Interpreting SNP heritability in admixed populations

Jinguo Huang[1,2], Saonli Basu[3], Mark D. Shriver[2], and Arslan A. Zaidi[†4,5]

[1]Bioinformatics and Genomics, Huck Institutes of the Life Sciences, Pennsylvania State University

[2]Department of Anthropology, Pennsylvania State University

[3]Department of Biostatistics, University of Minnesota

[4]Department of Genetics, Cell Biology, and Development, University of Minnesota

[5]Institute of Health Informatics, University of Minnesota

[†]Correspondence to A.A.Z (aazaidi@umn.edu)

November 7, 2023

Abstract

SNP heritability (h^2_{snp}) is defined as the proportion of phenotypic variance explained by genotyped SNPs and is believed to be a lower bound of heritability (h^2), being equal to it if all causal variants are known. Despite the simple intuition behind h^2_{snp}, its interpretation and equivalence to h^2 is unclear, particularly in the presence of population structure and assortative mating. It is well known that population structure can lead to inflation in $\hat{h}^2_{snp}$ estimates. Here we use analytical theory and simulations to demonstrate that h^2_{snp} estimated with genome-wide restricted maximum likelihood (GREML) can be biased in admixed populations, even in the absence of confounding and even if all causal variants are known. This is because admixture generates linkage disequilibrium (LD), which contributes to the genetic variance, and therefore to heritability. GREML implicitly assumes this component is zero, which may not be true, particularly for traits under divergent or stabilizing selection in the source populations, leading under- or over-estimates of h^2_{snp} relative to h^2. For the same reason, GREML estimates of local ancestry heritability (h^2_γ) will also be biased. We describe the bias in $\hat{h}^2_{snp}$ and $\hat{h}^2_\gamma$ as a function of admixture history and the genetic architecture of the trait and discuss its implications for genome-wide association and polygenic prediction.

Introduction

The ability to estimate (narrow-sense) heritability (h^2) from unrelated individuals was a major advance in genetics. Traditionally, h^2 was estimated from family-based studies in which the phenotypic resemblance between relatives could be modeled as a function of their expected genetic relatedness [1]. But this approach was limited to analysis of closely related individuals where pedigree information is available and the realized genetic relatedness is not too different from expectation [2]. With the advent of genome-wide association studies (GWAS), we hoped that many of the variants underlying this heritability would be uncovered. But when genome-wide significant SNPs explained a much smaller fraction of the phenotypic variance, it became important to explain the missing heritability – were family-based estimates inflated or were GWAS just underpowered, limited by variant discovery?

Yang *et al.* (2010) [3] made the key insight that one could estimate the portion of h^2 tagged by genotyped SNPs, regardless of whether or not they were genome-wide significant, by exploiting the subtle variation in the realized genetic relatedness among apparently unrelated individuals [3–5]. This quantity came to be known colloquially as 'SNP heritability' (h^2_{snp}) and it is believed to be equal to h^2 if all causal variants are included among genotyped SNPs [3]. Indeed, estimates of h^2_{snp} explain a much larger fraction of trait heritability than GWAS SNPs [3], approaching family-based estimates of h^2 when whole genome sequence data are used [6]. This has made it clear that GWAS have yet to uncover more variants with increasing sample size. Now, h^2_{snp} has become an important aspect of the design of genetic studies and is often used to define the power of variant discovery in GWAS and the upper limit of polygenic prediction accuracy.

Despite the utility and simple intuition of h^2_{snp}, there is much confusion about its interpretation and equivalence to h^2, particularly in the presence of population structure and assortative mating [7–12]. But much of the discussion of heritability in structured populations has focused on biases in $\hat{h}^2_{snp}$ – the estimator – due to confounding effects of shared environment and linkage disequilibrium (LD) with other variants [7, 9–11, 13]. There is comparatively little discussion, at least in human genetics, on the fact that LD due to population structure also contributes to genetic variance, and therefore, is a component of heritability [1] (but see [14–16] for a rigorous discussion). We think this is at least partly due to the fact that most studies are carried out in cohorts with primarily European ancestry, where the degree of population structure is minimal and large effects of LD can be ignored. However, that is not the case for diverse, multi-ethnic cohorts, which have historically been underrepresented in genetic studies, but thanks to a concerted effort in the field, are now becoming increasingly common [17–23]. The complex structure in these cohorts also brings unique methodological challenges and it is imperative that we understand whether existing methods, which have largely been evaluated in more homogeneous groups, generalize to more diverse cohorts.

Our goal in this paper is to study the behavior of $\hat{h}^2_{snp}$ in admixed populations. What is its interpretation in the ideal situation where causal variants are known? Is it an unbiased estimate of h^2? To answer these questions, we derived a general expression for the genetic variance in admixed populations, decomposing it in terms of the contribution of population structure, which influences both the genotypic variance at individual loci and the LD across loci. We used extensive simulations, where the ground truth

is known, to show that $\hat{h}^2_{snp}$ estimated with genome-wide restricted maximum likelihood (GREML) [3, 5] can be biased in admixed and other structured populations, even in the absence of confounding and when all causal variants are known. We explain this in terms of the discrepancy between the model assumed by GREML and the generative model from which the genetic architecture of the trait in the population may have been sampled. We describe the bias in $\hat{h}^2_{snp}$ as a function of admixture history and genetic architecture and discuss its implications for GWAS and polygenic prediction accuracy.

