frontiers in Genetics

BRIEF RESEARCH REPORT
published: 24 May 2021
doi: 10.3389/fgene.2021.673167

Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies

Meng Lin[1], Danny S. Park[2], Noah A. Zaitlen[3], Brenna M. Henn[4] and Christopher R. Gignoux[1]**

[1] Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [2] Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, United States, [3] Department of Neurology and Computational Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [4] Department of Anthropology, Center for Population Biology and the Genome Center, University of California, Davis, Davis, CA, United States

Genome-wide association studies (GWAS) are primarily conducted in single-ancestry settings. The low transferability of results has limited our understanding of human genetic architecture across a range of complex traits. In contrast to homogeneous populations, admixed populations provide an opportunity to capture genetic architecture contributed from multiple source populations and thus improve statistical power. Here, we provide a mechanistic simulation framework to investigate the statistical power and transferability of GWAS under directional polygenic selection or varying divergence. We focus on a two-way admixed population and show that GWAS in admixed populations can be enriched for power in discovery by up to 2-fold compared to the ancestral populations under similar sample size. Moreover, higher accuracy of cross-population polygenic score estimates is also observed if variants and weights are trained in the admixed group rather than in the ancestral groups. Common variant associations are also more likely to replicate if first discovered in the admixed group and then transferred to an ancestral population, than the other way around (across 50 iterations with 1,000 causal SNPs, training on 10,000 individuals, testing on 1,000 in each population, p = 3.78e-6, 6.19e-101, ∼0 for F_{ST} = 0.2, 0.5, 0.8, respectively). While some of these F_{ST} values may appear extreme, we demonstrate that they are found across the entire phenome in the GWAS catalog. This framework demonstrates that investigation of admixed populations harbors significant advantages over GWAS in single-ancestry cohorts for uncovering the genetic architecture of traits and will improve downstream applications such as personalized medicine across diverse populations.

Keywords: admixture, statistical power, complex trait genetics, polygenic score, genetic architecture

OPEN ACCESS

Edited by:
Fernanda De Souza Gomes Kehdy,
Oswaldo Cruz Foundation (Fiocruz), Brazil

Reviewed by:
Cesar Fortes-Lima,
Uppsala University, Sweden
Eaaswarkhanth Muthukrishnan,
Dasman Diabetes Institute, Kuwait

****Correspondence:***
Meng Lin
meng.lin@cuanschutz.edu
Christopher R. Gignoux
chris.gignoux@cuanschutz.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *26 February 2021*
Accepted: *27 April 2021*
Published: *24 May 2021*

Citation:
Lin M, Park DS, Zaitlen NA, Henn BM and Gignoux CR (2021) Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies. Front. Genet. 12:673167. doi: 10.3389/fgene.2021.673167

INTRODUCTION

Genome-wide association studies (GWAS) have allowed for significant progress in the field of human complex traits. However, groups with multiple ancestral origins have seldom been a primary focus in large scale genetic studies because: (1) admixed groups, along with other non-European populations, have largely been underrepresented in GWAS designs in the past

(Bustamante et al., 2011; Popejoy and Fullerton, 2016; Martin et al., 2017a), and (2) the population structure from heterogeneous ancestries in an admixed group, if not properly corrected, can result in spurious correlation signals and thus greater false positive rates (Rosenberg et al., 2010). However this mixture of ancestries present in admixed populations provides opportunities for novel discovery. Recent advancements in methodologies tailored for genetic mapping in admixed populations include disentangling of ancestry principal components and relatedness in the presence of admixture (Thornton et al., 2012; Conomos et al., 2015, 2016), combining local ancestry and allelic information to improve quantitative trait locus (QTL) mapping (Pasaniuc et al., 2011; Shriner et al., 2011; Atkinson et al., 2021), leveraging local ancestries for detection of epistasis (Aschard et al., 2015), and better fine mapping from linkage disequilibrium (LD) variability in diverse groups (Zaitlen et al., 2010; Asimit et al., 2016; Wojcik et al., 2019; Shi et al., 2020). Despite the fast development and practicality of these methods, they have not often been applied to sample sizes of hundreds of thousands to millions because study design and data collection in mega-scale cohorts routinely prioritize recruitment of participants of single ancestry (Atkinson et al., 2021). This greatly impedes downstream progress, such as polygenic risk score application across populations, where much lower accuracy is observed in non-European populations for many traits (Duncan et al., 2019; Martin et al., 2019; Cavazos and Witte, 2021).

In addition, complex traits in admixed groups potentially harbor differing genetic architectures and varying environmental exposures compared to most widely studied groups such as Europeans. Some biomedical traits have higher risk prevalence in admixed groups, such as prostate cancer in African Americans (Bhardwaj et al., 2017; Conti et al., 2021), asthma in Puerto Ricans (Lara et al., 2006; Pino-Yanes et al., 2015), obesity and type II diabetes in Native Hawaiians (Maskarinec et al., 2009), and active tuberculosis in a South African admixed population (Chimusa et al., 2014), which are likely attributed to elevated ancestry-specific risk allele frequency. Among anthropometric traits, skin pigmentation in groups with admixed ancestry harbor greater phenotypic variance than those with single ancestries (Martin et al., 2017b). Here, the larger phenotypic variance is likely caused by increased polygenicity in admixed groups, where in contrast some causal variants are nearly fixed in the single ancestry groups due to strong directional selection of skin pigmentation (e.g., rs1426654 in *SLC24A5*; Lin et al., 2018). The increase in minor allele frequencies in admixed populations compared to the populations of ancestral origin could be ubiquitous in traits that have been under differential processes of selection among ancestral populations or simply among populations that are deeply diverged. This would theoretically result in greater power of discovery in GWAS, as the analysis is most powerful for variants with higher minor allele frequency (MAF).

While genetic epidemiologists have typically focused on homogeneous populations, there are clear opportunities to improve discovery in admixed populations. For example, local ancestry can be leveraged to improve power in certain scenarios (e.g., Pasaniuc et al., 2011). In addition, Zhang and Stram (2014) observed a power gain in admixed individuals in dichotomous traits compared to pooled ancestral populations with stratification without environmental confounding. Here, we develop the genotype-phenotype simulation package *APRICOT*, Admixed Population poweR Inference Computed for phenOtypic Traits, a flexible mechanistic model, to address the question of power across a range of realistic scenarios of genotypic and ancestry-associated contributions. With *APRICOT* we compared an admixed population to a similar-sized ancestral population on its own across a range of allelic differentiation (as measured by F_{ST}; Weir and Cockerham, 1984) and varying narrow-sense heritability, allowing for a range of ancestry–phenotype associations, whether driven by genetics or environment. We further extend the insights gained by *APRICOT* to look at power for replication, whether from admixed populations to ancestral populations or *vice versa*, as well as opportunities to derive trans-ethnic polygenic scores.

