TYPE Brief Research Report
PUBLISHED 18 June 2025
DOI 10.3389/fgene.2025.1559496

OPEN ACCESS

EDITED BY
Kristina Allen-Brady,
The University of Utah, United States

REVIEWED BY
Ahmed Rebai,
Centre of Biotechnology of Sfax, Tunisia
Rachel Mester,
University of California, Los Angeles,
United States

*CORRESPONDENCE
Emile R. Chimusa,
emile.chimusa@northumbria.ac.uk

RECEIVED 12 January 2025
ACCEPTED 19 May 2025
PUBLISHED 18 June 2025

CITATION
Mugo JW, Mulder N and Chimusa ER (2025) Data simulation to optimize frameworks for genome-wide association studies in diverse populations.
Front. Genet. 16:1559496.
doi: 10.3389/fgene.2025.1559496

Data simulation to optimize frameworks for genome-wide association studies in diverse populations

Jacquiline W. Mugo[1], Nicola Mulder[2] and Emile R. Chimusa[3]*

[1]Allergology and Clinical Immunology Unit, Department of Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [2]Division of Computational Biology, Department of Biomedical Sciences, Medical School Cape Town, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa, [3]Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle Upon Tyne, United Kingdom

Whole-genome or genome-wide association studies (GWAS) have become a fundamental part of modern genetic studies and methods for dissecting the genetic architecture of common traits based on common polymorphisms in random populations. It is hoped that there would be many potential uses of these identified variants, including a better understanding of the pathogenesis of traits, disease risk prediction, discovery of biomarkers, and clinical prediction of drug treatments for populations and global health. Questions have been raised about whether associations that are largely discovered in European ancestry populations are replicable in diverse populations, can inform medical decision-making globally, and how efficiently current GWAS tools perform in populations of high genetic diversity, multi-wave genetic admixture, and low linkage disequilibrium, such as African populations. Here, we discuss some of the challenges in association mapping and leverage genomic data simulation to mimic structured African, European, and multi-way admixed populations to evaluate the replicability of association signals from current state-of-the-art GWAS tools. We use the results to discuss optimized frameworks for the analysis of GWAS data in diverse populations. Finally, we outline the implications, challenges, and opportunities these studies present for populations of non-European descent.

KEYWORDS

GWAS, whole-genome sequencing, genetic diversity, genetic risk, admixture, admixture mapping, population genetics

Highlights

- Variability in the predictive methods and categorization of functionally relevant genetic variants pose specific challenges in GWAS of diverse populations.
- Current GWAS mixed models may not fully control for sub-variant structure between affected and unaffected samples, especially if there is an environmental component to phenotypic associations with ancestry at local variants or locus-specific ancestry due to admixture and inadequate or closely related reference ancestral populations.
- Methodological development is still needed to directly control for local-specific ancestry tracts in variant-level GWAS, which may further improve the power and reduce false positives in mixed-ancestry or multi-ancestry samples.

Introduction

The frequent occurrence of population differences in phenotypic outcomes, drug, and treatment responses has important consequences for biomedical sciences. This has been shown to be result of variations in host genomes and differing environments (Tishkoff and Williams, 2002; Goetz et al., 2014; Mao et al., 2007; Martin et al., 2017; Campbell and Tishkoff, 2008). For more than two decades, genome-wide association studies (GWAS) have been common approaches in genetic studies to identify molecular variants underlying these variations and have been used successfully for detecting variants in linkage disequilibrium (LD) within causal genes (direct association) or genes associated to causal genes (indirect association) (Martin et al., 2017; Campbell and Tishkoff, 2008; Seldin et al., 2011). These approaches have become a fundamental part of modern genetic studies, and methods for dissecting the genetic architecture of common traits based on common polymorphisms in different populations have been developed (Seldin et al., 2011). They have enhanced our knowledge on the genetic architecture of many complex diseases, such as heritability estimation, the individual's genetic reliability of disease (polygenic risk scores, PRSs), the genetic correlation between diseases, number of loci, and their effect sizes (Brody et al., 2017; Chimusa et al., 2018; Sirugo and Tishkoff, 2019). So far, many new genetic associations to diseases have been identified (Chimusa et al., 2018; Sirugo and Tishkoff, 2019). Currently, approaches developed to identify the association between genetic variability and human phenotypes have mostly been designed to capture genomes with a long range of LD and haplotypes, such as those found in European descent populations, who have mostly undergone a population bottleneck (Seldin et al., 2011; Chimusa et al., 2018). For example, a review by Sirugo and Tishkoff (2019) of GWAS diversity in the GWAS catalog up to January 2019 revealed that the reported GWAS were dominated by two populations, with Europeans accounting for 52% and Asian populations accounting for 21% (Sirugo and Tishkoff, 2019). Although the contribution by the Asian population was commendable, further analysis of the individuals in the GWAS by ethnicity revealed a persistent gap: 78% were of European ancestry, 10% were of Asian origin, 10% were from Africa, and 1% were of Hispanic origin, while other ethnicities accounted for less than 1% of the individuals (Sirugo and Tishkoff, 2019; Mills and Rahal, 2019; Visscher et al., 2017). These have raised concerns about healthcare disparity when GWAS results are translated into clinical relevance for global health. Mills and Rahal (2019) analyzed 3,639 GWAS and found that 86.03% of discovery, 76.69% of replication, and 83.19% of combined ancestry diversity in GWAS were mostly from individuals of European descent. The finding corroborates those of other studies on Asian descent populations (Kim et al., 2011) (9.92% discovery, 17.97% replication, and 12.37% combined), African American or Afro-Caribbean populations (1.96% discovery, 1.96% replication, and 1.96% combined), Hispanic or Latin American populations (1.30% discovery, 1.33% replication, and 1.30% combined), other or mixed-ancestry individuals (0.48% discovery, 1.77% replication, and 0.87% combined), and African ancestry populations (0.31% discovery, 0.28% replication, and 0.30% combined). These studies suggest the inclusion of diversity in data and recognize the consequences of the lack of diversity. Kim et al. (2018) examined the risk posed by genetic disease across global populations using GWAS and showed that the ancestral risk allele discovered is 5.1% higher and the derived risk allele discovered is 5.40% lower in African populations. Further investigation using different populations showed that non-African groups yielded disease associations that have biased allele frequencies, while the African populations yielded disease associations that are relatively free of bias. Caution must, therefore, be taken when using GWAS results from one population to predict disease risk in another.

Research continues to reveal that current GWAS results from European cohorts cannot be generalized to diverse populations due to confounding environmental factors across populations, differing patterns of LD, differences in allelic architecture, and other contributing factors (Visscher et al., 2017; Shriner et al., 2011a; Shriner et al., 2011b; Shriner, 2017). Significant differences in European and diverse populations have also been observed in the genetic determinants of both common and rare diseases and their effect sizes (Visscher et al., 2017; Shriner et al., 2011a; Shriner et al., 2011b; Shriner, 2017). Nevertheless, GWAS are now slowly being extended to diverse populations. Non-European populations are now included in large disease-analysis studies, and new consortia have been established in countries with diverse populations. The Human Heredity and Health in Africa (H3Africa) consortium has spearheaded GWAS on the African continent (Mulder et al., 2018). The INdian DIabetes COnsortium (INDICO) (INdian DIabetes COnsortium, 2011) and the GenomeAsia 100K Project (Author anonymous, 2019) are other examples of consortia in diverse populations.

On the other hand, it is still being observed that large numbers of modern drugs approved by the Food and Drug Administration and similar organizations have been developed with relevance to European ancestry populations while not addressing the fact that subtle differences in the genetic make-up of other populations, such as Asian, South American, and African populations, can affect drug efficacy or response (Visscher et al., 2017; Hassan et al., 2020). This has been evidenced by the fact that hundreds of thousands of people still die each year because of adverse drug reactions, which may result from multiple factors, including disease determinants, environmental exposure, human microbiome profiles, and genetic factors (Visscher et al., 2017; Hassan et al., 2020). The use of genetic information to inform medical decision-making, however, raises questions as to whether such use could be equitable. Given differences in allelic architecture, differing patterns of LD, and confounding of environmental factors across populations, the richer mixtures of African genetic variants and differing environments are likely to contribute to wider phenotypic and individual microbiome profile variability (Tishkoff and Williams,

Abbreviations: EMMAX, Efficient Mixed-Model Association eXpedited; FAIR, findability, accessibility, interoperability, and reusability; FPR, false-positive rate; GEMMA, genome-wide efficient mixed model association; GCTA, genome-wide complex trait analysis; GWAS, genome-wide association studies; JasMAP, joint ancestry and SNP association method for a multi-way admixed population; LAI, local ancestry inference; LD, linkage disequilibrium; LMM, linear mixed model; LOCO, leave one chromosome out; PC(s), principal components; PCA, principal component analysis; PheGenI, Phenotype−Genotype Integrator; PRS, polygenic risk score; SNP, single-nucleotide polymorphism; TRS, transcriptional risk score.

2002; Awany et al., 2018). It is, therefore, crucial to advance GWAS research and assess how well current approaches can capture diverse global population cohorts (Campbell and Tishkoff, 2008; Chimusa et al., 2018; Visscher et al., 2017).

We hypothesize that understanding and then appropriately modeling different aspects of genetic architecture in the African population has the potential to achieve unbiased and powerful estimates of genetic risk in them, as well as in multi-ethnic and admixed populations such as African Americans. Here, we leverage genomic data simulation that mimics African, European, and multi-way admixed populations to evaluate the replicability of association signals from current state-of-the-art GWAS tools. We dissect reasons from the biological and methodological perspectives that account for the replicability of GWAS and identify the challenges ahead. In contrast to the exemplary success of disease gene discovery, currently, GWAS findings are not fully useful for predicting phenotypes. Finally, we provide an overview of the prospects for individual prediction of disease risk and its foreseeable impact on clinical practice in populations of non-European descent.