Model

Genetic architecture

We begin by describing a generative model for the phenotype. Let $y = g + e$, where y is the phenotypic value of an individual, g is the genotypic value, and e is random error. We assume additive effects such that $g = \sum_{i=1}^{m} \beta_i x_i$ where β_i is the effect size of the ith biallelic locus and $x_i \in \{0, 1, 2\}$ is the number of copies of the trait-increasing allele. Importantly, the effect sizes are fixed quantities and differences in genetic values among individuals are due to random variation in genotypes. Note, that this is different from the model assumed by GREML where genotypes are fixed and effect sizes are random [14].

We denote the mean, variance, and covariance with $\mathbb{E}(.)$, $\mathbb{V}(.)$, and $\mathbb{C}(.)$, respectively, where the expectation is measured over random draws from the population rather than random realizations of the evolutionary process. We can express the additive genetic variance of a quantitative trait as follows:

$$V_g = \mathbb{V}(\sum_{i=1}^{m} \beta_i x_i) = \sum_{i=1}^{m} \beta_i^2 \, \mathbb{V}(x_i) + \sum_{j \neq i} \beta_i \beta_j \, \mathbb{C}(x_i, x_j)$$

Here the first term represents the contribution of individual loci (genic variance) and the second term is the contribution of linkage disequilibrium (LD contribution). We make the assumption that loci are unlinked and therefore, the LD contribution is entirely due to population structure. We describe the behavior of V_g in a population that is a mixture of two previously isolated populations A and B that diverged from a common ancestor. To do this, we denote θ as the fraction of the genome of an individual with ancestry from population A. Thus, $\theta = 1$ if the individual is from population A, 0 if they are from population B, and $\theta \in (0, 1)$ if they are admixed. Then, V_g can be expressed in terms of ancestry as (Appendix):

$$V_g = 2\,\mathbb{E}(\theta) \sum_{i=1}^{m} \beta_i^2 \, f_{\mathrm{i}}^{\mathrm{A}}(1 - f_{\mathrm{i}}^{\mathrm{A}}) + 2\{1 - \mathbb{E}(\theta)\} \sum_{i=1}^{m} \beta_i^2 \, f_{\mathrm{i}}^{\mathrm{B}}(1 - f_{\mathrm{i}}^{\mathrm{B}}) \tag{1.1}$$

$$+2\,\mathbb{E}(\theta)\{1 - \mathbb{E}(\theta)\} \sum_{i=1}^{m} \beta_i^2 (f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2 \tag{1.2}$$

$$+2\,\mathbb{V}(\theta) \sum_{i=1}^{m} \beta_i^2 (f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2 \tag{1.3}$$

$$+4\,\mathbb{V}(\theta) \sum_{i \neq j} \beta_i \beta_j (f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}}) \tag{1.4}$$

where $f_{\mathrm{i}}^{\mathrm{A}}$ and $f_{\mathrm{i}}^{\mathrm{B}}$ are the allele frequencies in populations A and B, and $\mathbb{E}(\theta)$ and $\mathbb{V}(\theta)$ are the mean and variance of individual ancestry. The sum of the first three terms represents the genic variance and the last term represents the LD contribution.

Demographic history

From Eq. 1, it is clear that, conditional on the genetic architecture in the source populations (β, f^A, f^B), V_g is a function of the mean, $\mathbb{E}(\theta)$, and variance, $\mathbb{V}(\theta)$, of individual ancestry in the admixed population. We consider two demographic models that affect $\mathbb{E}(\theta)$ and $\mathbb{V}(\theta)$ in qualitatively different ways. In the first model, the source populations meet once t generations ago (we refer to this as $t = 0$) in proportions p and $1 - p$, after which there is no subsequent admixture (Fig. 1A). In the second model, there is continued gene flow in every generation from one of the source populations such that the mean overall amount of ancestry from population A is the same as in the first model (Fig. 1A). For brevity, we refer to these as the hybrid-isolation (HI) and continuous gene flow (CGF) models, respectively, following Pfaff *et al.* (2001) [24]. $\mathbb{V}(\theta)$ is also affected by assortative mating based on ancestry (hereafter referred to as assortative mating for brevity) and we model this following Zaitlen *et al.* (2017) using a parameter $P \in (0, 1)$, which represents the correlation of the ancestry of individuals across mating pairs in the population[25].

Under these conditions, the behavior of $\mathbb{E}(\theta)$ and $\mathbb{V}(\theta)$ has been described previously [25, 26] (Fig. 1B and C). Briefly, in the HI model, $\mathbb{E}(\theta)$ remains constant at p in the generations after admixture as there is no subsequent gene flow. $\mathbb{V}(\theta)$ is at its maximum at $t = 0$ when each individuals carries chromosomes either from population A or B, but not both. This genome-wide correlation in ancestry breaks down in subsequent generations as a function of mating, independent assortment, and recombination, leading to a decay in $\mathbb{V}(\theta)$, the rate depending on the strength of assortative mating (Fig. 1C). In the CGF model, both $\mathbb{E}(\theta)$ and $\mathbb{V}(\theta)$ increase with time as new chromosomes are introduced from the source populations. But while $\mathbb{E}(\theta)$ continues to increase monotonically, $\mathbb{V}(\theta)$ will plateau and decrease due to the countervailing effects of independent assortment and recombination which redistribute ancestry in the population, reaching equilibrium at zero if there is no more gene flow and the population is mating randomly. $\mathbb{V}(\theta)$ provides an intuitive and quantitative measure of the degree of population structure (along the axis of ancestry) in admixed populations.