METHODS

Simulation Framework

The framework of the simulator, now at https://github.com/menglin44/APRICOT, includes the main function of genotype-mediated simulation framework, and a side function to estimate simulation-based power estimate between a trait and global ancestry.

Simulation-Based Power Estimate Between a Trait and Global Ancestry

The first simulator we provide in this study builds phenotypes in admixed populations using only global ancestries, without involving genotype. The aim is to assess if the sample size is adequate for observing a dichotomous trait by ancestry correlation. The details are described in **Supplementary Notes**.

Genotype-Mediated Simulation Framework

The general simulation framework consists of two steps: first, we model ancestries and simulate genotypes based on ancestry specific frequencies and phenotypes (**Figure 1**); then, we test associations between the phenotype and causal variants via a linear model for a quantitative trait, or a logistic regression for a dichotomous trait, and summarize the statistical power. If the population is admixed, global ancestry is supplied as a fixed effect to correct for population structure.

Ancestry Modeling

Global ancestry in a 2-way admixed population is modeled as a beta distribution. The *i*th individual's ancestry θ is characterized as

$$\theta_i \sim Beta\left(m^2\left(\frac{1-m}{v}-\frac{1}{m}\right),\ \frac{1-m}{m}\alpha\right)$$

where *m* and *v* are the mean and variance of the global ancestry (from a presumptive population 1 in this framework) in an admixed population of interest.

The local ancestries, i.e., the source of ancestry of both the maternal and paternal copies of haplotypes at

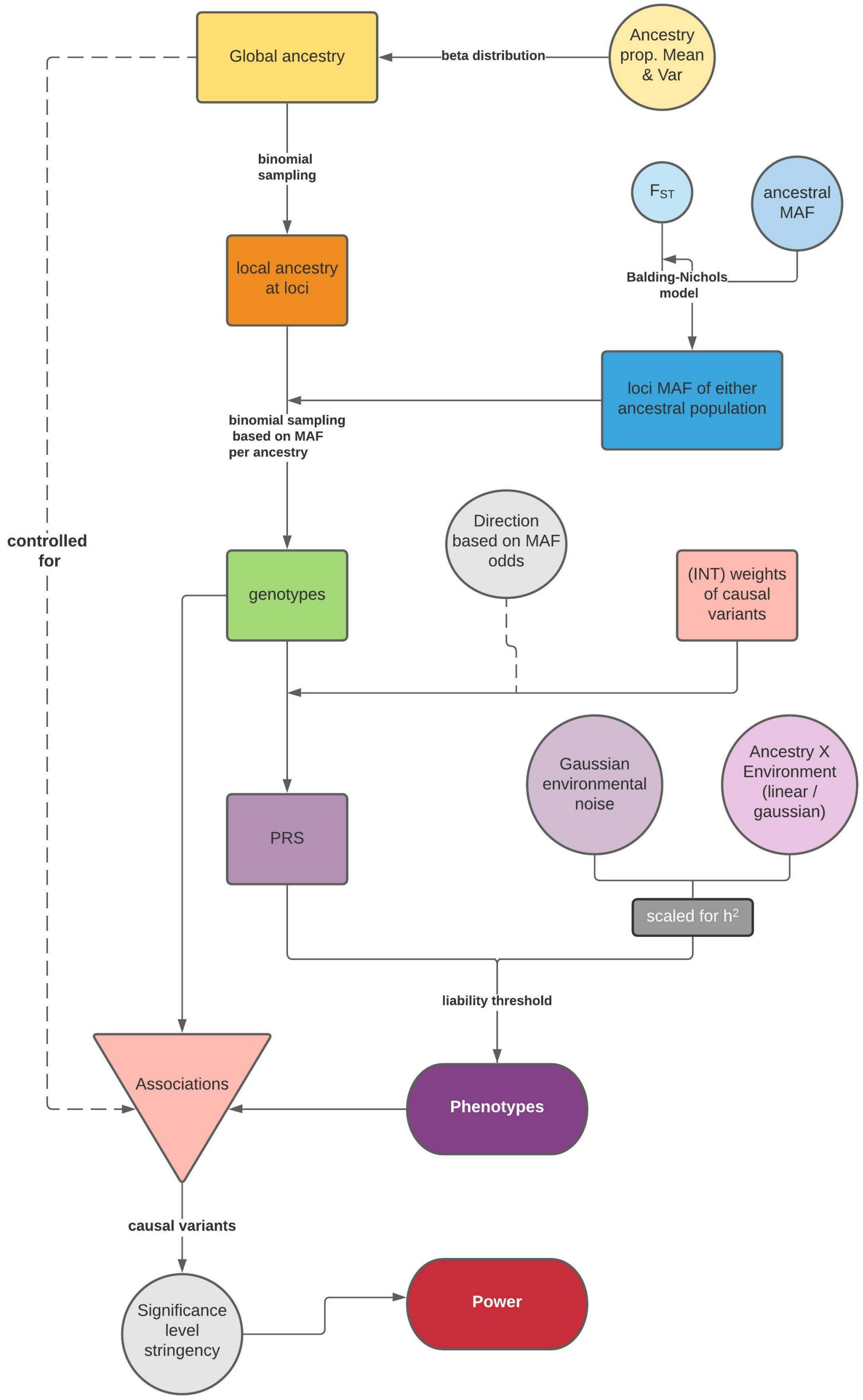
Global ancestry
beta distribution
Ancestry prop. Mean & Var
binomial sampling
F_{ST}
ancestral MAF
local ancestry at loci
Balding-Nichols model
loci MAF of either ancestral population
binomial sampling based on MAF per ancestry
controlled for
Direction based on MAF odds
(INT) weights of causal variants
genotypes
Gaussian environmental noise
Ancestry X Environment (linear / gaussian)
PRS
scaled for h^2
liability threshold
Associations
Phenotypes
causal variants
Significance level stringency
Power

FIGURE 1 | Flow chart of the simulation framework in *APRICOT*, including simulation of genotypes and phenotypes based on an admixture process and subsequent association tests for statistical power. A separate side function for testing association between a trait and global ancestry is illustrated in **Supplementary Figure 1**.

any genomic position, can be obtained from a binomial sampling with the probability equaling the global ancestry. The process is repeated independently for diploid chromosomes over a presumptive number (n) of LD-independent loci to form a n N local ancestry matrix, where N is the proposed sample size.