Genome disease mapping

The delineation of health and complex diseases from polymorphism-based association mapping holds promise to bridge the gap between clinical translation and statistical association, thereby improving diagnostics, screening, genetic testing, and counseling in global clinical populations (Sirugo and Tishkoff, 2019; Mills and Rahal, 2019; Martin et al., 2019). It has been shown that variants associated with diseases found in populations of European ancestry do not always replicate in diverse populations such as African populations (Chimusa et al., 2018; Kim et al., 2018; Martin et al., 2019) for several reasons, including confounding of environmental factors across populations, differing patterns of LD, and differences in the allelic architecture.

Although efforts such as the China Kadoorie Biobank (Chen et al., 2011), Global Screening Array (GSA) (Nelson et al., 2017; Kalra et al., 2018), Multi-Ethnic Global Array (MEGA) (Nelson et al., 2017), and H3Africa array (Mulder et al., 2018) have recently enabled effective genome-wide DNA microarrays in diverse populations, there are still many issues in GWAS, such as (1) GWAS small sample size in diverse populations, including in African populations (Swart et al., 2022); (2) stratification due to the correlation of environmental exposures and genetic correlation background due to common ancestry or multi-wave admixture and pre-/post-admixture selection pressure (Chimusa et al., 2018); (3) translation of associated loci into suitable biological hypotheses (Chimusa et al., 2018); (4) the understanding of how multiple modestly associated loci within genes interact to influence a phenotype (Chimusa et al., 2018). Control of population stratification in GWAS has been one of the biggest concerns to ensure that observed associations reflect the genetic effects of each genomic locus rather than correlations with ancestry (Author anonymous, 2019; Korte and Farlow, 2013).

For decades, mixed-model approaches have been attractive in GWAS as they allow the inclusion of all samples irrespective of ancestry. Mixed-model approaches control for population stratification by modeling distant relatedness between samples due to ancestry (Korte et al., 2012). Several implementations exist, and we list some of them chronologically up to the latest in 2020 in Figure 1. Mixed models may yield greater statistical power, both through increased sample size and by controlling for the variance explained by the genetic relatedness between individuals (i.e., a random-effect component) (Zhou and Stephens, 2012; Chimusa et al., 2014). However, there is evidence that variants with low frequency (1%–5%) or at the boundary may not often attain genome-wide significance in mixed models due to their imperfect asymptotic distribution (Chimusa et al., 2014). In addition, they may not fully control for sub-variant structure between affected and unaffected samples, especially if there is an environmental component to phenotypic associations with ancestry at local variants or locus-specific ancestry due to admixture (Seldin et al., 2011; Winkler et al., 2010; Brody et al., 2017; Shriner et al., 2011b).

Non-genetic factors such as environmental exposures may be correlated with genetic ancestry due to the shared local environment (familiar or community effects) or the relationship between ancestry and sociocultural factors such as ethnicity and religious background (McGrath et al., 2013). Effective methodological development is still needed to directly control local-specific ancestry tracts in variant-level GWAS, which may further improve power and reduce false positives in mixed-ancestry or multi-ancestry samples (Visscher et al., 2017; McGrath et al., 2013; Marigorta et al., 2018). However, it is worth noting that several efforts and advances have been made in leveraging the complementarity of single-nucleotide polymorphism (SNP) association signals captured through GWAS and admixture signals to calibrate and improve GWAS power in admixed populations (Shriner et al., 2011a; Shriner et al., 2011b; Shriner, 2017; Chimusa et al., 2014). In doing so, numerous studies have leveraged local-specific ancestry tracts in variant-level association analyses to African Americans (Peprah et al., 2015), Latinos (Gonzalez et al., 2016), South African Coloureds (Chimusa et al., 2014), and Hispanic cohorts (Chen et al., 2015), demonstrating added value beyond standard association testing (Geza et al., 2019; Chimusa et al., 2016). Admixture association is critically reliant on accurate locus ancestry inference (LAI), which requires well-specified founding population reference samples (Chakraborty and Weiss, 1988). Power can be optimized by combining admixture mapping and association testing (Kim et al., 2018), but this approach is rarely adopted because of the multi-stage process required and the challenge in application to complex multi-way admixed samples (Thornton and Bermejo, 2014).

For example, currently available joint ancestry and SNP association methods (Mugo et al., 2023; Atkinson et al., 2020; Atkinson et al., 2021) have primarily been tailored to African American populations (two-way admixed populations), despite the authors' suggestion that, in theory, these methods could be extended to more than three-way admixed populations. This has either been left for future research or not optimized for multi-way admixed populations in most of these tools. In addition, the accuracy of these approaches relies on the use of accurate knowledge of ancestry inference, particularly when applying to complex multi-way admixed populations (Mugo et al., 2023). Although recently introduced locus-specific ancestry methods for multi-way admixed populations achieved equivalent accuracy (Geza

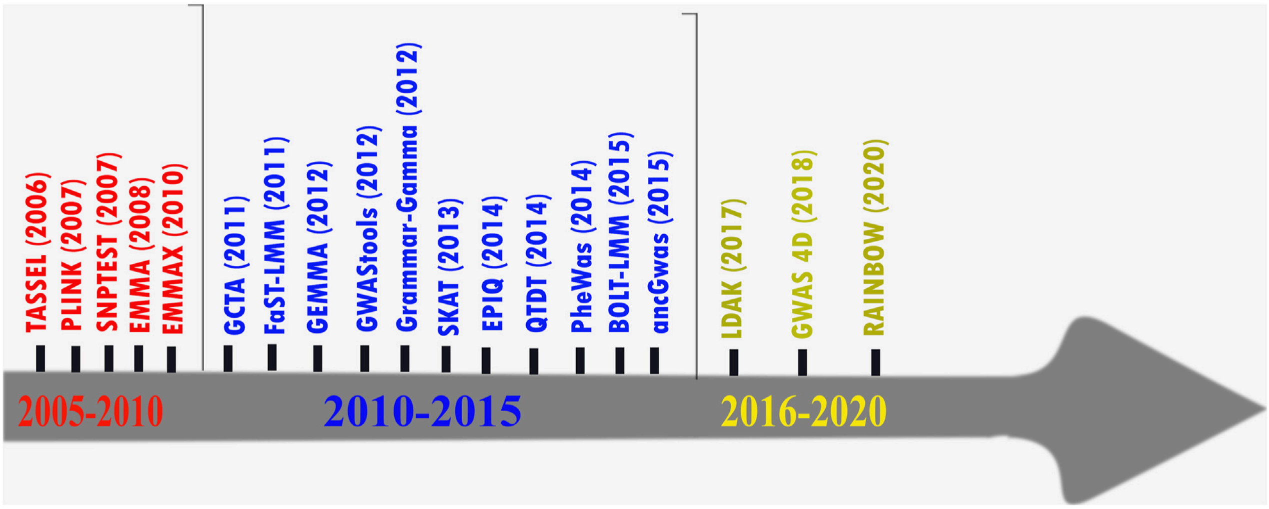

FIGURE 1
Examples illustrating the chronological order of a partial list of GWAS tools.

et al., 2019; Honorato-Mauer et al., 2025; Sun et al., 2025), they still suffer from spurious deviations in average local ancestry at some chromosomal locations of cases and controls, where the modeled ancestral population is unusually different from the true ancestral population due to historical actions such as natural selection and also inaccurate or closely related reference ancestral populations (Chimusa et al., 2014; Geza et al., 2019; Sun et al., 2025). Furthermore, application of these methodologies to multi-way admixed populations, like most African populations known to have high diversity within differing environments, remains mostly less explored (Author anonymous, 2019). There remains a critical need to (1) improve LAI accuracy (Chimusa et al., 2014; Geza et al., 2019; Thornton and Bermejo, 2014; Mugo et al., 2023; Atkinson et al., 2020); (2) develop methods for optimizing the power of association testing (Chimusa et al., 2014; Geza et al., 2019; Mugo et al., 2023; Atkinson et al., 2020; Atkinson et al., 2021) and PRSs in admixed data (Sirugo and Tishkoff, 2019; Coram et al., 2017; Marnetto et al., 2020); (3) build integrative software for running a multi-stage admixture analysis pipeline (Geza et al., 2019).

Replication in genetic association studies

In general, the replication of reported results is the most reliable validation of scientific discoveries (National Academies of Sciences et al., 2019; Kraft et al., 2009). In complex trait genetics, replication occurs when the same genetic marker is consistently associated with the same phenotype in independent cohorts. In addition, a positive association at variants in strong LD with the original marker SNP is often used as evidence of replication, even if formal exact replication is not achieved. Cross-ancestry replicability has the power to shed light on the genetic architecture of complex traits, informing the reliability of effect estimations and their variability across human ancestries.

GWAS meta-analysis has increasingly that leverages association summary statistics to facilitate and encourage *silico* replication to maintain reliability in genetic association findings has increasingly been adopted. A meta-analysis framework combines results from different GWAS cohorts into a single analysis framework to recover signals that a single GWAS cohort study might miss and to address the between-study and between-population heterogeneity (Chimusa et al., 2016). Recently, meta-analysis has shown remarkable discovery results and helped us better understand and validate association results from different studies. Meta-analysis is considered a post-genome-wide association study method; however, heterogeneity among GWAS meta-analyses remains an issue, particularly with increasing number of studies (Kraft et al., 2009). Variation in the cohort size across independent studies is challenging, especially when these studies are conducted from distinct populations of different ancestry and patterns of LD (Kraft et al., 2009). Similarly, the list of new post-GWAS tools, such as multi-marker analyses, which go beyond single-SNP tests, or the inclusion of functional evidence to reweight GWAS results, is growing by the day (Han and Eskin, 2011). The heterogeneity in these methodologies will necessarily complicate the evaluation of replicability.

Another caveat is that GWAS conducted in non-European ancestry populations usually include fewer samples (Mieth et al., 2016), making the current picture of genetic association to disease across diverse populations incomplete (Manolio, 2017). This creates a challenge for the power of GWAS meta-analysis across diverse population cohorts of differing genetic ancestry. Moreover, caution is required as incomplete replication can also be informative; several studies have reported a lack of interpopulation replicability, indicating that some risk variants are population-specific (Martin et al., 2017; Nakagome et al., 2012). For example, comparing Asian and European associations with major depression, the failure to replicate is largely due to differences in patterns of LD, which reduced power in one population since the proportion of attributable risk decreases with population-specific minor allele frequency (Wang et al., 2007).