Results

Genetic variance in admixed populations

To understand the expectation of genetic variance in admixed populations, it is first worth discussing its behavior in the source populations. In Eq. 1, the first term represents the within-population component (V_{gw}) and the last three terms altogether represent the component of genetic variance between populations A and B (V_{gb}). Note that $V_{gb} = \frac{(\bar{g_A} - \bar{g_B})^2}{2}$ is positive only if there is a difference in the mean genotypic values (Fig. 2). This variance increases with genetic divergence since the expected values of both $(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2$ and $(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}})$ are functions of F_{ST}. But while $(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2$ is expected

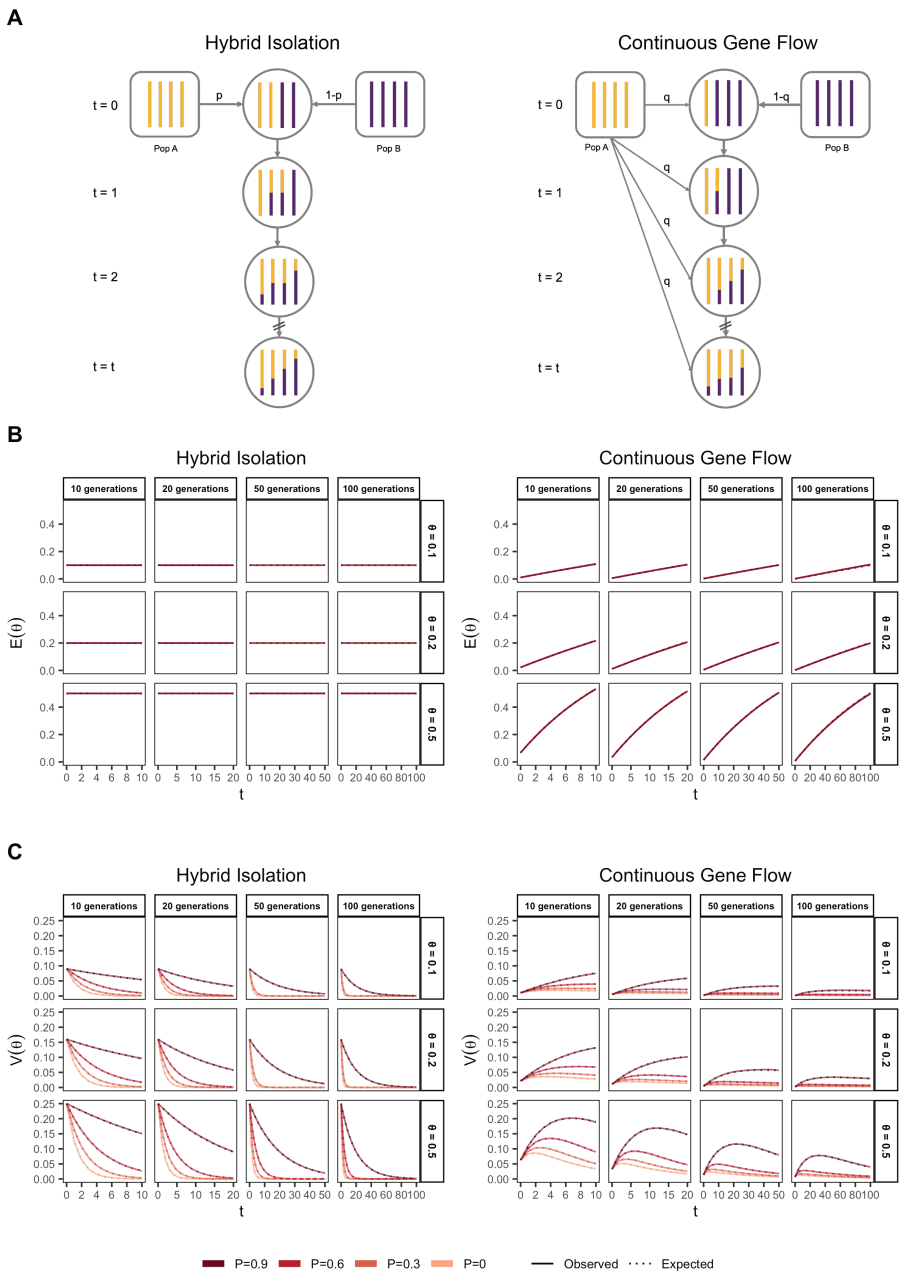

Figure 1: The behavior of mean and variance of individual ancestry as a function of admixture history. (A) Shows the demographic models under which simulations were carried out. Admixture might occur once (Hybrid Isolation, HI, left column) or continuously (Continuous Gene Flow, CGF, right column). (B) The mean individual ancestry, $\mathbb{E}(\theta)$ remains constant over time in the HI model and increases in the CGF model with continued gene flow. (C) The variance in individual ancestry, $\mathbb{V}(\theta)$ is maximum at $t = 0$ in the HI model, decaying subsequently. $\mathbb{V}(\theta)$ increases with gene flow in the CGF model and will subsequently decrease with time. P measures the strength of assortative mating, which slows the decay of $\mathbb{V}(\theta)$. P=0.6 is missing for simulations run for 50 and 100 generations and $\theta \in \{0.1, 0.2\}$ due to the difficulty in finding mate pairs (Methods).

to increase monotonically with increasing divergence, $(f_i^A - f_i^B)(f_j^A - f_j^B)$ is expected to be zero under neutrality because the direction of frequency change will be uncorrelated across loci. In this case, the LD contribution, i.e., (1.4), is expected to be zero and $V_{gb} = (1.1) + (1.2) + (1.3)$. However, this is true only in expectation over the evolutionary process and the realized LD contribution may be non-zero even for neutral traits.