Genotype Simulation

We first draw allele frequencies in the two ancestries (Population 1 and 2) from a beta distribution under the Balding–Nichols model (Balding and Nichols, 1995) with a given F_{ST}

$$p_{1s}, p_{2s} \sim Beta\left(\frac{p_s(1-F_{ST})}{F_{ST}}, \frac{(1-p_s)(1-F_{ST})}{F_{ST}}\right)$$

where p_s is the allele frequency at an independent locus s in an ancestral population prior to the divergence and drawn from a uniform distribution *unif* (0.001, 0.999). Within the model we additionally provide additional distributions if the focus is not on common variants as it is here. We set F_{ST} as a flexible value to increase from a baseline genome-wide F_{ST} when the ancestral allele becomes rare. This is to reflect that a rarer variant in the ancestral population is easier to drift to different frequencies in diverged populations, especially if one population has undergone a severe bottleneck (as would be expected to increase F_{ST}).

$$F_{ST} = F_{ST_G} + (1 - MAF_a)\delta$$

where F_{ST_G} is the genome-wide background F_{ST} between the two populations of ancestral origin and is considered lower than the F_{ST} between trait-causal loci because of the difference in directional selection. MAF_a is the MAF of the variant in ancestral populations and δ is the increment with regard to MAF decrease, set as 0.3 in this study. Alternatively, we also test for fixed F_{ST} under the genome-wide background value when exploring the effect on power from various F_{ST} values ranging from 0.1 to 0.9. The genotypes are then drawn from binomial sampling using the allele frequency corresponding to the local ancestry assigned at the locus (i.e., p_{1s} or p_{2s}) across all loci.

Genetic Contribution to Trait

We randomly assign w out of the total n loci to be causal variants, where w is the proposed polygenicity of the trait. The weights for the causal variants are drawn from a standard normal distribution N (0, 1), and the signs of the weights are tied to the prevalence of the allele in the two populations of ancestral origin: the direction of the weight, positive or negative, at a locus is decided by the binary outcome of trial with probability $\frac{p_{1s}}{p_{2s}}$. In this way, a difference in directional selection of the complex trait in the two populations is introduced to facilitate a correlation between the trait and ancestry in the admixed group. Then, polygenic risk scores (PRS) in samples can be calculated based on the weights and the genotypes at causal loci.

Non-genetic Contribution to Trait

The non-genetic component is treated as the sum of two parts in admixed populations: (1) random environmental variation modeled as Gaussian noise and (2) environmental confounders correlated with ancestry, such as socioeconomic status and education, modeled as ancestry by environment interaction. Details are described in **Supplementary Notes**.

Phenotype

For quantitative traits, the phenotype is the direct sum of the genetic component (i.e., PRS) and the non-genetic score. For dichotomous traits, the phenotype is converted from the sum of genetic and non-genetic scores to binary case and control status based on the given liability threshold of the case prevalence.

Association Testing

Association between the trait and a variant is tested via a linear regression for a quantitative trait, or a logistic regression for a dichotomous trait in all three populations. The global ancestry is corrected in the admixed group. Power is defined as the proportions of causal variants with a significant p-value above a given stringency threshold.

Extended Analyses Based on Simulations

Estimation of False Positives

A false positive rate in association tests is verified against the association stringency by calculating the proportions of non-causal variants being discovered with significant association p-values. This is calculated separately in each population. The background F_{ST} (at non-causal variants) between the source populations was set to a constant 0.2, while a range of F_{ST} from 0.1 to 0.9 at trait causal loci that reflect trait divergence were tested. Other parameters were set as *standard* as described in the result section, with 100 causal variants, heritability 0.5, and environment by ancestry effect modeled as the sum of ancestral Gaussian environmental noise proportional to global ancestry. Each set of parameters was run with 50 repetitions.

PRS Estimation

For training purposes, we obtained the "estimated" weights of common (MAF = 5%), causal variants by conducting association analyses in 10,000 individuals in each simulated population (i.e., admixed population with ancestry proportions approximating those in African Americans, and two source populations: Pop 1 as the major ancestry source representing West Africans, Pop2 as the minor ancestry source representing Europeans). The simulations were run with 1,000 causal variants, F_{ST} at 0.2 and an increment associated with the rarity of the ancestral MAF, and other parameters the same as *standard* as above. We then used these weights to estimate PRS in another 1,000 individuals in each population as a test. We tested the PRS construction in two ways: firstly, we only used significant ($p < 0.05$) causal-variants from the training set (true positives); secondly, we included all significant variants (all positives) over a range of different stringency (p = 0.05, 5e-4, 5e-6, 5e-8, respectively). The true PRS of the individuals in the test set were available through an intermediate step in the simulations (**Figure 1**) and were used to test the accuracy of the estimated PRS via correlation coefficients.

Calculation of FST for Traits From the GWAS Catalog

We used the full NHGRI-EBI GWAS catalog "All associations v1.0" (Buniello et al., 2018) to extract variants that are significant genome-wide (<5e-8). We restricted traits to 899 that have more than 10 significantly associated variants that can be found in the 1000 Genome Project Phase 3

(1000 Genomes Project Consortium et al., 2015), and computed Weir and Cockerham's F_{ST} (Weir and Cockerham, 1984) between 99 Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) and 108 Yorubans in Ibadan, Nigeria (YRI) samples using PLINK v1.9[1] (Chang et al., 2015). The genomic background weighted F_{ST} was calculated on common variants (MAF > 5%) only.

RESULTS

Correlation Between a Trait and Ancestry Is Common

Complex trait studies in groups with heterogeneous ancestries usually require a correction for population structure. The implicit assumption is often a correlation between global ancestry and the trait that is commonly observed *a priori*. The estimated ancestries, or typically ancestry informative principal components, are included as a fixed effect to adjust for phenotypic variance from non-genetic confounders (e.g., social and cultural factors correlated with population structure), and to avoid spurious associations (Price et al., 2006). The correlations between ancestries and complex traits can also be due to changes in genetic architectures among ancestral groups either due to differential selection or deep divergence among populations. This in turn forms one of the basic motivations of multi-ancestry genetic studies, including admixture mapping (loci with ancestry deviating from genome-wide expectation), and cross-population transferability of genetic predictors. Therefore, we provide a power estimate for whether a significant correlation with ancestries can be observed, within a given incidence rate and ancestry distributions (section "Methods" and **Supplementary Figure 1**).