GWAS application for disease risk prediction

The exceptional polygenicity of human traits makes unraveling mechanisms from whole-genome or GWAS daunting (Mills and Rahal, 2019). PRSs, which estimate an individual's genetic liability to disease or traits compared to others with a different genetic constitution (Wray et al., 2021; Choi et al., 2020), are still mostly derived from European ancestry GWAS data, making their predictive power substantially lower when computed in non-European samples, particularly those of African ancestry (Sirugo and Tishkoff, 2019). Furthermore, the development of disease association studies and PRS methods (Choi et al., 2020; Euesden et al., 2015; Choi and O'Reilly, 2019; Vilhjálmsson et al., 2015), their applications to understand disease etiology, and their evaluation for clinical utility have been explored almost entirely in European ancestry populations (Sirugo and Tishkoff, 2019).

PRS portability and generalizability have been widely reported in recent years (Martin et al., 2017), yet PRSs in non-European ancestry samples are still routinely calculated using the same European GWAS data and PRS methods as applied to Europeans. This takes no account of known population genetic factors affecting the data, such as marked LD differences, genetic drift, natural selection, daily nutrition, family history, and gene–environment interactions. Consequently, the clinical utility and etiological insights provided by PRS may have little relevance to Africans and African Americans. In addition, PRS calculations are inherently dependent on the quality of the underlying GWAS data. If the GWAS used to derive the PRS model is underpowered or has biases, the PRS may not be accurate or reliable and makes its application to a multi-way admixed population even worse. Furthermore, current PRS methods are limited in their ability to integrate epigenetic factors and interactions between different genetic regions. Although PRS can summarize the effects of individual genetic variants, they do not yet account for how gene expression may be regulated by epigenetic modifications or how different genetic variants may interact to influence disease risk. While it is worth noting the advances currently being made in more sophisticated PRS models (Wray et al., 2021; Klau et al., 2023), there is an increasing risk of overfitting, particularly when the models incorporate many genetic variants. Overfitting occurs when a model is too closely aligned with the training data, resulting in a model that does not generalize well to new data sets (Wray et al., 2021; Klau et al., 2023). There is still a need to ensure that PRS models maintain robustness and a similar magnitude of accuracy in different populations through careful validation and cross-validation strategies, as well as by integrating an explainable predictive model within PRS (Klau et al., 2023).

All these raise the question as to how the clinical utility of these methods can be made equitable across multi-ethnic populations and, specifically, how to accurately predict health and disease risk in multiway admixed and African populations.

Leveraging the data simulation framework to dissect GWAS in diverse populations

Simulation of homogeneous and mixed-ancestry case/control populations with well-known structures that mimic real populations may help better understand the genetic variation of these populations and evaluate different existing GWAS tools of genetic variations undermining complex diseases.

The genetic structure of populations, as well as other controllable factors, including allele frequency and LD patterns of genetic markers, is important for the simulation of genotyping data for GWAS (Mugo et al., 2017). It is important to note that the power of a statistical test to detect a risk locus relies heavily on the allelic spectrum (numbers and frequencies of alleles) and the LD structure around the locus (Mugo et al., 2017). Therefore, it has been suggested that simulated data should possess both local and long-range LD (LRLD) patterns and maintain allelic frequencies like real data (Turgut and Koca, 2024). The resampling approach starts with real data and avoids the use of an evolutionary process. It has been shown that this method has its advantages, compared to other approaches, in retaining real data properties, such as allele frequency and LD of the initial pool data (Mugo et al., 2017).

Simulation of multiple disease loci

To facilitate benchmarking common GWAS tools, we simulated homogeneous and heterogeneous GWAS datasets based on haplotypes from the 1000 Genomes Project spanning the genome and realistic enough to mimic African, European, and admixed populations to challenge the statistical methods for association testing under real-world conditions.

We used resampling and population growth models with recombination breakpoints while mimicking mutation rates as described in our previous simulation tool FractalSIM (Mugo et al., 2017). African and European populations were simulated under a homogeneous simulation model. We merged five European and two West African populations to form the reference population for the simulation. Merged populations, the corresponding sample sizes, and the abbreviations for the populations used are listed in Supplementary Table S1.

We selected 9,139,969 common biallelic SNPs in both European and African populations. Two sets of case/control datasets with an equal number of cases and controls (500 cases, 500 controls and 2,500 cases, 2,500 controls) were simulated for each merged population, while maintaining population substructure (Mugo et al., 2017), in generating simulated genome datasets and mimicking each reference panel. Although it is well known that GWAS power is correlated with increasing samples, these sample sizes were modestly chosen to (1) reflect current African GWAS sample size affordability (Fatumo and Inouye, 2023; Olono et al., 2024) and (2) allow a realistic evaluation of the association power on the most popular and commonly used GWAS tools through comparing GWAS summary statistics in European versus African populations. Our decision here relies on the fact that most GWAS in non-European populations still suffer from small sample sizes, and we, therefore, base our evaluation on the minimum and maximum

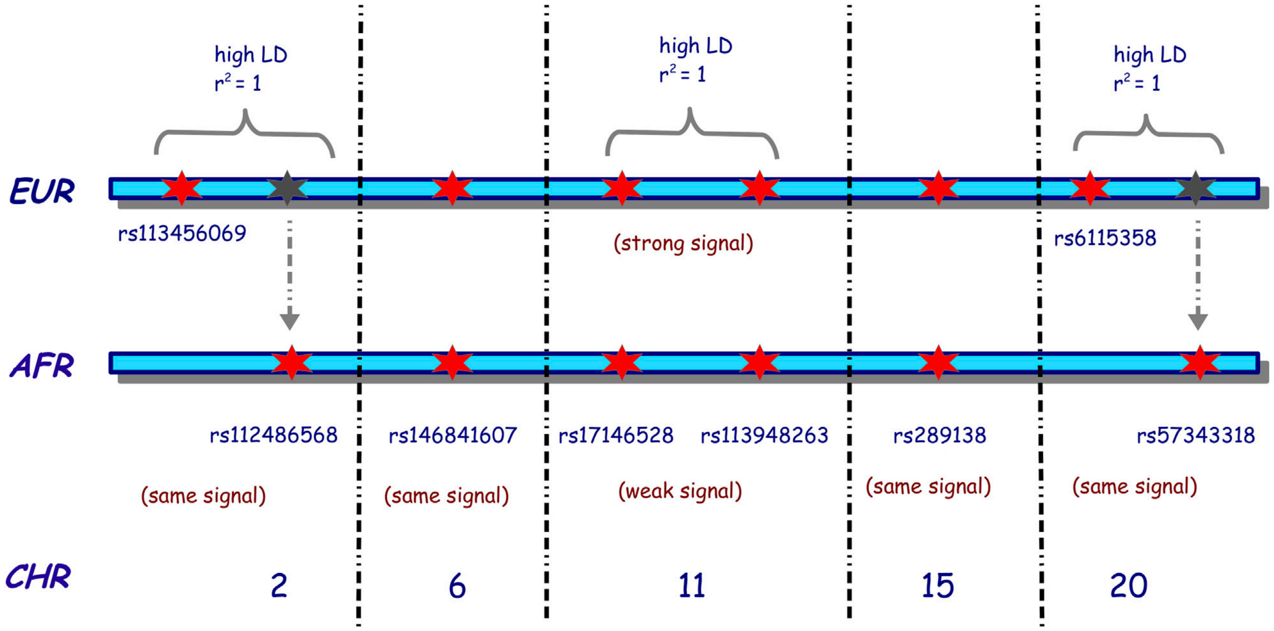

FIGURE 2
Illustration of the choice of the risk SNPs selected for the simulation of homogeneous European and African populations. EUR indicates European, AFR indicates African, and CHR indicates the chromosome. A red star indicates that the SNP was simulated as a risk SNP in that population. In contrast, a gray star indicates that an SNP is present in the European population, in high LD with the risk SNP simulated on that chromosome but simulated as a risk SNP in the African population. The black dotted vertical lines indicate the presence of other chromosomes between the two chromosomes. The risk SNPs on chromosomes 2, 6, 15, and 20 were simulated with the same risk strength in both populations, while on chromosome 11, the risk SNPs of Europeans were simulated with a strong risk signal, and those of Africans were simulated with a weak risk signal.

sample sizes that African studies can currently afford, as indicated in the current literature (Mills and Rahal, 2019; Zhou and Stephens, 2012; Turgut and Koca, 2024; Fatumo and Inouye, 2023; Olono et al., 2024).

To assess the associated risk effect, a total of eight fixed SNPs were simulated with risk effects through all simulation scenarios. These SNPs were randomly selected across the genome, and as such, we simulated the risk effect on SNPs on chromosomes 2, 6, 11, 15, and 20. Figure 2 illustrates our choice of the tag SNPs (representative SNPs that represent a group of SNPs in a genomic region) and risk SNPs across different chromosomes. On chromosome 2, we chose two SNPs, rs113456069 and rs112486568, which were selected, such that they were in complete LD ($r^2 = 1$) in the simulated European dataset. SNP rs113456069 was then simulated with a risk effect in the simulated European population, while rs112486568 was simulated with a risk effect in the simulated African population and no effect in the simulated European population. Both SNPs were simulated with the same signal strength in both simulated datasets. A similar process was applied in simulating SNPs with risk effects on chromosome 20. SNPs rs6115358 and rs7343318 were in complete LD in the European reference dataset, but only rs6115358 was simulated with a risk effect in the simulated European dataset, and rs57343318 was simulated with a risk effect in the African populations. The objective of this design in simulating the risk-effect SNPs on chromosomes 2 and 20 was to enable investigating GWAS replicability from simulated European into simulated African GWAS datasets and evaluate the rate of misassociation and misreplication. The SNPs with risk effects on chromosome 11 were simulated such that they were in complete LD in both simulated datasets, but in the European dataset, they were simulated to have a strong signal, while in the simulated African dataset, they were simulated to have a weak signal. This was to enable examining the power from different tools for capturing weak association signals in simulated African population and understanding if increasing samples may have contributions. On chromosomes 6 and 15, both SNPs were simulated with the same risk signal strength in both simulated datasets. In addition, we considered the same homozygosity and heterozygosity relative risks for these eight risk tag SNPs in all simulation scenarios. The list of these SNPs and the corresponding relative risks in the simulated European and African populations are listed in Supplementary Table S2. The cases and controls were then simulated using a multiple logistic regression model implemented in FractalSIM (Mugo et al., 2017).