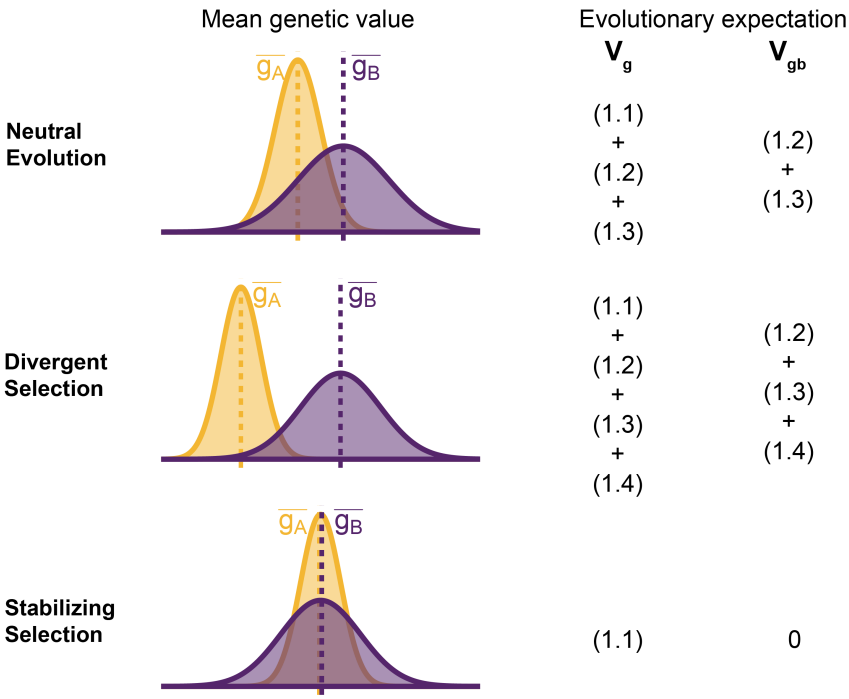

Figure 2: Decomposing genetic variance in a two-population system. The plot illustrates the expected distribution of genetic values in two populations under different selective pressures and the terms on the right list the total (V_g) and between-population genetic variance (V_{gb}) expected over the evolutionary process. For neutrally evolving traits (top row), we expect there to be an absolute difference in the mean genetic values ($|\bar{g_A} - \bar{g_B}|$) that is proportional to F_{ST}. For traits under divergent selection (middle), $|\bar{g_A} - \bar{g_B}|$ is expected to be greater than that expected under genetic drift. For traits under stabilizing selection, $|\bar{g_A} - \bar{g_B}|$ will be less than that expected under genetic drift, and zero in the extreme case.

For traits under selection, the LD contribution is expected to be greater or less than zero, depending on the type of selection. Under divergent selection, trait-increasing alleles will be systematically more frequent in one population over the other, inducing positive LD across loci [27, 28], increasing the LD contribution, i.e., term (1.4). Stabilizing selection, on the other hand, induces negative LD [29, 30]. In the extreme case, the mean genetic values of the two populations are exactly equal and $V_{gb} = (1.2) + (1.3) + (1.4) = 0$. For this to be true, (1.4) has to be negative and equal to (1.2) + (1.3), which are both positive, and the total genetic variance is reduced to the within-population variance, i.e., term (1.1) (Fig. 2). This is relevant because, as we show in the following sections, the behavior of the genetic variance in admixed populations depends on the magnitude of V_{gb} between the source populations.

We illustrate this by tracking the genetic variance in admixed populations for two traits, both with the same mean F_{ST} at causal loci but with different LD contributions (term 1.4): one where the LD

contribution is positive (Trait 1) and the other where it is negative (Trait 2). Thus, traits 1 and 2 can be thought of as examples of phenotypes under divergent and stabilizing selection, respectively, and we refer to them as such from hereon. To simulate the genetic variance of such traits, we drew the allele frequencies (f^{A} and f^{B}) in populations A and B for 1,000 causal loci with $F_{ST} \sim 0.2$ using the Balding-Nichols model [31]. We drew their effects (β) from $\mathcal{N}(0, \frac{1}{2m\bar{f}(1-\bar{f})})$ where $\bar{f}$ is the mean allele frequency between the two populations, m is the number of loci. To simulate positive and negative LD, we permuted the effect signs across variants 100 times and selected the combinations that gave the most positive and negative LD contribution to represent the genetic architecture of traits that might be under directional (Trait 1) and stabilizing (Trait 2) selection, respectively (Methods). We simulated the genotypes of 10,000 individuals under the HI and CGF models for $t \in \{10, 20, 50, 100\}$ generations post-admixture and calculated genetic values for both traits using $g = \sum_{i=1}^{m} \beta_i x_i$, where $m = 1,000$ (Method). The observed genetic variance at any time can then be calculated simply as the variance in genetic values, i.e. $V_g = \mathbb{V}(g)$.

In the HI model, $\mathbb{E}(\theta)$ does not change (Fig. 1B) so terms (1.1) and (1.2) are constant through time. Terms (1.3) and (1.4) decay towards zero as the variance in ancestry goes to zero and V_g ultimately converges to (1.1) + (1.2) (Fig. 3). This equilibrium value is equal to the $\mathbb{E}(V_g|\theta)$ (Appendix) and the rate of convergence depends on the strength of assortative mating, which slows the rate at which $\mathbb{V}(\theta)$ decays. V_g approaches equilibrium from a higher value for traits under divergent selection and lower value for traits under stabilizing selection because of positive and negative LD contributions, respectively, at $t = 0$ (Fig. 3). In the CGF model, V_g increases initially for both traits with increasing gene flow (Fig. 3). This might seem counter-intuitive at first because gene flow increases admixture LD, which leads to more negative values of the LD contribution for traits under stabilizing selection (Fig. S1). But this is outweighed by positive contributions from the genic variance – terms (1.1) + (1.2) + (1.3) – all of which initially increase with gene flow (Fig. S1). After a certain point, the increase in V_g slows down as any increase in $\mathbb{V}(\theta)$ due to gene flow is counterbalanced by recombination and independent assortment. Ultimately, V_g will decrease if there is no more gene flow, reaching the same equilibrium value as in the HI model, i.e., $\mathbb{E}(V_g|\theta) = (1.1) + (1.2)$. Because the loci are unlinked, we refer to the sum (1.3) + (1.4) as the contribution of population structure.