The Power of Genetic Discovery in an Admixed Population Is Higher Than in Ancestral Populations

We primarily focused on a genotype-mediated simulation framework to investigate the GWAS setting in an admixed group. We started by modeling global ancestries, then generating LD-independent genotypes based on population divergence, and subsequently the corresponding phenotypes under an additive model (section "Methods" and **Figure 1**). We set up the model in a 2-way admixed group with similar proportions to African Americans, here an average of ∼75% West African ancestry (denoted as Population 1 in simulations) and ∼25% European ancestry (denoted as Population 2) (Bryc et al., 2015; Baharian et al., 2016). We simulated a complex trait assuming 50% narrow sense heritability with 100 causal variants, either as a quantitative or a dichotomous trait with a liability threshold of 5%, in both the admixed population of interest and the homogenous populations of ancestral origin (N = 1,000 each). To induce a difference between ancestral phenotypic distributions and a correlation between a trait and global ancestries, we tied the direction of effect sizes to the minor allele frequencies in the two populations of ancestral origin (section "Methods" and **Figure 2**). In this study, each independent setting was repeated in 50 runs.

In the standard setting, we modeled the parameters described above, and F_{ST} across the 100 causal variants between Population 1 and 2 as a flexible value with a baseline equaling the genome-wide F_{ST} of 0.2 and an increment associated with the rarity of the ancestral MAF. This is referred to as Fst = 0.2+ in the text. The aim of this is to mirror the larger stochasticity in the frequency change of an ancestrally rare variant in diverged populations, especially when one of the derived populations has experienced the severe out-of-Africa bottleneck. We then tested associations between the phenotype and each locus while correcting for global ancestries over various sample sizes ranging from 1,000 to 1,000,000. We found the power to discover a causal variant at a canonical threshold of $p \leq 0.05$ significantly higher in an admixed population than the average in either of the populations of ancestral origin (Wilcoxon p = 1.23e-20 and 5.79e-10 across the range of sample sizes for quantitative and dichotomous traits described in **Figure 2**, respectively). In addition, we observed similarly high power in admixed populations with a slightly different ancestry composition to approximate the mixture of Indigenous American and European major ancestries in Chileans (Homburger et al., 2015), and an F_{ST} of 0.18 between the two source populations (Vidal et al., 2019; **Supplementary Figure 2**).

In addition to the standard setting where an environment by ancestry effect (Env × Anc) is modeled as ancestry-weighted Gaussian noise, we explored an alternative where we model Env × Anc as linearly dependent on the ancestry percentages, which would explain a range of proportions of phenotypic variance from 0% to $1 - h^2$ (**Supplementary Figure 3**). The power advantage in admixed populations remains consistent between the default Gaussian Env × Anc and linear modeling, where the latter was set as up to 10% of non-genetic components (**Supplementary Figure 4**).

The comparatively high power in admixed populations is more pronounced when the trait distributions have greater distance between Population 1 and 2, or the two populations are more deeply diverged, reflected by the larger F_{ST} at causal variants (**Figure 3**). To relate to real-world GWAS, we compared our levels of differentiation to the NHGRI-EBI GWAS catalog (Buniello et al., 2018). Among the 899 traits that have more than 10 genome-wide significant hits found in 1000 Genomes Project, the majority (N = 877) have at least one associated variant beyond the background F_{ST} of 0.155 (**Figure 3**), we provide a list of the most-differentiated traits between CEU and YRI in **Supplementary Table 1**. In contrast to the response to varying F_{ST}, the statistical power does not obviously change when the narrow sense heritability of the trait differs (**Supplementary Figure 5**). When increasing the overall stringency of the type I error rate up to a conventional genome-wide significance of 5e-8, the power advantage remains very similar across different thresholds, despite the expected decrease in power value on the absolute scale in all populations (**Supplementary Figures 6, 7**). Thus we picked the canonical threshold of $p \leq 0.05$ for the remaining analyses, as it can represent all stringency levels when this study focuses on the relative power comparison, and this