The heterogeneous (admixed) datasets were generated under a single-point admixture scenario, where the admixture process occurs at a single point in history, such that the current generation is the offspring of the admixed population that has interbred over subsequent generations. Considering a random mating model where interbreeding has occurred for 10 generations, the admixture simulation first mimicked the isolated growth of each parental (reference) population, where a disease model (risk or null) was simulated in the isolated homogeneous simulation for each of the parental populations, like the case–control homogeneous simulation of the European and African datasets detailed above. At generation 0, the isolated parental populations were allowed to interbreed. We simulated both three-way and five-way admixture scenarios. Supplementary Table

S3 lists the reference parental populations used in these two scenarios, their corresponding initial sample sizes, and the proportion of ancestry contribution from each of these parental/reference populations.

In the three-way simulation, we included 466,142 biallelic SNPs that were the intersection between the three parental populations. We simulated eight risk SNPs as described above and generated 2,500 cases and 2,500 controls. In the five-way admixture scenario, we incorporated 623,330 biallelic SNPs that were present in the intersection between all five parental populations and simulated eight risk SNPs on chromosomes 2, 6, 11, 15, and 20 as described above. In the five-way scenario, however, we simulated two sets of datasets of different sample sizes: a dataset of 500 cases and 500 controls and another of 2,500 cases and 2,500 controls.

We simulated different ancestry risk scenarios on different chromosomes by varying the presence and strength of genotype risk on the risk variant simulated and the ancestry risk on the genomic region containing the variant. We simulated ancestry risk by simulating ancestry deviation between cases and controls in the region that contained risk variants. In the three-way simulation on chromosomes 2 and 11, we simulated strong genotype and ancestry risks; on chromosome 6, we simulated very strong ancestry risk and weak genotype risk; and on chromosome 15, we simulated weak genotype and ancestry risks. All the other chromosomes were simulated under a null model in this scenario. In the five-way simulation, we simulated similar levels of risk in the 500 cases and 500 controls, as well as 2,500 cases and 2,500 control sample sizes. On chromosome 2, we simulated strong genotype and ancestry risks; on chromosomes 6 and 20, we simulated a strong genotype and no ancestry risk; on chromosomes 11 and 15, we simulated weak genotype and ancestry risks, along with a null model on all the other chromosomes. The risk SNPs simulated in the three-way and five-way scenarios, and their respective homozygosity and heterozygosity relative risks specified in cases are listed in Supplementary Table S4. Depending on the MAF of the risk SNPs, the specified risk effects introduced risk signal strength, as indicated in Supplementary Table S5.

Assessing simulated GWAS datasets through population structure

We first assessed the structure of the simulated data for both the homogeneous and admixed populations. Since the simulation process was similar for the two sets of case–control datasets in the homogeneous populations and the five-way admixture simulation, we used the 500 cases and 500 controls for this assessment.

In the three simulated datasets, European, African, and admixed, we first merged the simulated GWAS datasets with their corresponding reference populations used in generating simulated data. We then obtained the first 10 principal components (PCs) using principal component analysis (PCA) implemented in genome-wide complex trait analysis (GCTA) and proceeded to plot the first and second PCs using the GENESIS tool (Buchmann and Hazelhurst, 2014). We used two approaches to assess the global ancestry in the admixture simulation. We first ran the ADMIXTURE tool (Alexander et al., 2009), using the supervised option, for the merged admixed datasets; second, we calculated the simulated global ancestry from the local ancestry block estimates generated through FractalSIM. We then plotted the two admixture tract plots for each scenario using the GENESIS tool.

Supplementary Figures S2, S3 show PCA plots for the African and European population simulations, respectively, while Supplementary Figures S4, S6 show the PCA plots for the simulated three-way and five-way admixed populations, respectively. The admixture tract plots for the three-way and five-way admixture simulations are illustrated in Supplementary Figures S5, S7, respectively. From Supplementary Figures S2, S3, we observed that from the simulated African and European populations, both cases and controls, clustered together, as would be expected in a homogenous population, with no population sub-structures. The simulated populations were also positioned between merged reference populations on the PCA 2 axis for the real African population and the PCA 1 axis for the real European population. On the PCA 1 axis, the simulated African population was very close to the reference population by considering the range of the axis, and similarly, on the PCA 2 axis, the simulated European population was also very close to the reference population, based on the range of the axis. This implies that the simulated cases/controls were genetically close to the respective merged African and European reference samples. For the admixed population, we observed on the PCA plots in Supplementary Figures S4, S6 that the admixed samples were confined within their respective reference parental populations, for both the three-way and five-way populations. Furthermore, we observed that the simulated population was spread out, as would be expected for an admixed population. The simulated three-way admixed population was closer to the YRI population, which contributed 70% of the ancestry, while the five-way admixed population is spread out further away from the EAS population but closer to the MAFR and SAS populations, which contributed larger proportions of the ancestry. For the admixture tract, for both the three-way and five-way scenarios, Supplementary Figures S5, S7 demonstrate that the ADMIXTURE tool estimates the global ancestry close to the true estimates but performs better in a three-way simulation than in a five-way simulation. This is expected as the model accuracy decreases with increasing parameter space (Zhou and Stephens, 2012). The PCA and admixture plots indicate that the structure of the simulated populations met the criteria of the population that we required for the downstream analysis.

Assessing genome association studies

We further examined state-of-the-art and commonly used GWAS tools representing major GWAS models using the simulated GWAS datasets described above. We included linear mixed-model (LMM)-based approaches Efficient Mixed-Model Association eXpedited (EMMAX) (Kang et al., 2010), GCTA (Yang et al., 2011) and genome-wide efficient mixed model association (GEMMA) (Zhou and Stephens, 2012), as well as the most widely used GWAS analysis tools PLINK (Purcell et al., 2007) and SNPTEST (Marchini et al., 2007). For the homogeneous African and European population, we considered the standard PLINK association test under a logistic model that allowed the inclusion of covariates, which we labeled PLINK-Logistic.

For the admixture simulations, we only considered PLINK-Logistic. For GCTA, we considered two association approaches included in the tool. In the first approach, the GRM used includes the chromosome with the SNP being tested for association, which we labeled GCTA, while the second approach uses a GRM that excludes the chromosome that contains the SNP being tested for association, which we label GCTA-LOCO (leave one chromosome out). This approach is an extension of GCTA to eliminate proximal contamination that may be introduced in the association analysis when this chromosome is included in the calculation of the GRM. Similarly, in SNPTEST, we considered both the frequentist association approach (which we refer to as SNPTEST-Frequentist) and the Bayesian approach (which we refer to as SNPTEST-Bayesian). We also assessed TRACTOR (Atkinson et al., 2020; Atkinson et al., 2021) and joint ancestry and SNP association method for a multi-way admixed population (JasMAP) (Mugo et al., 2023), two recently joint SNP and admixture association approaches tailored for admixed populations. TRACTOR and JasMAP require LAIs as input, which we obtained by running RFMIX (Maples et al., 2013). We used RFMIX for ancestry deconvolution as this is the recommended tool by TRACTOR and JasMAP.

First, we obtained the first 10 PCs under each simulated set of data using GCTA. For the homogeneous populations, we included five PCs as covariates when running PLINK-Logistic and SNPTEST; based on the PCA plots, we did not observe any population structure in the homogeneous cohorts. In the admixture populations, however, we included 10 PCs as covariates in the association test to control for global ancestry. No missingness was observed in the datasets, and all the simulated samples were retained for the association analysis.

We considered only common SNPs when running the association tests. We thus ran the association analysis using eight disease-scoring statistics for the homogeneous population and seven for the admixed population. We then obtained the corresponding summary statistics and Manhattan plots. To correct for multiple tests, we used the Bonferroni correction approach. Since the number of SNPs in our homogeneous population was >1, 000, 000, we used a genome-wide significance of 5.0×10^{-08} for all the frequentist tests. The significance threshold for the three-way admixed population was 1.57×10^{-07}; for the five-way admixed population, it was 8.48×10^{-08} for the smaller sample size and 8.47×10^{-08} for the larger sample size. We used log(BF) of 4.61 as the significant threshold for the Bayes factor (BF) for the SNPTEST-Bayesian test, using Jeffrey's scale of evidence (Marchini et al., 2007; Kosheleva et al., 2021; Wakefield, 2009; Jeffreys, 1961). JasMAP outputs the posterior probability of association (PPA) as the final summary statistics, and a significance threshold of PPA = 0.5 is used.

Evaluating European and African simulated GWAS

Results in Supplementary Figures S8–S11 are based on the homogeneous European and African populations, and the corresponding summary statistics for the simulated SNPs with risk effects are displayed in Supplementary Tables S6–S21. In both simulations, we observed that, for all the tools assessed with the small sample size, none detected the signals on chromosome 11. However, for the European population, the LMM-based tools, GEMMA, GCTA, and GCTA-LOCO, and the standard PLINK approach captured the signals on four of the chromosomes, while EMMAX, PLINK-Logistic, and SNPTEST detected significant signals on three of the chromosomes. Despite EMMAX and SNPTEST detecting three out of the five simulated risk regions at this sample size for the European population, they eliminated the SNP with simulated risk effects on chromosome 6 from the analysis as part of internal quality controls, and thus, no significant SNP was observed. In comparison, in the African simulated GWAS dataset with the smaller sample size, we observed that all the tools, except for SNPTEST, were only able to capture the signals on chromosomes 2, 6, and 20 at significant levels and the signals on chromosomes 15 only at marginal significance thresholds.