GREML estimation of h^2_{snp}

In their original paper, Yang *et al.* (2010) defined h^2_{snp} as the variance explained by genotyped SNPs and not as heritability [3]. This is because h^2 is the genetic variance explained by causal variants, which are unknown. Genotyped SNPs may not overlap with or tag all causal variants and thus, h^2_{snp} is understood to be a lower bound of h^2, both being equal if causal variants are known [3]. Our goal is to demonstrate that this may not be true in structured populations and quantify the bias in $\hat{h}^2_{snp}$, even in the ideal situation when causal variants are known.

We used GREML, implemented in GCTA [3, 5], to estimate the genetic variance for our simulated traits. GCTA assumes the following model: $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$ where $\boldsymbol{Z}$ is an $n \times m$ standardized genotype matrix such that the genotype of the kth individual at the ith locus is $z_{ik} = \frac{x_{ik} - 2f_i}{\sqrt{2f_i(1-f_i)}}$, f_i being the

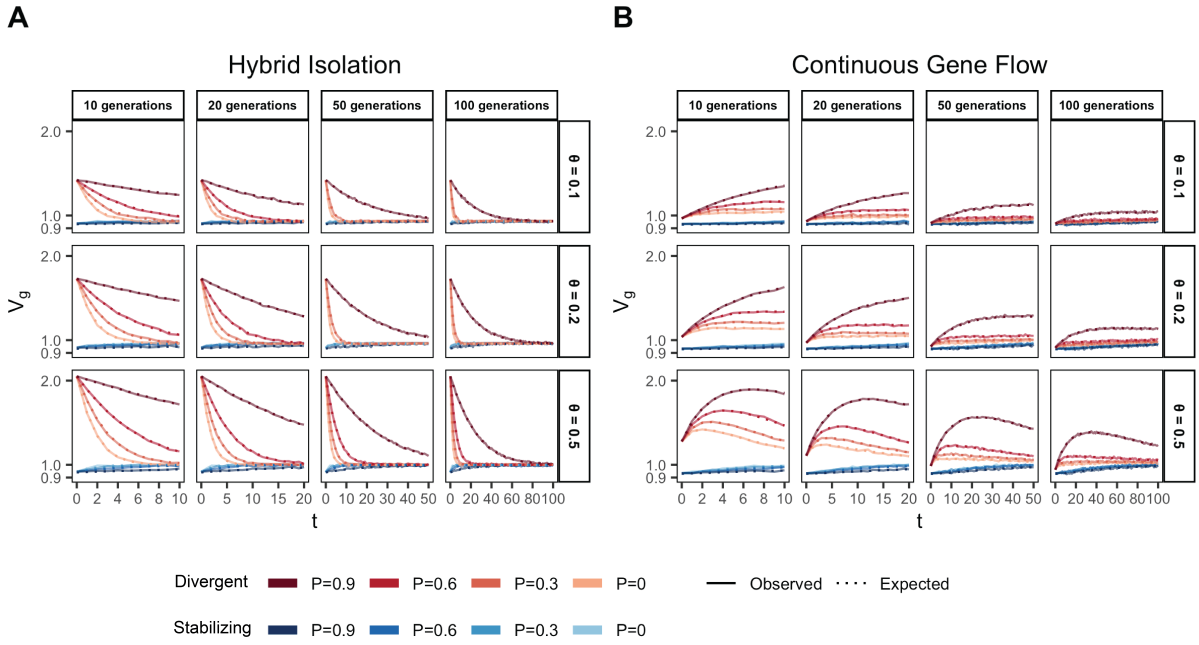

Figure 3: Genetic variance in admixed populations under the (A) HI and (B) CGF models. Dotted lines represent the expected genetic variance based on Eq. (1) and solid lines represent results of simulations averaged over ten replicates. Red and blue lines represent traits under divergent and stabilizing selection, respectively. $P = 0.6$ is missing for simulations run for 50 and 100 generations and $\theta \in \{0.1, 0.2\}$ due to the difficulty in finding mate pairs (Methods)

allele frequency. The SNP effects corresponding to the scaled genotypes are assumed to be random and independent such that $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{I}\frac{\sigma_u^2}{m})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}\sigma_\epsilon^2)$ is random environmental error. Then, the phenotypic variance can be decomposed as:

$$\begin{aligned}\mathbb{V}(\boldsymbol{y}) &= \mathbb{V}(\boldsymbol{Z}\boldsymbol{u}) + \mathbb{V}(e) \\ &= \frac{\boldsymbol{Z}\boldsymbol{Z}'}{m}\sigma_u^2 + \sigma_\epsilon^2\end{aligned}$$

where $\frac{\boldsymbol{Z}\boldsymbol{Z}'}{m}$ is the genetic relationship matrix (GRM), the variance components σ_u^2 and σ_ϵ^2 are estimated using restricted maximum likelihood, and $\hat{h}^2_{snp}$ is calculated as $\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\hat{\sigma}_\epsilon^2}$. We are interested in asking whether $\hat{\sigma}_u^2$ is an unbiased estimate of V_g. To answer this, we constructed the GRM with causal variants and estimated $\hat{\sigma}_u^2$ using GCTA [3, 4].