[1] www.cog-genomics.org/plink/1.9/

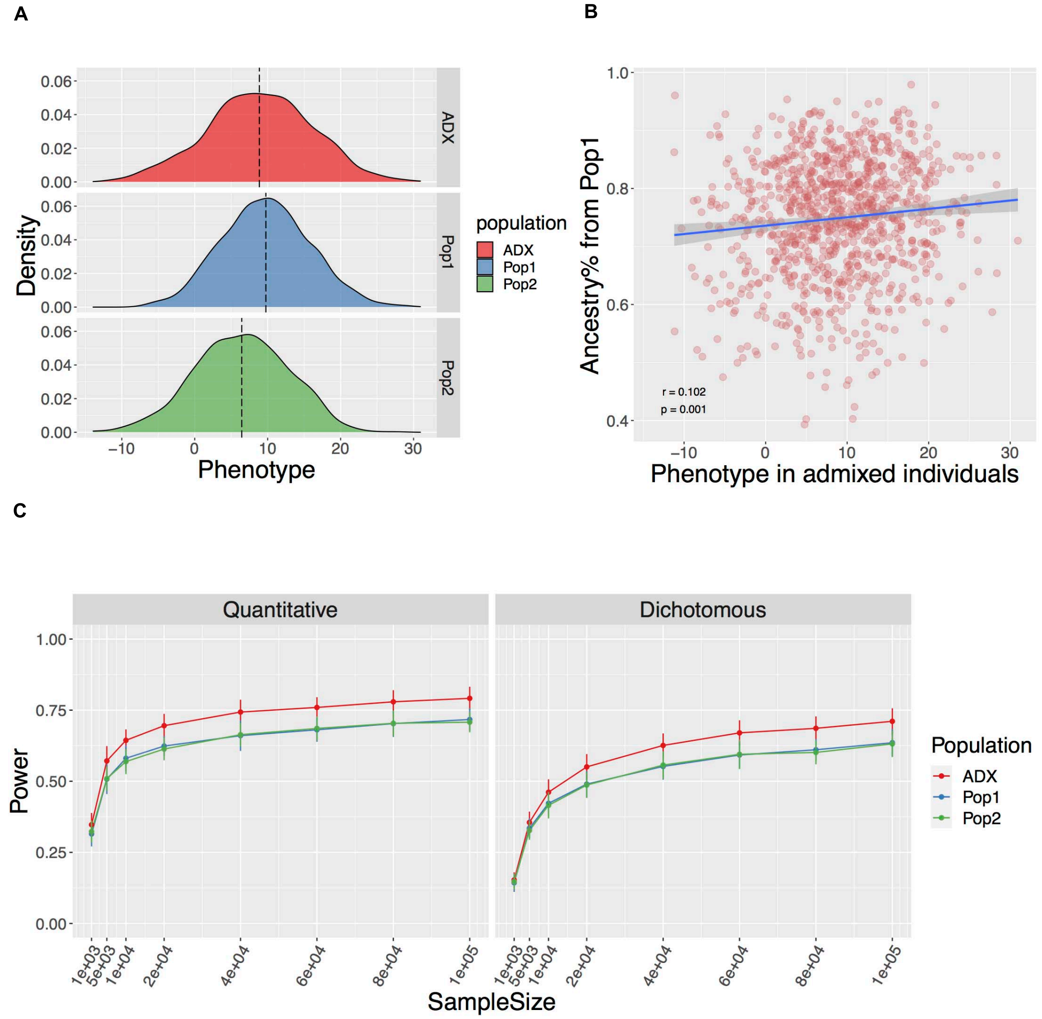

FIGURE 2 | Genotype mediated simulation under an example condition. The simulated trait has 100 causal variants with a narrow sense h^2 = 0.5, F_{ST} at causal variants 0.2+ (explained in Result), and environment by ancestry effect modeled as the sum of ancestral Gaussian environmental noise proportional to global ancestry. **(A)** Simulated quantitative phenotype distribution of populations of ancestral origin (Pop1, Pop2, blue, and green, respectively) and admixed population (ADX, red) of 1,000 samples each. **(B)** Correlation between simulated phenotype in admixed population and the global ancestry from Population 1. **(C)** Power to discover a causal variant over a range of sample sizes in a quantitative and dichotomous trait. Data point and error bars represent the mean and standard deviation across 50 repetitions, respectively.

more relaxed cutoff would include a larger number of causal variants for further discussion. The actual false positive rate of associations, as calculated from the 900 non-causal variants from the simulation when setting the background F_{ST} to 0.2, remained at approximately 5% in all three populations across the full F_{ST} and h^2 range for the causal variants of a trait (**Supplementary Figure 8**).

Cross-Population Replication and Transferability Is Asymmetric Between the Admixed Group and Homogenous Groups

As GWAS is conventionally focused on common variants, to investigate replication and transferability we then increased the polygenicity of a trait to 1,000 causal variants, and set MAF filtering at 5% for each population's genotypes prior to testing associations. We compare discovery in the major ancestral population (Population 1) relative to the admixed population. The proportion of significant signals that replicate in the reciprocal group is asymmetric between the two populations. Discovery in the admixed samples was more likely to replicate in Population 1 than the other way around, and this trend becomes more exaggerated as trait F_{ST} increases (one-way Wilcoxon p = 3.78e-6, < 2.2e-16, and < 2.2e-16 for F_{ST} = 0.2, 0.5, and 0.8, respectively; **Figure 4**).

We tested the cross-population transferability of polygenic scores (or polygenic risk scores, PRS) constructed from discovered loci with MAF > 5% in each population by increasing the sample sizes of the training set to 10,000

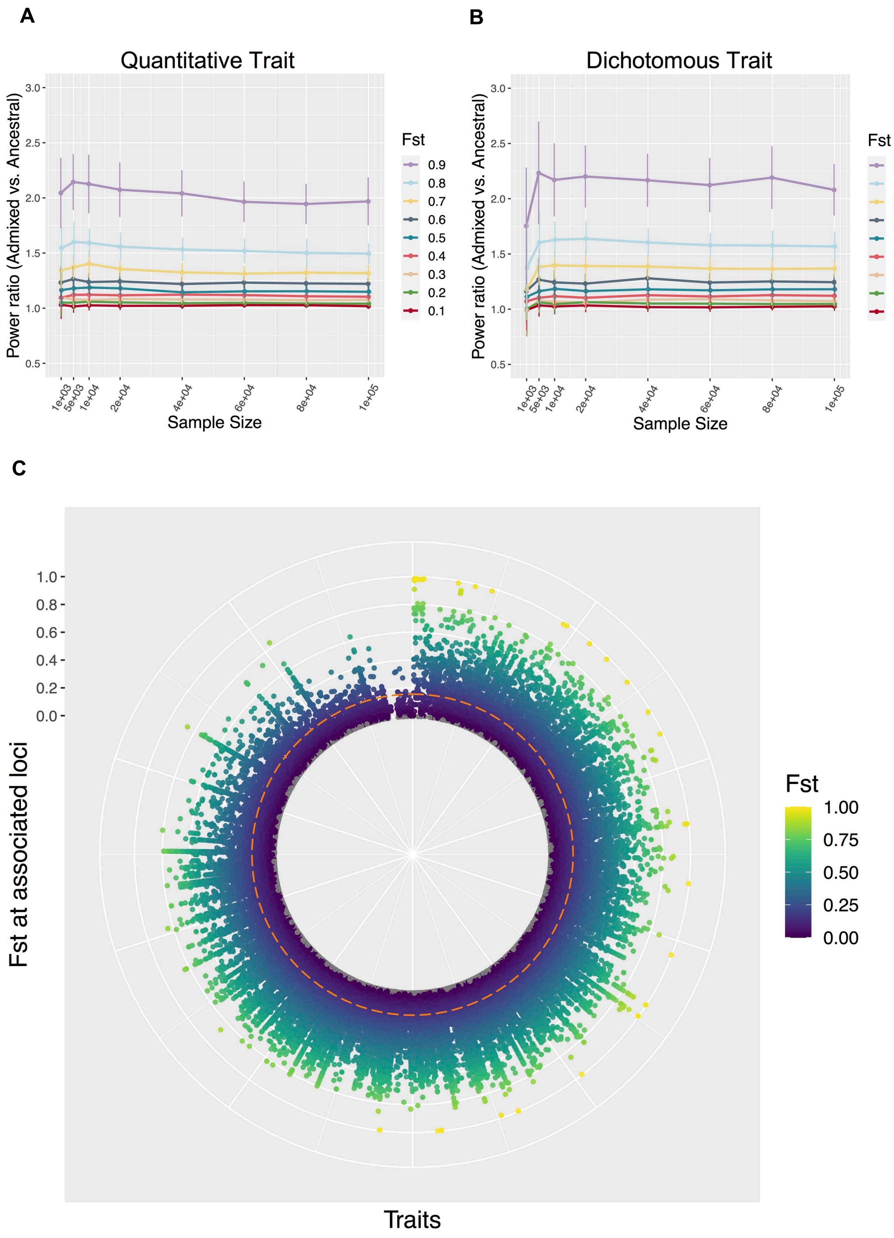

FIGURE 3 | Varying F_{ST} at trait associated loci. Ratio of power in admixed population over the average in the two populations of ancestral origin, with different F_{ST} at causal loci in **(A)** a quantitative trait and **(B)** a dichotomous trait. F_{ST} was set to constant during simulations per a specified value. The trait was assumed to have 100 causal loci and a narrow sense heritability of 0.5, with environment by ancestry effect modeled as a sum of ancestral Gaussian noise proportional to the global ancestry. Data points and error bars represent the mean and standard deviation across 50 repetitions, respectively. **(C)** F_{ST} at genome-wide significant hits for 899 traits from the GWAS catalog, between CEU and YRI from the 1000 Genomes Project Phase 3. Traits are spread along the radian (x-axis), with variant F_{ST} shown along the radius (y-axis). The dashed line represents the genomic background F_{ST}.