On increasing the sample size for the European population, we observed that all the tools were able to capture the simulated disease signals, and although EMMAX and SNPTEST excluded the risk SNP on chromosome 6 by internal quality control, SNPs in LD with these risk SNPs were captured for this population, and a significant signal was detected. However, in the African population, the signals at chromosomes 15 and 20 showed improvement with increased sample sizes, but with less significant thresholds compared to the European population. We also note that at chromosome 11, where a weak signal was simulated for the African simulation, even with a larger sample size, the signals were still at a marginal significance threshold with all assessed scoring statistics.

The findings suggested that in a homogeneous European population with small sample sizes, GEMMA, GCTA, GCTA-LOCO, and PLINK were more robust in capturing most of the simulated risk variants at significant levels, with PLINK-Logistic following suit. With large sample sizes, all the tools were effective in capturing the simulated risk at significance levels. We also noted that internal quality control checks implemented in EMMAX, SNPTEST-Frequentist, and SNPTEST-Bayesian might remove risk variants, especially in analyses with small sample sizes, and thus miss out significant associations. In addition, our results revealed that with a small sample size, most tools were underpowered to detect some of the risk variants present at a significant level in the African GWAS, and even with an increased sample size, as observed on chromosome 11, some risk variants struggled to reach the stringent standard GWAS threshold in the presence of a signal. Similar significant thresholds were observed for the risk SNPs simulated on chromosome 2, rs113456069 and rs112486568, in the European and African populations, respectively, and similarly on chromosome 20, rs6115358 and rs57343318, in the European and African populations, respectively. It is noteworthy that only SNPs rs113456069 on chromosome 2 and rs6115358 on chromosome 20 were simulated with strong risk effects in the European dataset, but not in the African data set, while SNPs rs112486568 on chromosome 2 and rs7343318 on chromosome 20 were simulated with strong risk effects in the African population, but not in the European dataset (Figure 2). Although SNPs rs113456069 and rs112486568 on chromosome 2 were simulated to be in high LD in the European population, we observed that these

SNPs were also in high LD in the African population. Similarly, SNPs rs6115358 and rs57343318 on chromosome 20 were also in high LD in the African datasets. We thus deduce that if strong risk signals exist in both European and African populations with high-powered studies, cross-population replication is possible using most of the tools assessed. In addition, we noticed that the local LD pulls out several non-risk variants (variants that were not simulated with risk effect) to reach genome-wide significance, suggesting that current tools might not distinguish between background LD (correlation due to nearby markers) and the linkage or correlation to true SNPs with risk effect.

Assessment of association tests from simulated admixed GWAS data

The association tests of GWAS tools assessed using the three-way admixed population simulation are displayed in Supplementary Figure S12, while Supplementary Tables 22–S28 present the summary statistics of the risk SNPs simulated. LMM-based tools EMMAX, GEMMA, and GCTA performed quite similarly in detecting the simulated risk variants and captured the risk variants on chromosomes 2 and 6 at a significant threshold. These three tools detected the risk variants simulated on chromosomes 11 and 15 at marginal significance thresholds; the SNPs in LD with the risk variant on chromosome 11 were detected as significant. GCTA-LOCO, an LMM-based approach, performed quite similarly to PLINK-Logistic, SNPTEST-Frequentist, and SNPTEST-Bayesian in capturing the risk variants on chromosomes 2, 6, and 11 as significant, while capturing the signal on chromosome 15 at a marginal significance threshold. On chromosome 11, however, we note that the four approaches detected a second region that was not simulated with a risk variant and, thus, a false-positive association that could be due to admixture was observed. The four approaches also captured a significant signal on chromosome 12 that was not simulated as significant but detected at a marginal significance threshold using the other tools. We, therefore, noted that the LMM-based approaches EMMAX, GEMMA, and GCTA were more robust in capturing a wide range of population structures, which enabled them to control for any spurious associations. However, GCTA-LOCO, also an LMM-based approach, was ineffective in capturing the sample structures, and we hypothesize that the LOCO approach might have missed accounting for a significant level of the sample structure in the analysis.

The association tests using five-way admixed population simulation are displayed in Supplementary Figures S13–S14, while Supplementary Tables S29–S35 present the summary statistics of the risk SNPs simulated. We observed that for the small sample size of 500 cases and 500 controls, all the tools could capture the simulated risk variants on chromosomes 2, 6, and 20. However, none of the tools captured the risk variants on chromosomes 11 and 15 at a significant level. With a large sample size, we observed that all the tools could capture one of the risk variants on chromosome 11, but the signal at chromosome 15 could still not reach the significance threshold. We thus noted that when the genotype risk was strong, irrespective of the presence and strength of the ancestry association, all the tools were also able to detect the risk variant at a significant level, as observed on chromosomes 2 and 6 in the three-way simulation and on chromosomes 2, 6, and 20 in the five-way simulation analysis. This was true for most tools, even with the smaller sample size in the five-way simulation analysis. However, when the genotype risk was weak and the ancestry risk present was weak or strong, most of the tools were limited in their ability to detect the simulated risk variant at a significant level, as observed on chromosome 15 in the three-way simulation and on chromosomes 11 and 15 in the five-way simulation.

Although GCTA-LOCO, PLINK-Logistic, SNPTEST-Frequentist, and SNPTEST-Bayesian were able to detect the risk SNPs simulated on chromosome 11 in the three-way admixed simulation at significance thresholds, they were limited in capturing the admixture-LD on this chromosome and resulted in spurious association signals, which GEMMA, EMMAX, and GCTA successfully controlled for; however, they detected this risk variant at marginal significance thresholds. By increasing the sample size, one simulated risk SNP on chromosome 11 in the five-way admixed population association was also detected as significant by all tools. The simulated ancestry risk on this chromosome was weak, which implied that the increase in power to detect the risk variant was highly likely due to the increase in sample size and not associated with ancestry risk.

We observed that TRACTOR performed quite similarly to the PLINK-Logistic, GCTA-LOCO, SNPTEST-Frequentist, and SNPTEST-Bayesian in capturing the simulated risk SNPs in the three-way admixed population (Supplementary Figure S15; Supplementary Table S36) and in the smaller sample size for the five-way admixed population (Supplementary Figures S16, S17; Supplementary Tables S37–S38). JasMAP (Supplementary Figures S18–S20) was able to improve the power to detect the risk SNP as significant when both the genotypic and ancestry risk signals were marginal in both simulated three-way (Supplementary Table S39) and five-way admixed datasets (Supplementary Tables S40, S41). Using the larger sample size, TRACTOR was able to capture the risk region that was close to the simulated risk SNP as significant but not the simulated risk SNP. We noted that for quite a number of ancestry backgrounds, TRACTOR was not able to generate a result for most SNPs, possibly due to the fact that it has not yet been optimized for multi-way admixed populations. Overall, joint association implemented in TRACTOR and JasMAP showed significant improvement in association power when the genotype risk effects are strong, irrespective of the strength of the ancestry risk and sample size (Supplementary Table S41).

Our results support the need for better calibrated methods of association in a multi-way admixed population that control population structure not only at a global level but also at a more local level by incorporating the effect of local ancestry.

Evaluating the replication of European GWAS in African populations

To assess the false-positive rate (FPR) for the association analysis for the different tools, using the GWAS summary statistics generated from each tool, we computed the proportion of non-risk SNPs that reached genome-wide significance; FPR =

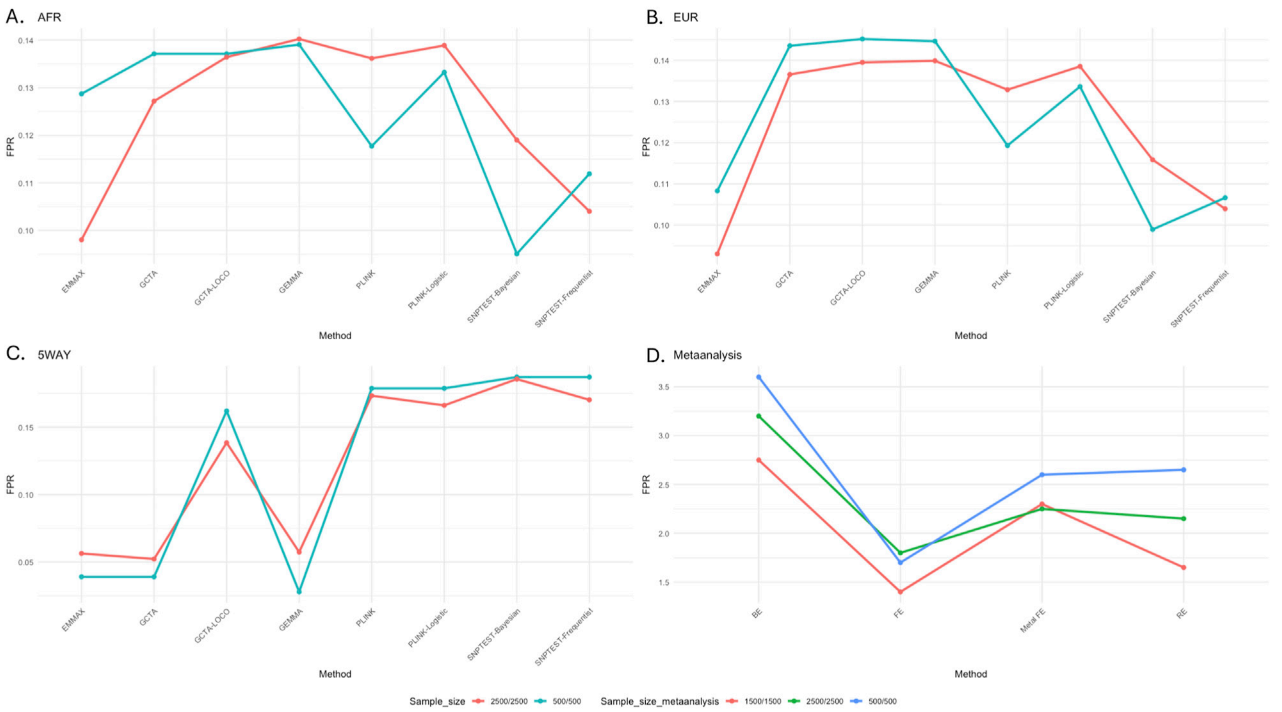

FIGURE 3
Proportion of non-risk simulated variants reaching genome-wide significance across GWAS tools. **(A)** From African (AFR); **(B)** from European (EUR); **(C)** admixed (five-way simulation); **(D)** cross meta-analysis of AFR and EUR simulated datasets.

false positive/(false positives + true negatives). Based on our results, all the tools had some percentages of false positives in the association test of the African (Figure 3A), European (Figure 3B), and admixed (Figure 3C) populations. In the simulated European and African populations, FPR seemed to be lower for EMMAX and SNPTEST, but the difference among the tools was very marginal. In the admixture simulation, based on five-way simulated datasets, it was clear that the LMM-based tools, except for GCTA-LOCO, had relatively lower FPR compared to the other tools. By eliminating a substantial number of SNPs in the calculation of the GRM, the LOCO approach possibly missed accounting for a significant level of the population structure in the analysis, which could have resulted in the relatively high FPR.

