. Ekoru, R.N. Nsubuga, G. Ssenyomo, A. Kamali, P.Kaleebu, C. Widmer, M.S. Sandhu, Linear mixed model for heritability estimation that explicitly addresses environmental variation, *Proc. Natl. Acad. Sci. U.S.A.* 113 (27) 7377-7382, https://doi.org/10.1073/pnas.1510497113 (2016).

164. Sun, Q., Rowland, B.T., Chen, J. *et al.* Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. *Nat Commun***15**, 1016 (2024). https://doi.org/10.1038/s41467-024-45135-z

165. Bitarello, B.D., Mathieson, I., Polygenic Scores for Height in Admixed Populations, *G3 Genes|Genomes|Genetics*, 10 (11), 4027–4036 (2020)

166. Haynes, W.A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci Rep* **8**, 1362 (2018). https://doi.org/10.1038/s41598-018-19333-x

167. O'Connor, Timothy D., et al. "Rare variation facilitates inferences of fine-scale population structure in humans." *Molecular Biology and Evolution* 32 (3) 653-660 (2015).

168. Fan, C., Mancuso, N., & Chiang, C. W. A genealogical estimate of genetic relationships. *The American Journal of Human Genetics*, *109*(5), 812-824 (2022).

169 Kelleher, J., Wong, Y., Wohns, A.W. *et al.* Inferring whole-genome histories in large population datasets. *Nat Genet* **51**, 1330–1338 (2019). https://doi.org/10.1038/s41588-019-0483-y

170. Speidel, L., Forest, M., Shi, S. *et al.* A method for genome-wide genealogy estimation for

thousands of samples. *Nat Genet* **51**, 1321–1329 (2019). https://doi.org/10.1038/s41588-019-0484-x

171. Zhang, B.C., Biddanda, A., Gunnarsson, Á.F. *et al.* Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat Genet* **55**, 768–776 (2023). https://doi.org/10.1038/s41588-023-01379-x

172. Kessler, M., Yerges-Armstrong, L., Taub, M. *et al.* Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. *Nat Commun* **7**, 12521 (2016). https://doi.org/10.1038/ncomms12521

173. Burgess, S., Foley, C.N., Allara, E. *et al.* A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat Commun* **11**, 376 (2020). https://doi.org/10.1038/s41467-019-14156-4

174. Rattray, N.J.W., Deziel, N.C., Wallach, J.D. *et al.* Beyond genomics: understanding exposotypes through metabolomics. *Hum Genomics* **12**, 4 (2018). https://doi.org/10.1186/s40246-018-0134-x

175. Bak, M., Madai, V.I., Celi, L.A. *et al.* Federated learning is not a cure-all for data ethics.*Nat Mach Intell* **6**, 370–372 (2024). https://doi.org/10.1038/s42256-024-00813-x

176. Alderman, Joseph E., et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *The Lancet Digital Health* 7(1) e64-e88 (2025):.

177. Pfohl, S.R., Cole-Lewis, H., Sayres, R. *et al.* A toolbox for surfacing health equity harms and biases in large language models. *Nat Med* **30**, 3590–3600 (2024). https://doi.org/10.1038/s41591-024-03258-2

178. Mello, M. M., & Wolf, L. E. The Havasupai Indian tribe case--lessons for research involving stored biologic samples. *The New England journal of medicine*, *363*(3), 204–207. (2010).

179. Lee, S. S., Cho, M. K., Kraft, S. A., Varsava, N., Gillespie, K., Ormond, K. E., Wilfond, B. S., & Magnus, D. "I don't want to be Henrietta Lacks": diverse patient perspectives on donating biospecimens for precision medicine research. *Genetics in medicine : official journal of the American College of Medical Genetics*, *21*(1), 107–113. (2019). https://doi.org/10.1038/s41436-018-0032-6

180. Israel, B. A., Coombe, C. M., Cheezum, R. R., Schulz, A. J., McGranaghan, R. J., Lichtenstein, R., Reyes, A. G., Clement, J., & Burris, A. Community-based participatory research: a capacity-building approach for policy advocacy aimed at eliminating health disparities. *American journal of public health*, *100*(11), 2094–2102. (2010). https://doi.org/10.2105/AJPH.2009.170506

181. Kaye J. The tension between data sharing and the protection of privacy in genomics research. *Annual review of genomics and human genetics*, *13*, 415–431. (2012). https://doi.org/10.1146/annurev-genom-082410-101454