per population and separately estimating the GWAS-based PRS in an additional testing set of 1,000 samples in each population. We measured the PRS accuracy as the correlation coefficient *r* between the estimated values and the true value in each test group across 50 repeats of simulations. Interestingly, the prediction accuracy is also asymmetric between admixed and homogenous samples. When we constructed PRS using only true positive signals at an alpha of 0.05, the accuracy of estimating PRS in Population 1 or 2 using weights and loci trained from the admixed population is significantly higher than the other way around. This holds true when using all (both true and false) positive signals at various stringency levels (**Figure 4**, **Supplementary Figure 9**, and **Supplementary Table 2**), suggesting another advantage in conducting GWAS in admixed populations.

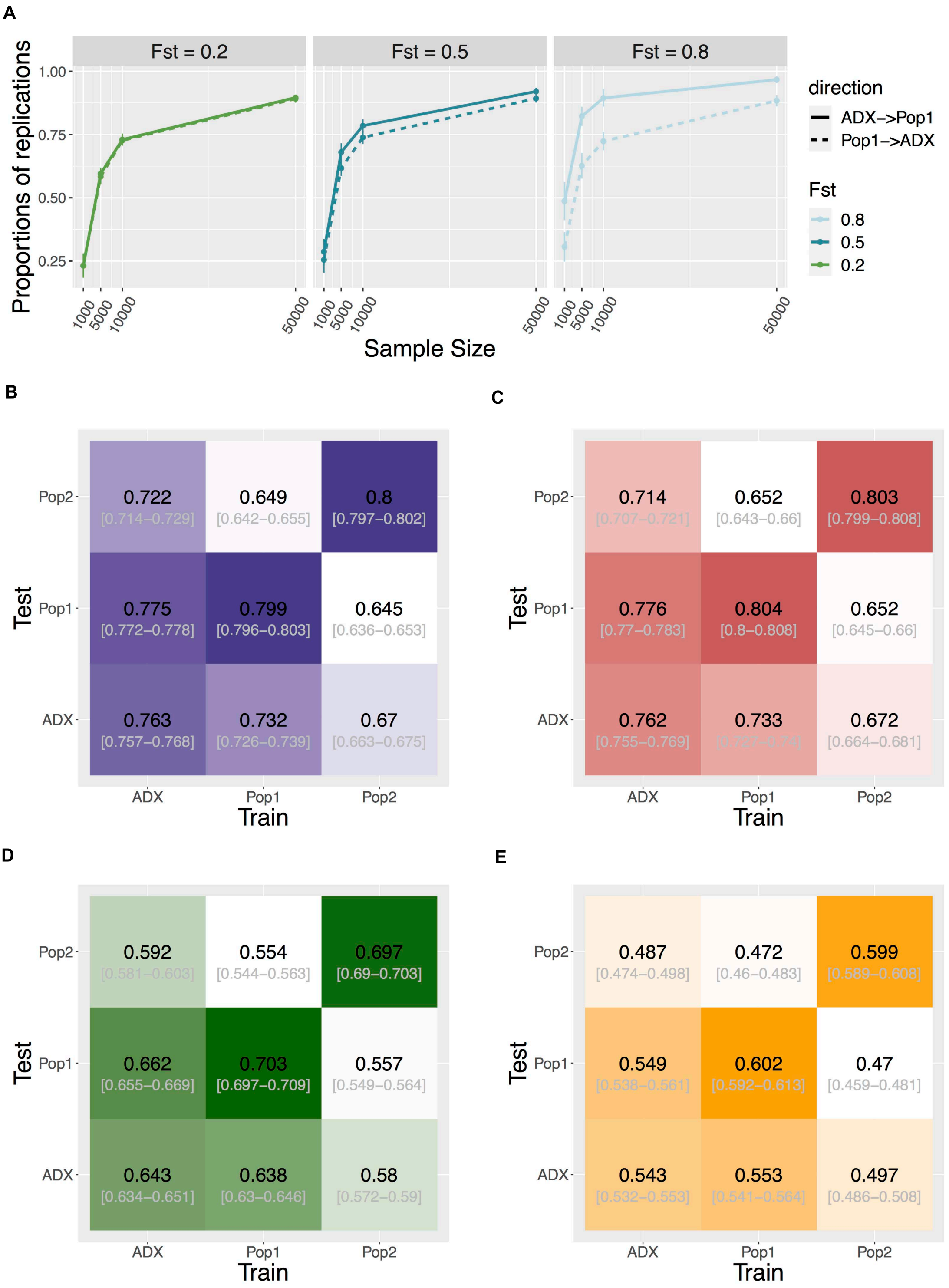

FIGURE 4 | Transferability of GWAS variants across populations. **(A)** Replication of individual signals that are common in both ancestral Population 1 and the admixed group. Direction of replication is shown as a solid or dashed line: the former indicates loci are discovered in an admixed population and replicated in Population 1; the latter loci are discovered in Population 1 and replicated in the admixed population. Data point and error bars represent the mean and standard deviation across 50 repetitions. **(B–E)** Heat map of accuracy of PRS using signals above different stringency of significance level at 0.05, 5e-4, 5e-6, and 5e-8, respectively. The accuracy is measured as the correlation coefficient between the estimated PRS against the true PRS. The training population where the weights and variants were identified, and the test population in which to construct PRS, are specified on the x- and y-axis. Central numbers in black within each cell are the average correlation coefficient across 50 independent simulations, with the 95% confidence interval of the mean acquired from bootstrapping (*n* = 1,000).

DISCUSSION

Our simulation framework, *APRICOT*, demonstrated that GWAS in admixed populations has greater power for discovery than in the homogenous populations of ancestral origin, given the same sample sizes. The difference in power increases when the trait is under more differentiated polygenic selection in the two populations of ancestral origin, reflected by F_{ST}. This is because when a trait is driven by more-differentiated variants, its causal variants are likely to be pushed to more extreme allele frequencies, thus weakening the statistical power of discovery in that population. In contrast, the frequency of the same causal loci in admixed populations likely have become more intermediate due to variation in ancestries, making them much easier to detect. An extreme yet classic example that echoes with the observation would be skin pigmentation, where selection is in the opposite direction in populations at high latitude and those living near the equator. A non-synonymous, skin-lightening mutation at rs1426654 is fixed in European descendants, with a high F_{ST} of 0.985 between CEU and YRI. This mutation would never have been discovered through GWAS if analyses were only conducted in European populations, but it is highly detectable through association analyses in admixed populations (Martin et al., 2017b).