To evaluate the replication of European GWAS results in African GWAS, we generated three GWAS (sample sizes 500, 1,500, and 2,500) in each of the simulated African and European datasets. For an easier presentation, we focused on these simulated SNPs on chromosomes 2 and 20 (Figure 2) and used the resulting summary statistics from the LMM model EMMAX to carry out meta-analysis across African and European datasets. We assessed the level of false replication by comparing fixed, random, and binary effects (Figure 3D). The results from cross meta-analysis suggest that replication between European and African GWAS is possible if the effect exists across the two populations.

Considering our findings in Supplementary Tables S6–21, we observe that cross-population GWAS replication through meta-analysis is possible and increases with sample size when a similar magnitude of risk effect exists across populations, and a fixed effect can be applied. Random effects are commonly used in meta-analysis as we often do not know the existence of the risk effect; however, our simulation results show that the random effect may suffer from false positives due to heterogeneity and differing population-specific LD patterns (Figure 3D). In addition, the implemented fixed effect in METASOFT (Han and Eskin, 2011) has fewer false positives than that in METAL (Willer et al., 2010). Our recommendation is to use METASOFT as it enables reporting on fixed, random, and binary effects with m-values [posterior distribution (≥0.9) that effects exist in the GWAS cohort] from each GWAS cohort in the meta-analysis.

What lessons can be drawn from European GWAS to benefit diverse populations?

In the last decade, the increase in human genomic data has led to more than 3,700 GWAS (Welter et al., 2014; Buniello et al., 2019). These studies predicted thousands of genetic risk variants, enabling gene discovery, biological function analysis, and the prediction of the genetic liability of various human phenotypes. The majority of these GWAS have been conducted in European decent populations. The variability in the predictive methods and categorization of functionally relevant genetic variants still pose specific challenges in diverse populations, particularly mixed-ancestry populations (Kim et al., 2018; Shriner et al., 2011a; Shriner et al., 2011b; Shriner, 2017). In addition, GWAS tools and pipelines commonly used in European descent populations may lead to high rates of false-positive/-negative results if stratification is not carefully controlled, particularly in African genomes that harbor the highest genetic diversity and are currently disproportionately under-represented in public databases and reference panels (Buniello et al., 2019). The lesson learnt from various GWAS from European descent populations is that one should consider (i) applying population-specific GWAS pipelines; (ii) choosing appropriate GWAS tool among the existing tools or possibly casting a vote of the association results from running multiple GWAS tools to allow genome-wide level of significance to have a consensus across many tools; (iii) the direction of effect size in each study in meta-analysis with African populations to replicate European GWAS; and (iv) reporting the minor allele frequency, effect size, standard error of the effect size, and LD of the associated variants in African versus non-African populations to enable improved interpretation of the results.

Despite most of the current well-established GWAS tools (Figure 1) being benchmarked using populations of European ancestry, several studies demonstrated that the high genetic diversity found in African and admixed populations makes their genomic studies more likely to detect many novel variants that are yet to be described in current public databases such as the GWAS Catalog (Sirugo and Tishkoff, 2019; Mills and Rahal, 2019; Welter et al., 2014; Buniello et al., 2019) and the Phenotype–Genotype Integrator (PheGenI) (Javed et al., 2014). Due to differing patterns of LD, population-specific allele frequencies, and the proportion of derived/ancestral risk alleles, caution must be used in (1) performing a meta-analysis which combines data from participants across multiple datasets within/between African and non-African populations to analyze millions of variants to increase the power of GWAS and (2) interpreting results from risk prediction and estimation of heritability.

Consistent with other recent studies (Shriner et al., 2011a; Shriner et al., 2011b; Shriner, 2017; Patron et al., 2019), the lesson learnt from various benchmarks of GWAS analysis is that one should consider A) applying significance thresholds and population-specific GWAS pipelines, B) choosing appropriate GWAS tools among the existing tools or considering a consensus approach by possibly running multiple GWAS tools to allow a genome-wide level of significance to have consensus across many tools, C) reporting population-specific minor allele frequency, effect size, standard error of the effect size, and LD of the associated variants in diverse populations to enable improved interpretation of the results, and D) the direction of effect size in each study in meta-analysis with diverse populations to replicate European or other population-specific GWAS.

Although increasing sample size improved the association power for most current GWAS tools, our simulation demonstrated that some risk variants could not reach the genome threshold level. As GWAS extends to diverse populations, the following should be noted:

1) Although increasing sampling may improve the association power, the more the samples, the more genetic variability within such data is swallowed, resulting in current GWAS tools possibly failing to detect some risk variants in large-scale GWAS data (Tam et al., 2019).
2) Diverse or admixture populations may harbor several disease-relevant rare, unique, or population-specific variants compared to Europeans who have undergone population bottlenecks (with more disease-specific common variants); thus, the current GWAS assumption based on "*common disease – common variant*" (Peng and Kimmel, 2007) may have a reduced benefit to diverse populations, mostly characterized by high genetic variation and low LD, or to admixture populations (Tishkoff and Williams, 2002; Martin et al., 2017; Campbell and Tishkoff, 2008; Chimusa et al., 2018; Chimusa et al., 2016; Saint Pierre and Génin, 2014).
3) Despite joint SNP and admixture association tests improving the association power and demonstrating added value beyond standard GWAS tools (Figure 1), they critically rely on accurate LAI, which also requires well-specified founder (reference) populations (Mugo et al., 2023; Atkinson et al., 2020; Atkinson et al., 2021).
4) It remains a challenge to construct appropriate reference or founders' panels for LAI that accurately characterize admixed populations. A consensus has not yet been reached about best practices for reference panels, including the use of continental versus sub-continental reference populations (Geza et al., 2019).
5) LAI methods in multi-way admixed populations may suffer from spurious deviations in average local ancestry at some chromosomal locations of cases/controls, where the modeled ancestral population is unusually different from the true ancestral population, due to historical actions such as natural selection (Chimusa et al., 2014; Secolin et al., 2019). This is still a serious unresolved weakness of admixture association (Chimusa et al., 2014; Secolin et al., 2019; Mani, 2017) in most multi-way admixed populations, which worsens when two or more reference populations are genetically and closely related, resulting in the ancestry being inaccurately assigned or misassigned to admixed individuals (Chimusa et al., 2014; Mugo et al., 2023; Atkinson et al., 2020; Atkinson et al., 2021; Secolin et al., 2019; Mani, 2017).
6) There are very few joint SNP and admixture tools, possibly because of their multi-stage process requiring (A) improved LAI accuracy (Mugo et al., 2023; Atkinson et al., 2020; Atkinson et al., 2021); (B) building integrative software for running multi-way admixture deconvolution analysis (Geza et al., 2019); (C) persisting dilemma in modeling effect sizes conditional on local ancestry, resulting in significant reduction in association power (Mester et al., 2023).
7) It is critical to develop new or adapted pipelines for diverse genetic data or to evaluate existing bioinformatics pipeline tools using diverse populations to account for diverse genetic and environmental characteristics that could differently shape phenotypic variation.

Concluding remarks and future perspectives

GWAS have significantly contributed to medical genomics and understanding of complex traits; however, large numbers of false positives and the small effect size of genetic risk variants have induced a need for calibrated sample sizes and a culture of FAIR (findability, accessibility, interoperability, and reusability) data and sharing. Although GWAS has not yet fully translated into an ability to predict phenotypes in real-world applications based on genetic markers, polygenic and transcriptional risk scores (PRSs and TRSs) for complex diseases hold potential for stratification according to risk, and there is a critical need for new approaches, methodologies, and diverse large data to address questions about the genetic architecture of complex traits and applicability of findings to clinical settings. In this study, we leveraged FractalSIM to generate simulated GWAS datasets mimicking European, African, and admixed populations to evaluate commonly used GWAS tools, as well as newly joint SNP association and admixture tools on their performance, using our simulated GWAS datasets. Our results suggested that LMM-based tools were more robust in capturing risk variants present in the European population with smaller samples but with increased samples in African and admixed populations. All the tools performed similarly and were limited in their ability to capture risk variants present in small sample sizes when using simulated African-specific data. Although increasing the sample size did improve the power to capture risk variants, when the signal was weak, some risk variants still struggled to reach the significance levels set in GWAS. Given the increased frequency of independent testing in simulated African population GWAS analysis due to the generally higher number of SNPs and short LD blocks, it has been suggested and shown that a stricter significance threshold should be considered (Choi et al., 2020). In considering this, it raises the question of whether the risk signals observed at the near marginal significance thresholds in the African GWAS simulated datasets with increased sample sizes would still be significant with more stringent thresholds. This, therefore, emphasizes the dire need for increased sampling in African populations if African GWAS is to catch up with European GWAS, given that small sample sizes still pose a limitation for African GWAS.