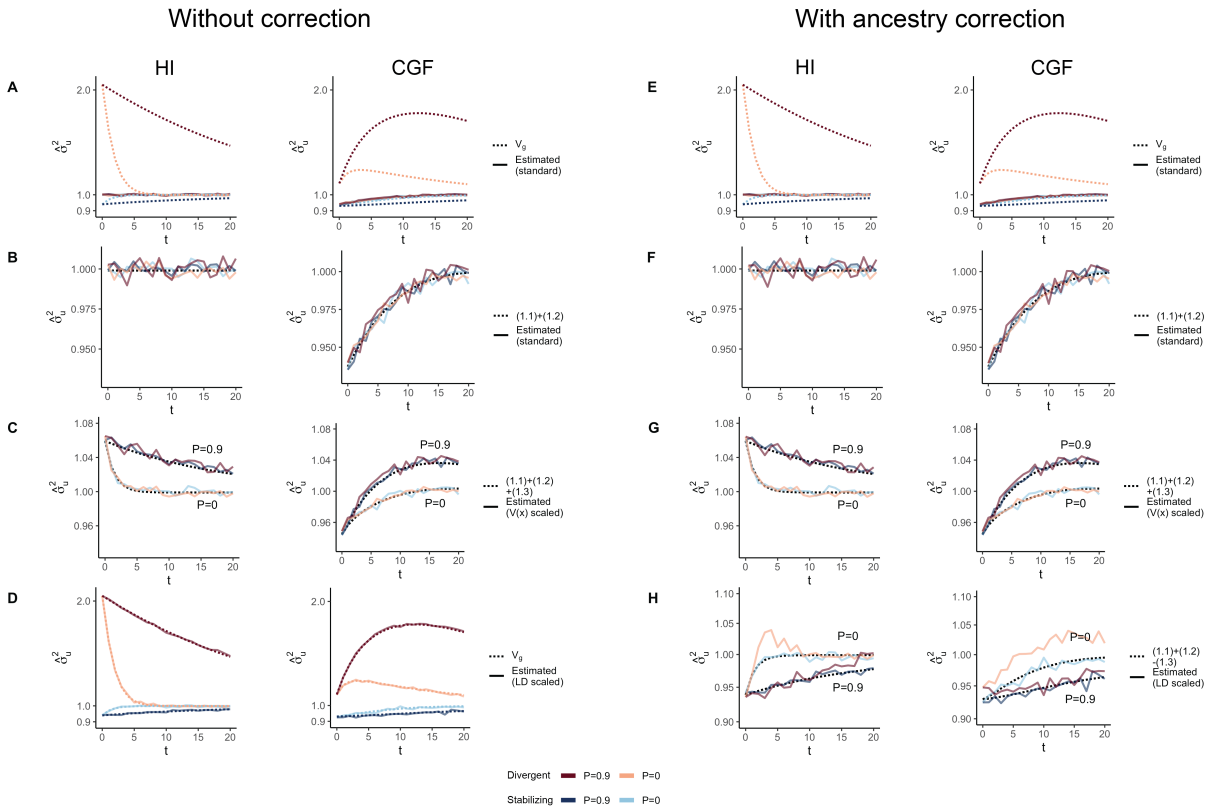

Figure 4: The behavior of GREML estimates of the genetic variance ($\hat{\sigma}_u^2$) in admixed populations under the HI (left column) and CGF (right column) models either without (A-D) or with (E-H) individual ancestry as a fixed effect. The solid lines represent estimates from simulated data averaged across ten replicates with red and blue colors representing estimates for traits under divergent and stabilizing selection, respectively. P indicates the strength of assortative mating. The dotted lines either represent the expected variance in the population based on Eq. 1 (A & B) or the expected estimate for three different ways of scaling genotypes (B-D & F-H). (A-B & E-F) show the behavior of $\hat{\sigma}_u^2$ for the default scaling, (C, G) shows $\hat{\sigma}_u^2$ when the genotype at a locus is scaled by its sample variance ($\mathbb{V}(x)$ scaled), and (D, H) when it is scaled by the sample covariance (LD scaled).

GCTA under- and over-estimates the genetic variance in admixed populations for traits under diver-

gent (Trait 1) and stabilizing selection (Trait 2), respectively, when there is population structure, i.e., when $\mathbb{V}(\theta) > 0$ (Fig. 4A). One reason for this bias is that the GREML model assumes that the effects are independent, and therefore the LD contribution is zero. This, as discussed in the previous section, is not true for traits under divergent or stabilizing selection between the source populations, and only true for neutral traits in expectation. Because of this, $\hat{\sigma}_u^2$ does not capture the LD contribution, i.e. term (1.4) (Fig. 4A). But $\hat{\sigma}_u^2$ can be biased even if the LD contribution is zero if the genotypes are scaled with $\sqrt{2f_i(1-f_i)}$ – the standard practice – where f_i is the frequency of the allele in the population. This scaling assumes that $\mathbb{V}(x_i) = 2f_i(1-f_i)$, which is true only if the population were mating randomly. In an admixed population $\mathbb{V}(x_i) = 2f_i(1-f_i) + 2\,\mathbb{V}(\theta)(f_i^A - f_i^B)^2$, where f_i, f_i^A, and f_i^B correspond to frequency in the admixed population, and source populations, A and B, respectively (Appendix). Alternatively, if the genotypes are scaled, $\mathbb{V}(z_i) = 1 + 2\,\mathbb{V}(\theta)F_{st}^{(i)}$ where $F_{st}^{(i)}$ is the F_{st} at the ith locus. We show that this assumption biases $\hat{\sigma}_u^2$ downwards by a factor of $2\,\mathbb{V}(\theta)(f_i^A - f_i^B)^2$ (or $2\,\mathbb{V}(\theta)F_{st}^{(i)}$ if genotypes are scaled) – term (1.3) (Fig. 4B, Appendix). Thus, with the standard scaling, $\hat{\sigma}_u^2$ gives a biased estimate in the presence of population structure, even of the genic variance.