Additionally, the power advantage in admixed populations may persist even for traits that have not been under such strong differentiation: for almost all the 899 traits we examined from the GWAS catalog, some associated SNPs can have a much larger than background F_{ST} between CEU and YRI, even when the traits themselves on average show limited differentiation (**Figure 3**). We note, however, that the high F_{ST} across these trait-associated variants could partially be attributed to ascertainment bias, where the "tagging SNPs" by design are common in Europeans, making the corresponding genetic component of these traits seemingly more differentiated across populations (Novembre and Barton, 2018). The true causal variants that were tagged by these signals could have moderately attenuated F_{ST}, yet the differences in allele frequency likely remain larger than expected, as previously observed from GWAS on simulated whole genome sequences between Africans and Europeans (Kim et al., 2018). Therefore, attempts to discover variants similar to these "F_{ST} outlier" signals would benefit from GWAS designed in admixed samples.

In this study, we provide a mechanistic framework to explore the relationship between power gain in single variant associations and variation in ancestries, mediated by the nature of intermediate allele frequencies in admixed populations. A similar hypothesis of power increase was also explored via simulations in Zhang and Stram (2014), though the focus of their study was to explore the role of local ancestry in genetic associations; therefore, the assumptions of architecture for comparison between admixed and ancestral populations were simplified, where non-genetic components (such as environmental effect and environment by ancestry interactions) and heritability were not considered in the model, and a constant effect size was assumed for all causal variants. Under this model, Zhang and Stram (2014) observed a power increase in admixed groups when compared to stratified analyses in the ancestral populations pooled with a proportion identical to the mean global ancestry percentage. Our simulations extended this framework to dive deeper into more-realistic scenarios across various ranges of environmental effect, trait divergence, and heritability. Moreover, we modeled the ancestry-phenotype association observed in many real-world traits under various different distributions. With a more adjustable genetic architecture in the model, we were able to quantify power advantage in admixed populations within different circumstances in order to investigate practical applications as replication portability and PRS.

In realistic practice, some confounders and restrictions beyond the model assumptions exist: first, some non-additive genetic components, such as genetic by environment interactions (G × E) and epistasis (Park et al., 2018; Rau et al., 2020), could potentially induce effect size heterogeneity at causal loci with or among populations (Rosenberg et al., 2019), thus obscuring the prediction of power advantage in admixed populations because the power of discovery would be variant-specific and balanced by the gain *vs.* loss from the increase in frequency and change in effect size. However, the increase in power is still expected to be substantial from additive components that are usually considered major in a genetic architecture, with effect sizes highly similar across populations (Wojcik et al., 2019). Additionally, currently the contribution from epistasis or G × E components to most trait variability is estimated to be relatively small (Wang et al., 2019; Dahl et al., 2020; Hivert et al., 2021). For variants with heterogeneous effect sizes per ancestry, other local ancestry-aware regression methods could potentially improve the power of detection in admixed populations (Atkinson et al., 2021). Second, the observations in this study that admixed populations harbor a greater power of discovery in GWAS than the ancestral populations is credited to the existence of ancestry variance, independent of specific demographic history of either the admixed or the ancestral population. It is possible that the demographic details or specific assumptions of the genetic architecture would affect the absolute value of power estimate on a finer scale, which has not been the focus of this study, yet is worth being further explored through forward or coalescent simulations with additional details (including modeling differential linkage disequilibrium patterns) in the future. Third, we focused on a single admixture scenario, albeit one reflecting a realistic scenario. We would anticipate our observed patterns to be exaggerated in populations with even contributions from Populations 1 and 2. Further our framework could be extended to k-way admixed populations, albeit with possibly elevated computation burden from the step of sampling population specific allele frequency that needs to meet all pairwise F_{ST} relationships. Additionally, the interpretation and degree of population-specific interpretation become more complex to be described here. Lastly, while we did not model selection that happens after admixture, it may be possible to investigate changes in post-admixture dynamics via bias in local ancestry, and the development of that approach would be a relevant future direction of *APRICOT*.

Despite the underrepresentation of admixed groups in large GWAS, recent research has highlighted the importance of conducting genetic research with more diversity. Our work joins burgeoning efforts to quantify the statistical benefits of complex trait studies in diverse populations, especially populations of mixed ancestry. Our work suggests another advantage for conducting genetic studies in admixed populations, which comes from elevated allele frequencies when traits are moderately to highly differentiated. Moreover, discoveries from such studies aid improvement in cross-population PRS, which is critical in clinical prediction in personalized medicine yet presently has suboptimal performance for many biomedical traits in non-European populations (Martin et al., 2019; Rosenberg et al., 2019; Cavazos and Witte, 2021). We therefore highlight that insights gained from admixed populations provide improved and appealing generalizable properties compared to homogeneous populations. As the field increasingly moves toward personalized medicine applications we must be mindful of opportunities to incentivize novel studies and analyses in diverse and, particularly as we highlight here, populations of mixed ancestry.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: 1000 Genomes Project: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/; Software availability: https://github.com/menglin44/APRICOT.

AUTHOR CONTRIBUTIONS

ML conducted analyses with supervision from BH and CG. ML, BH, and CG wrote the manuscript with algorithmic insights from DP and NZ. All authors edited and approved the submitted version.