When the genotype risk was weak in the presence of a strong ancestry risk, the association between joint SNPs and admixture was successful in leveraging the ancestry risk to enhance the power to detect the signal even when the sample size was small. We observed that joint associations implemented in TRACTOR and JasMAP are calibrated for admixed populations and have significantly improved the association's power in detecting risk effect signals when the genotype risk was strong, irrespective of the strength of the ancestry risk and sample size.

We observed that cross-population replication in the presence of strong risk signals in both European and African populations is possible when applying most of the current state-of-the-art tools from homogeneous population-based association analysis, provided the studies are high-powered. However, caution should be exercised while using EMMAX, SNPTEST-Frequentist, and SNPTEST-Bayesian approaches as internal quality control procedures in these tools may eliminate risk variants from the analysis. Furthermore, we observed that the LMM-based models, except for GCTA-LOCO, performed better at controlling for spurious associations in the admixture context. However, they were limited in detecting the simulated risk variant when the genotype risk was very weak, irrespective of whether the ancestry risk was very high or moderate at the genomic region containing the risk SNP.

Overall, GWAS reproducibility is critical, and it is also important to keep fostering a culture of replication to maintain reliability in findings. As sequencing and, consequently, the availability of genomic information on African populations increases, there are new opportunities to design next-generation disease scoring statistical models that capture not only common variants but also rare and population-specific variants. These new approaches should be tailored to and leverage the characterization of diverse populations with longer histories, high genetic diversity, and environmental heterogeneity, as well as varying types of LD patterns. This will enable us to better understand and elucidate the genetic architecture of African complex traits, variation in drug/treatment response, and disease outcomes. There are potentially many uses of novel disease-scoring statistics models that further leverage the environmental diversity, such as on the African continent, including a better understanding of the pathogenesis of diseases of global health relevance, new leads for studying underlying risk prediction, and advancing clinical prediction of global treatment. The findings from these African-specific disease scoring statistics on African data will pave the way for a new, more diverse research dimension in public health translation[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19].

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

[1] TASSEL, https://www.maizegenetics.net/tassel

[2] PLINK, https://www.cog-genomics.org/plink/

[3] SNPTEST, https://www.chg.ox.ac.uk/~gav/snptest/

[4] EMMA, https://academic.oup.com/genetics/article/178/3/1709/6,061,473

[5] EMMAX, https://genome.sph.umich.edu/wiki/EMMAX

[6] GCTA, https://yanglab.westlake.edu.cn/software/gcta/#Overview

[7] GEMMA, https://github.com/genetics-statistics/GEMMA

[8] FaST-LMM, https://fastlmm.github.io/

[9] GWAStools, https://www.bioconductor.org/packages/release/bioc/html/GWASTools.html

[10] Grammar-Gamma, https://lomereiter.github.io/2015/03/27/overview.html

[11] SKAT, http://www.hsph.harvard.edu/~xlin/software.html

[12] EPIQ, https://bioinformaticshome.com/tools/gwas/descriptions/EPIQ.html#gsc.tab=0

[13] QTDT, https://bioinformaticshome.com/tools/gwas/descriptions/QTDT.html#gsc.tab=0

[14] PheWAS, https://phewascatalog.org/phewas/#home

[15] BOLT-LMM, https://alkesgroup.broadinstitute.org/BOLT-LMM/BOLT-LMM_manual.html

[16] ancGWAS, https://github.com/echimusa/ancGWAS

[17] LDAK, https://dougspeed.com/downloads/

[18] GWAS 4D, http://www.mulinlab.org/gwas4d/gwas4d/gwas4d/gwas4d_about

[19] RAINBOW, https://github.com/KosukeHamazaki/RAINBOWR

Ethics statement

The studies involving humans were approved by the University of Cape Town, South Africa. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JM: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. NM: Formal Analysis, Investigation, Supervision, Validation, Visualization, Writing – review and editing. EC: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the University of Cape Town–Africa Institute for Mathematical Sciences (UCT-AIMS) Scholarship, DAAD German Academic Exchange Service Fund No. A/91628092, the Integrative Biomedical Sciences Departmental Fund, the NRF/RCUK Newton Grant, the National Institutes of Health Common Fund under grant number U24HG006941 (H3ABioNet), and the Academy of Medical Sciences Professorship (Grant # APR9\1024), United Kingdom. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Acknowledgments

The authors acknowledge the High-Performance Computing resources at the University of Cape Town, the CHPC (https://www.chpc.ac.za/), and the National Integrated Cyber Infrastructure System (NICIS) for providing access to their Centre for High-Performance Computing resources that were used to run all the analyses in this research work. The authors are grateful to H3ABioNet consortium members for helpful discussions.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2025.1559496/full#supplementary-material

References

Author anonymous, (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576 (7785), 106–111. doi:10.1038/s41586-019-1793-z

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., et al. (2020). Tractor: a framework allowing for improved inclusion of admixed individuals in large-scale association studies. *BioRxiv*, 2020–05. doi:10.1101/2020.05.17.100727

Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., et al. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* 53 (2), 195–204. doi:10.1038/s41588-020-00766-y

Awany, D., Allali, I., Dalvie, S., Hemmings, S., Mwaikono, K. S., Thomford, N. E., et al. (2018). Host and microbiome genome-wide association studies: current state and challenges. *Front. Genet.* 9, 637. doi:10.3389/fgene.2018.00637

Brody, J. A., Morrison, A. C., Bis, J. C., O'Connell, J. R., Brown, M. R., Huffman, J. E., et al. (2017). Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat. Genet.* 49 (11), 1560–1563. doi:10.1038/ng.3968

Buchmann, R., and Hazelhurst, S. (2014). *Genesis manual*. Johannesburg: University of the Witwatersrand.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47 (Database issue), D1005-D1012–D1012. doi:10.1093/nar/gky1120

Campbell, M. C., and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genom. Hum. Genet.* 9, 403–433. doi:10.1146/annurev.genom.9.081307.164258

Chakraborty, R., and Weiss, K. M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci.* 85 (23), 9119–9123. doi:10.1073/pnas.85.23.9119

Chen, W., Brehm, J. M., Manichaikul, A., Cho, M. H., Boutaoui, N., Yan, Q., et al. (2015). A genome-wide association study of chronic obstructive pulmonary disease in Hispanics. *Ann. Am. Thorac. Soc.* 12 (3), 340–348. doi:10.1513/AnnalsATS.201408-380OC

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., et al. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* 40 (6), 1652–1666. doi:10.1093/ije/dyr120

Chimusa, E. R., Dalvie, S., Dandara, C., Wonkam, A., and Mazandu, G. K. (2018). Post genome-wide association analysis: dissecting computational pathway/network-based approaches. *Briefings Bioinforma.* 20, 690–700. doi:10.1093/bib/bby035

Chimusa, E. R., Mbiyavanga, M., Mazandu, G. K., and Mulder, N. J. (2016). ancGWAS: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations. *Bioinformatics* 32 (4), 549–556. doi:10.1093/bioinformatics/btv619

Chimusa, E. R., Zaitlen, N., Daya, M., Möller, M., van Helden, P. D., Mulder, N. J., et al. (2014). Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Hum. Mol. Genet.* 23 (3), 796–809. doi:10.1093/hmg/ddt462

Choi, S. W., Mak, T. S. H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* 15 (9), 2759–2772. doi:10.1038/s41596-020-0353-1

Choi, S. W., and O'Reilly, P. F. (2019). PRSice-2: polygenic Risk Score software for biobank-scale data. *Gigascience* 8 (7), giz082. doi:10.1093/gigascience/giz082

Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L., and Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *Am. J. Hum. Genet.* 101, 638–226. doi:10.1016/j.ajhg.2017.09.005

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2015). PRSice: polygenic risk score software. *Bioinformatics* 31 (9), 1466–1468. doi:10.1093/bioinformatics/btu848

Fatumo, S., and Inouye, M. (2023). African genomes hold the key to accurate genetic risk prediction. *Nat. Hum. Behav.* 7 (3), 295–296. doi:10.1038/s41562-023-01549-1

Geza, E., Mulder, N. J., Chimusa, E. R., and Mazandu, G. K. (2019). FRANC: a unified framework for multi-way local ancestry deconvolution with high density SNP data. *Briefings Bioinforma.* 21, 1837–1845. doi:10.1093/bib/bbz117

Goetz, L. H., Laura, H., Uribe-Bruce, L., Quarless, D., Libiger, O., and Nicholas, J. (2014). Admixture and clinical phenotypic variation. *Hum. Hered.* 77 (1-4), 73–86. doi:10.1159/000362233

Gonzalez, S., Gupta, J., Villa, E., Mallawaarachchi, I., Rodriguez, M., Ramirez, M., et al. (2016). Replication of genome-wide association study (GWAS) susceptibility loci in a Latino bipolar disorder cohort. *Bipolar Disord.* 18 (6), 520–527. doi:10.1111/bdi.12438

Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* 88 (5), 586–598. doi:10.1016/j.ajhg.2011.04.014

Hassan, R., Allali, I., Agamah, F. E., Elsheikh, S. S., Thomford, N. E., Dandara, C., et al. (2020). Drug response in association with pharmacogenomics and pharmacomicrobiomics: towards a better personalized medicine. *Briefings Bioinforma.* 22, bbaa292. doi:10.1093/bib/bbaa292

Honorato-Mauer, J., Shah, N. N., Maihofer, A. X., Zai, C. C., Belangero, S., Nievergelt, C. M., et al. (2025). Characterizing features affecting local ancestry inference performance in admixed populations. *Am. J. Hum. Genet.* 112, 224–234. doi:10.1016/j.ajhg.2024.12.005

INdian DIabetes COnsortium (2011). INDICO: the development of a resource for epigenomic study of Indians undergoing socioeconomic transition. *HUGO J.* 5 (1-4), 65–69. doi:10.1007/s11568-011-9157-2

Javed, A., Agrawal, S., and Ng, P. C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat. methods* 11 (9), 935–937. doi:10.1038/nmeth.3046

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press, 95–103.