The overall bias in $\hat{\sigma}_u^2$ is determined by the relative magnitude and direction of terms (1.3) and (1.4), both of which are functions of $\mathbb{V}(\theta)$, and therefore, of the degree of structure in the population. The contribution of term (1.3) will be modest, even in highly structured populations (Fig. S1) and therefore, the overall bias is largely driven by the LD contribution. If there is no more gene flow, $\mathbb{V}(\theta)$ will ultimately go to zero and V_g will converge towards $\hat{\sigma}_u^2$. Thus, $\hat{\sigma}_u^2$ is more accurately interpreted as the genetic variance expected if the LD contribution were zero and if the population were mating randomly. In other words, $\mathbb{E}(\hat{\sigma}_u^2) = (1.1) + (1.2) \neq V_g$ (Fig. 4B).

In principle, we can recover the missing components of V_g by scaling the genotypes appropriately. For example, we can recover term (1.3) by scaling the genotype at each variant i by its sample variance, i.e., $z_{ik} = \frac{x_{ik}-2f_i}{\sqrt{\mathbb{V}(x_i)}}$ (Fig. 4C) (Appendix). We can also recover term (1.4) by scaling the genotypes with the covariance between SNPs, i.e., the LD matrix, as previously proposed [32, 33] (Methods). In matrix form, the 'LD-scaled' genotypes can be written as $\boldsymbol{Z} = \boldsymbol{L}^{-1}(\boldsymbol{X} - 2\boldsymbol{P})'$ where $\boldsymbol{P}$ is an $n \times m$ matrix such that all elements of the ith column contain the frequency of the ith SNP and $\boldsymbol{L}$ is the (lower diagonal) square root matrix of the LD matrix $\boldsymbol{\Sigma}, i.e., \boldsymbol{\Sigma} = \boldsymbol{LL'}$ [32]. GREML recovers the LD contribution under this scaling, resulting in unbiased estimates of V_g for both traits (Fig. 4D).

In practice, however, the LD contribution may not be fully recoverable for two reasons. One, the LD-scaled GRM requires computing the inverse of $\boldsymbol{\Sigma}$ or $\boldsymbol{L}$ which may not exist, especially if the number of markers is greater than the sample size – the case for most human genetic studies. Second, it is common to include individual ancestry or principal components of the GRM as fixed effects in the model to correct for population stratification. This should also have the effect of removing the components of genetic variance along the ancestry axes, the residual variance being equal to $\mathbb{E}\{\mathbb{V}(g|\theta)\} = (1.1) + (1.2) - (1.3)$ (Appendix). Indeed, this is what we observe in Fig. 4H. Thus, if ancestry is included as a fixed effect, we expect V_g to be underestimated in the presence of population structure, regardless of genetic architecture.

Local ancestry heritability

A related quantity of interest in admixed populations is local ancestry heritability (h^2_γ), which is defined as the proportion of phenotypic variance that can be explained by local ancestry. Zaitlen *et al.* (2014) [34] showed that this quantity is related to, and can be used to estimate, h^2 in admixed populations. The advantage of this approach is that local ancestry segments shared between individuals are identical by descent and are therefore, more likely to tag causal variants compared to array markers, allowing one to potentially capture the contributions of rare variants [34]. Here, we show that (i) the relationship between h^2_γ and h^2 is not straightforward and (ii) $\hat{h}^2_\gamma$ may not be an unbiased estimate of local ancestry heritability under the random effects model for the same reasons that $\hat{h}^2_{snp}$ is not unbiased.

We define local ancestry $\gamma_i \in \{0, 1, 2\}$ as the number of alleles at locus i that trace their ancestry to population A. Thus, ancestry at the ith locus in individual k is a binomial random variable with $\mathbb{E}(\gamma_{ik}) = 2\theta_k$, θ_k being the ancestry of the kth individual. Similar to genetic value, the 'ancestry value' of an individual can be defined as $\sum_{i=1}^m \phi_i\gamma_i$, where $\phi_i = \beta_i(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})$ is the effect size of local ancestry (Appendix). Then, the genetic variance due to local ancestry can be expressed as (Appendix):

$$
\begin{aligned}
V_\gamma =& \mathbb{V}\left(\sum_{i=1}^m \phi_i\gamma_i\right) = \sum_{i=1}^m \phi_i^2\,\mathbb{V}(\gamma_i) + \sum_{i=1}^m\sum_{j\neq i} \phi_i\phi_j\,\mathbb{C}(\gamma_i,\gamma_j)\\
=&2\,\mathbb{E}(\theta)\{1-\mathbb{E}(\theta)\}\sum_{i=1}^m \phi_i^2 + 2\,\mathbb{V}(\theta)\sum_{i=1}^m \phi_i^2 + 4\,\mathbb{V}(\theta)\sum_{i=1}^m\sum_{j\neq i}\phi_i\phi_j\\
=&2\,\mathbb{E}(\theta)\{1-\mathbb{E}(\theta)\}\sum_{i=1}^m \beta_i^2(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2\\
&+2\,\mathbb{V}(\theta)\sum_{i=1}^m \beta_i^2(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2\\
&+4\,\mathbb{V}(\theta)\sum_{i=1}^m\sum_{j\neq i}\beta_i\beta_j(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}})
\end{aligned}
$$

and heritability explained by local ancestry is simply the ratio of V_γ and the phenotypic variance. Note that $V_\gamma = (1.2) + (1.3) + (1.4)$ and therefore its behavior is similar to V_g in that the terms (1.3) and (1.4) decay towards zero as $\mathbb{V}(\theta) \to 0$, and V_γ converges to (1.2) (Fig. S2). Additionally, the dependence of V_γ on both $\mathbb{E}(\theta)$ and $\mathbb{V}(\theta)$ precludes a straightforward derivation between local ancestry heritability and h^2.