FUNDING

This research was supported by NIH Grants R35GM133531 (to BH), R01HG010297, U01HG009080/S1 (to NZ and CG), and R01HG011345 (to NZ and CG).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.673167/full#supplementary-material

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526:68. doi: 10.1038/nature15393

Aschard, H., Gusev, A., Brown, R., and Pasaniuc, B. (2015). Leveraging local ancestry to detect gene-gene interactions in genome-wide data. *BMC Genet.* 16:124. doi: 10.1186/s12863-015-0283-z

Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P., and Zeggini, E. (2016). Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* 24, 1330–1336. doi: 10.1038/ejhg.2016.1

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, 195–204. doi: 10.1038/s41588-020-00766-y

Baharian, S., Barakatt, M., Gignoux, C. R., Shringarpure, S., Errington, J., Blot, W. J., et al. (2016). The Great migration and african-american genomic diversity. *PLoS Genet.* 12:e1006059. doi: 10.1371/journal.pgen.1006059

Balding, D. J., and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96, 3–12. doi: 10.1007/bf01441146

Bhardwaj, A., Srivastava, S. K., Khan, M. A., Prajapati, V. K., Singh, S., Carter, J. E., et al. (2017). Racial disparities in prostate cancer a molecular perspective. *Front. Biosci.* 22:772–782. doi: 10.2741/4515

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., and Mountain, J. L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* 96, 37–53. doi: 10.1016/j.ajhg.2014.11.010

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Bustamante, C. D., Vega, F. M. D. L., and Burchard, E. G. (2011). Genomics for the world. *Nature* 475:163. doi: 10.1038/475163a

Cavazos, T. B., and Witte, J. S. (2021). Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Hum. Genet. Genom. Adv.* 2:100017. doi: 10.1016/j.xhgg.2020.100017

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 1–16. doi: 10.1186/s13742-015-0047-8

Chimusa, E. R., Zaitlen, N., Daya, M., Möller, M., Helden, P. D., van 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, 796–809. doi: 10.1093/hmg/ddt462

Conomos, M. P., Miller, M. B., and Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* 39, 276–293. doi: 10.1002/gepi.21896

Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* 98, 127–148. doi: 10.1016/j.ajhg.2015.11.022

Conti, D. V., Darst, B. F., Moss, L. C., Saunders, E. J., Sheng, X., Chou, A., et al. (2021). Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat. Genet.* 53, 65–75. doi: 10.1038/s41588-020-00748-0

Dahl, A., Nguyen, K., Cai, N., Gandal, M. J., Flint, J., and Zaitlen, N. (2020). A robust method uncovers significant context-specific heritability in diverse complex traits. *Am. J. Hum. Genet.* 106, 71–91. doi: 10.1016/j.ajhg.2019.11.015

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., et al. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* 10, 3328. doi: 10.1038/s41467-019-11112-0

Hivert, V., Sidorenko, J., Rohart, F., Goddard, M. E., Yang, J., Wray, N. R., et al. (2021). Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *Am. J. Hum. Genetics*. 108, 786–798. doi: 10.1016/j.ajhg.2021.02.014

Homburger, J. R., Moreno-Estrada, A., Gignoux, C. R., Nelson, D., Sanchez, E., Ortiz-Tello, P., et al. (2015). Genomic insights into the ancestry and demographic history of south America. *PLoS Genet.* 11:e1005602. doi: 10.1371/journal.pgen.1005602

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.* 19:179. doi: 10.1186/s13059-018-1561-7

Lara, M., Akinbami, L., Flores, G., and Morgenstern, H. (2006). Heterogeneity of childhood asthma among hispanic children: puerto rican children bear a disproportionate burden. *Pediatrics* 117, 43–53. doi: 10.1542/peds.2004-1714

Lin, M., Siford, R. L., Martin, A. R., Nakagome, S., Möller, M., Hoal, E. G., et al. (2018). Rapid evolution of a skin-lightening allele in southern African KhoeSan. *Proc. Natl. Acad. Sci.* 115:201801948. doi: 10.1073/pnas.1801948115

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017a). Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100, 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, 584–591. doi: 10.1038/s41588-019-0379-x

Martin, A. R., Lin, M., Granka, J. M., Myrick, J. W., Liu, X., Sockell, A., et al. (2017b). An Unexpectedly complex architecture for skin pigmentation in Africans. *Cell* 171, 1340–1353e14. doi: 10.1016/j.cell.2017.11.015

Maskarinec, G., Grandinetti, A., Matsuura, G., Sharma, S., Mau, M., Henderson, B. E., et al. (2009). Diabetes prevalence and body mass index differ by ethnicity: the Multiethnic Cohort. *Ethnic Dis.* 19, 49–55.

Novembre, J., and Barton, N. H. (2018). Tread lightly interpreting polygenic tests of selection. *Genetics* 208, 1351–1355. doi: 10.1534/genetics.118.300786

Park, D. S., Eskin, I., Kang, E. Y., Gamazon, E. R., Eng, C., Gignoux, C. R., et al. (2018). An ancestry−based approach for detecting interactions. *Genet. Epidemiol.* 42, 49–63. doi: 10.1002/gepi.22087

Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. H. L., et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using african americans from care and a breast cancer consortium. *PLoS Genet.* 7:e1001371. doi: 10.1371/journal.pgen.1001371

Pino-Yanes, M., Thakur, N., Gignoux, C. R., Galanter, J. M., Roth, L. A., Eng, C., et al. (2015). Genetic ancestry influences asthma susceptibility and lung function among Latinos. *J. Aller. Clin. Immun.* 135, 228–235. doi: 10.1016/j.jaci.2014.07.053

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

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Rau, C. D., Gonzales, N. M., Bloom, J. S., Park, D., Ayroles, J., Palmer, A. A., et al. (2020). Modeling epistasis in mice and yeast using the proportion of two or more distinct genetic backgrounds: evidence for "polygenic epistasis.". *PLoS Genet.* 16:e1009165. doi: 10.1371/journal.pgen.1009165

Rosenberg, N. A., Edge, M. D., Pritchard, J. K., and Feldman, M. W. (2019). Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evol. Med. Publ. Heal* 2019, 26–34. doi: 10.1093/emph/eoy036

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* 11, 356–366. doi: 10.1038/nrg2760

Shi, H., Burch, K. S., Johnson, R., Freund, M. K., Kichaev, G., Mancuso, N., et al. (2020). Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *Am. J. Hum. Genet.* 106, 805–817. doi: 10.1016/j.ajhg.2020.04.012

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

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., and Risch, N. (2012). Estimating kinship in admixed populations. *Am. J. Hum. Genet.* 91, 122–138. doi: 10.1016/j.ajhg.2012.05.024

Vidal, E. A., Moyano, T. C., Bustos, B. I., Pérez-Palma, E., Moraga, C., Riveras, E., et al. (2019). Whole genome sequence, variant discovery and annotation in mapuche-huilliche native south Americans. *Sci. Rep.*U.K. 9:2132. doi: 10.1038/s41598-019-39391-z

Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci. Adv.* 5:eaaw3538. doi: 10.1126/sciadv.aaw3538

Weir, B. S., and Cockerham, C. C. (1984). Estimating F−statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570, 514–518. doi: 10.1038/s41586-019-1310-4

Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* 86, 23–33. doi: 10.1016/j.ajhg.2009.11.016

Zhang, J., and Stram, D. O. (2014). The role of local ancestry adjustment in association studies using admixed populations. *Genet. Epidemiol.* 38, 502–515. doi: 10.1002/gepi.21835

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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.