Kalra, S., Kaur, R. P., Ludhiadch, A., Shafi, G., Vashista, R., Kumar, R., et al. (2018). Association of CYP2C19* 2 and ALDH1A1* 1/* 2 variants with disease outcome in breast cancer patients: results of a global screening array. *Eur. J. Clin. Pharmacol.* 74 (10), 1291–1298. doi:10.1007/s00228-018-2505-6

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42 (4), 348–354. doi:10.1038/ng.548

Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J., and Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. *Genome Biol.* 19, 1–14. doi:10.1186/s13059-018-1561-7

Kim, Y. J., Go, M. J., Hu, C., Hong, C. B., Kim, Y. K., Lee, J. Y., et al. (2011). Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat. Genet.* 43 (10), 990–995. doi:10.1038/ng.939

Klau, J. H., Maj, C., Klinkhammer, H., Krawitz, P. M., Mayr, A., Hillmer, A. M., et al. (2023). AI-based multi-PRS models outperform classical single-PRS models. *Front. Genet.* 14, 1217860. doi:10.3389/fgene.2023.1217860

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* 9 (1), 29–9. doi:10.1186/1746-4811-9-29

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* 44 (9), 1066–1071. doi:10.1038/ng.2376

Kosheleva, O., Kreinovich, V., Trung, N. D., and Autchariyapanitkul, K. (2021). *How to make a decision based on the minimum bayes factor (MBF): explanation of the Jeffreys scale*. Springer: Data Science for Financial Econometrics, 115–120.

Kraft, P., Zeggini, E., and Ioannidis, J. P. (2009). Replication in genome-wide association studies. *Stat. Sci. a Rev. J. Inst. Math. Statistics* 24 (4), 561–573. doi:10.1214/09-STS290

Mani, A. (2017). Local ancestry association, admixture mapping, and ongoing challenges. *Circ. Cardiovasc. Genet.* 10 (2), e001747. doi:10.1161/CIRCGENETICS.117.001747

Manolio, T. A. (2017). In Retrospect: a decade of shared genomic associations. *Nature* 546 (7658), 360–361. doi:10.1038/546360a

Mao, X., Bigham, A., Mei, R., Gutierrez, G., Weiss, K., Brutsaert, T., et al. (2007). A genomewide admixture mapping panel for hispanic/latino populations. *Am. J. Hum. Genet.* 80 (6), 1171–1178. doi:10.1086/518564

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93 (2), 278–288. doi:10.1016/j.ajhg.2013.06.020

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39 (7), 906–913. doi:10.1038/ng2088

Marigorta, U. M., Rodríguez, J. A., Gibson, G., and Navarro, A. (2018). Replicability and prediction: lessons and challenges from GWAS. *Trends Genet.* 34 (7), 504–517. doi:10.1016/j.tig.2018.03.005

Marnetto, D., Pärna, K., Läll, K., Molinaro, L., Montinaro, F., Haller, T., et al. (2020). Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.* 11 (1), 1628–1629. doi:10.1038/s41467-020-15464-w

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100 (4), 635–649. doi:10.1016/j.ajhg.2017.03.004

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

McGrath, J. J., Mortensen, P. B., Visscher, P. M., and Wray, N. R. (2013). Where GWAS and epidemiology meet: opportunities for the simultaneous study of genetic and environmental risk factors in schizophrenia. *Schizophr. Bull.* 39 (5), 955–959. doi:10.1093/schbul/sbt108

Mester, R., Hou, K., Ding, Y., Meeks, G., Burch, K. S., Bhattacharya, A., et al. (2023). Impact of cross-ancestry genetic architecture on GWASs in admixed populations. *Am. J. Hum. Genet.* 110 (6), 927–939. doi:10.1016/j.ajhg.2023.05.001

Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci. Rep.* 6 (1), 36671. doi:10.1038/srep36671

Mills, M. C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. *Commun. Biol.* 2 (1), 9–11. doi:10.1038/s42003-018-0261-x

Mugo, J. W., Chimusa, E. R., and Mulder, N. (2023). JasMAP: a joint ancestry and SNP association method for a multi-way admixed population. pp.2023–2110.

Mugo, J. W., Geza, E., Defo, J., Elsheikh, S. S., Mazandu, G. K., Mulder, N. J., et al. (2017). A multi-scenario genome-wide medical population genetics simulation framework. *Bioinformatics* 33 (19), 2995–3002. doi:10.1093/bioinformatics/btx369

Mulder, N., Abimiku, A. L., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., et al. (2018). H3Africa: current perspectives. *Pharmacogenomics personalized Med.* 11, 59–66. doi:10.2147/PGPM.S141546

Nakagome, S., Mano, S., Kozlowski, L., Bujnicki, J. M., Shibata, H., Fukumaki, Y., et al. (2012). Crohn's disease risk alleles on the NOD2 locus have been maintained by natural selection on standing variation. *Mol. Biol. Evol.* 29 (6), 1569–1585. doi:10.1093/molbev/mss006

National Academies of Sciences, Medicine, Policy, Global Affairs, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, Theoretical Statistics and Board on Mathematical Sciences (2019). *Reproducibility and replicability in science*. Washington (DC): National Academies Press.

Nelson, S. C., Romm, J. M., Doheny, K. F., Pugh, E. W., and Laurie, C. C. (2017). Imputation-based genomic coverage assessments of current genotyping arrays: illumina HumanCore, OmniExpress, multi-ethnic global array and sub-arrays, global screening array, Omni2. 5M, Omni5M, and affymetrix UK biobank. *BioRxiv*, 150219.

Olono, A., Mitesser, V., Happi, A., and Happi, C. (2024). Building genomic capacity for precision health in Africa. *Nat. Med.* 30 (7), 1856–1864. doi:10.1038/s41591-024-03081-9

Patron, J., Serra-Cayuela, A., Han, B., Li, C., and Wishart, D. S. (2019). Assessing the performance of genome-wide association studies for predicting disease risk. *PloS one* 14 (12), e0220215. doi:10.1371/journal.pone.0220215

Peng, B., and Kimmel, M. (2007). Simulations provide support for the common disease–common variant hypothesis. *Genetics* 175 (2), 763–776. doi:10.1534/genetics.106.058164

Peprah, E., Xu, H., Tekola-Ayele, F., and Royal, C. D. (2015). Genome-wide association studies in Africans and african Americans: expanding the framework of the genomics of human traits and disease. *Public health genomics* 18 (1), 40–51. doi:10.1159/000367962

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Saint Pierre, A., and Génin, E. (2014). How important are rare variants in common disease? *Briefings Funct. genomics* 13 (5), 353–361. doi:10.1093/bfgp/elu025

Secolin, R., Mas-Sandoval, A., Arauna, L. R., Torres, F. R., de Araujo, T. K., Santos, M. L., et al. (2019). Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci. Rep.* 9 (1), 13900. doi:10.1038/s41598-019-50362-2

Seldin, M., Pasaniuc, B., and Price, A. (2011). New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* 36, S21–S27.

Shriner, D. (2017). Overview of admixture mapping. *Curr. Protoc. Hum. Genet.* 94 (1), 1.23.1–1.23.8. doi:10.1002/cphg.44

Shriner, D., Adeyemo, A., Ramos, E., Chen, G., and Rotimi, C. N. (2011a). Mapping of disease-associated variants in admixed populations. *Genome Biol.* 12 (5), 223. doi:10.1186/gb-2011-12-5-223

Shriner, D., Adeyemo, A., and Rotimi, C. N. (2011b). Joint ancestry and association testing in admixed individuals. *PLoS Comput. Biol.* 7 (12), e1002325. doi:10.1371/journal.pcbi.1002325

Sirugo, S. M. W., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *CELL* 177 (4), 1080. doi:10.1016/j.cell.2019.04.032

Sun, Q., Horimoto, A. R., Chen, B., Ockerman, F., Mohlke, K. L., Blue, E., et al. (2025). Opportunities and challenges of local ancestry in genetic association analyses. *Am. J. Hum. Genet.* 112 (4), 727–740. doi:10.1016/j.ajhg.2025.03.004

Swart, Y., van Eeden, G., Uren, C., van der Spuy, G., Tromp, G., and Möller, M. (2022). GWAS in the southern African context. *PloS one* 17 (9), e0264657. doi:10.1371/journal.pone.0264657

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20 (8), 467–484. doi:10.1038/s41576-019-0127-1

Thornton, T. A., and Bermejo, J. L. (2014). Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genet. Epidemiol.* 38 (S1), S5-S12–S12. doi:10.1002/gepi.21819

Tishkoff, S., and Williams, S. (2002). Genetic analysis of african populations: human evolution and complex disease. *Nat.Rev Genet.* 3 (8), 611–621. doi:10.1038/nrg865

Turgut, A. O., and Koca, D. (2024). The effects of case/control ratio and sample size on genome-wide association studies: a simulation study. *Veterinary Med. Sci.* 10 (3), e1444. doi:10.1002/vms3.1444

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97 (4), 576–592. doi:10.1016/j.ajhg.2015.09.001

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of gwas discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101 (1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol. Official Publ. Int. Genet. Epidemiol. Soc.* 33 (1), 79–86. doi:10.1002/gepi.20359

Wang, Y. F., Zhang, H., and Ouyang, Q. (2007). Clinical manifestations of inflammatory bowel disease: east and West differences. *J. Dig. Dis.* 8 (3), 121–127. doi:10.1111/j.1443-9573.2007.00296.x

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids Res.* 42 (D1), D1001–D1006. doi:10.1093/nar/gkt1229

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26 (17), 2190–2191. doi:10.1093/bioinformatics/btq340

Winkler, C., Nelson, G., and Smith, M. (2010). Admixture mapping comes of age. *Annu.Rev.Genomics Hum. Genet.* 11, 65–89. doi:10.1146/annurev-genom-082509-141523

Wray, N. R., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., Murray, G. K., et al. (2021). From basic science to clinical application of polygenic risk scores: a primer. *JAMA psychiatry* 78 (1), 101–109. doi:10.1001/jamapsychiatry.2020.3049

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88 (1), 76–82. doi:10.1016/j.ajhg.2010.11.011

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44 (7), 821–824. doi:10.1038/ng.2310