GREML estimation of $\hat{h}^2_\gamma$ is similar to that of $\hat{h}^2_{snp}$, the key difference being that the former involves constructing the GRM using local ancestry instead of genotypes [34]. The following model is assumed: $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{v} + \boldsymbol{\xi}$ where $\boldsymbol{W}$ is an $n \times m$ standardized local ancestry matrix, $\boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{I}\frac{\sigma_v^2}{m})$ are local ancestry effects, and $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}\sigma_\xi^2)$. Note that σ_ξ^2 captures both environmental noise as well as any genetic variance independent of local ancestry. The phenotypic variance is decomposed as $\mathbb{V}(\boldsymbol{y}) = \mathbb{V}(\boldsymbol{W}\boldsymbol{v}) + \mathbb{V}(\boldsymbol{\xi}) = \frac{\boldsymbol{W}\boldsymbol{W}'}{m}\sigma_v^2 + \sigma_\xi^2$ where $\frac{\boldsymbol{W}\boldsymbol{W}'}{m}$ is the local ancestry GRM and σ_v^2 is the parameter of interest, which is believed to be equal to V_γ – the genetic variance due to local ancestry.

We show that, in the presence of population structure, i.e., when $\mathbb{V}(\theta) > 0$, $\hat{\sigma}_v^2$ is biased downwards relative to V_γ for traits under divergent selection and upwards for traits under stabilizing selection because it does not capture the contribution of LD (Fig. 5A). But there is another source of bias in $\hat{\sigma}_v^2$,

which tends to be inflated in the presence of population structure if individual ancestry is not included as a covariate, even with respect to the expectation of V_γ under equilibrium (seen more clearly in Fig. 5B-C). We suspect this inflation is because of strong correlations between local ancestry – local ancestry disequilibrium – across loci that inflates $\hat{\sigma}_v^2$ in a way that is not adequately corrected even when all causal variants are included in the model [4, 10]. Scaling local ancestry by its covariance removes this bias and recovers the contribution of LD (Fig. 5D) presumably because this accounts for the correlation in genotypes across loci. Including individual ancestry as a fixed effect also corrects for the inflation in $\hat{\sigma}_v^2$ (Fig. 5E-H). But as with $\hat{\sigma}_u^2$, this practice will underestimate the genetic variance due to local ancestry in the presence of population structure because it removes the variance along the ancestry axis (Fig. 5E-H).

Based on the above, $\hat{h}_\gamma^2$ and corresponding estimates of h^2 are more accurately interpreted as the heritability due to local ancestry and heritability, respectively, expected in the population in the absence of gene flow and under random mating. We believe $\hat{h}_\gamma^2$ is still useful in that, because it should capture the effects of rare variants, it can be used to estimate the upper bound of $\hat{h}^2_{snp}$.

In a previous paper, we suggested that local ancestry heritability could potentially be used to estimate the genetic variance between populations [35]. Our results suggest this is not possible for two reasons. First, the GREML estimator of local ancestry heritability, as we show in this section is biased and does not capture the LD contribution. But even if we were able to recover the LD component, our decomposition shows that local ancestry is equal to the genetic variance between populations (V_{gb}) only when $\mathbb{E}(\theta) = 0.5$ and $\mathbb{V}(\theta) = \mathbb{E}(\theta)\{1-\mathbb{E}(\theta)\} = 0.25$, which is only possible at $t = 0$ in the HI model. After admixture, $\mathbb{V}(\theta)$ decays and the equivalence between V_γ and V_{gb} is lost, making it impossible to estimate the latter from admixed populations, especially for traits under divergent or stabilizing selection, even if the environment is randomly distributed with respect to ancestry. We note that this conclusion was recently reached independently by Schraiber and Edge (2023) [36].

How much does population structure contribute in practice?

In the previous sections, we showed theoretically that $\hat{h}^2_{snp}$ may be biased in admixed populations even if the causal variants are known and in the absence of confounding. Ultimately, whether or not this is true in practice is an empirical question, which is difficult to answer because the causal variants, their F_{ST}, and the correlation between their effect sizes are unknown. Here, we sought to answer a related question: to what extent does population structure contribute to the variance explained by GWAS SNPs among Americans with mixed African and European ancestry? To answer this, we calculated the total genetic variance explained in the 1000 Genomes [37] ASW (African Americans from the Southwest) using independent genome-wide significant SNPs for 26 quantitative traits from the GWAS catalog [38]. We decomposed the variance explained into the four components in Equation 1 using allele frequencies (f^A and f^B) from the 1000 Genomes YRI (Yorubans from Ibadan Nigeria) and CEU (Northern Europeans from Utah), and the mean ($\mathbb{E}(\theta) \approx 0.77$) and variance ($\mathbb{V}(\theta) \approx 0.02$) of individual ancestry from ASW (Methods).

We show that for skin pigmentation – a trait under strong divergent selection – the LD contribution,

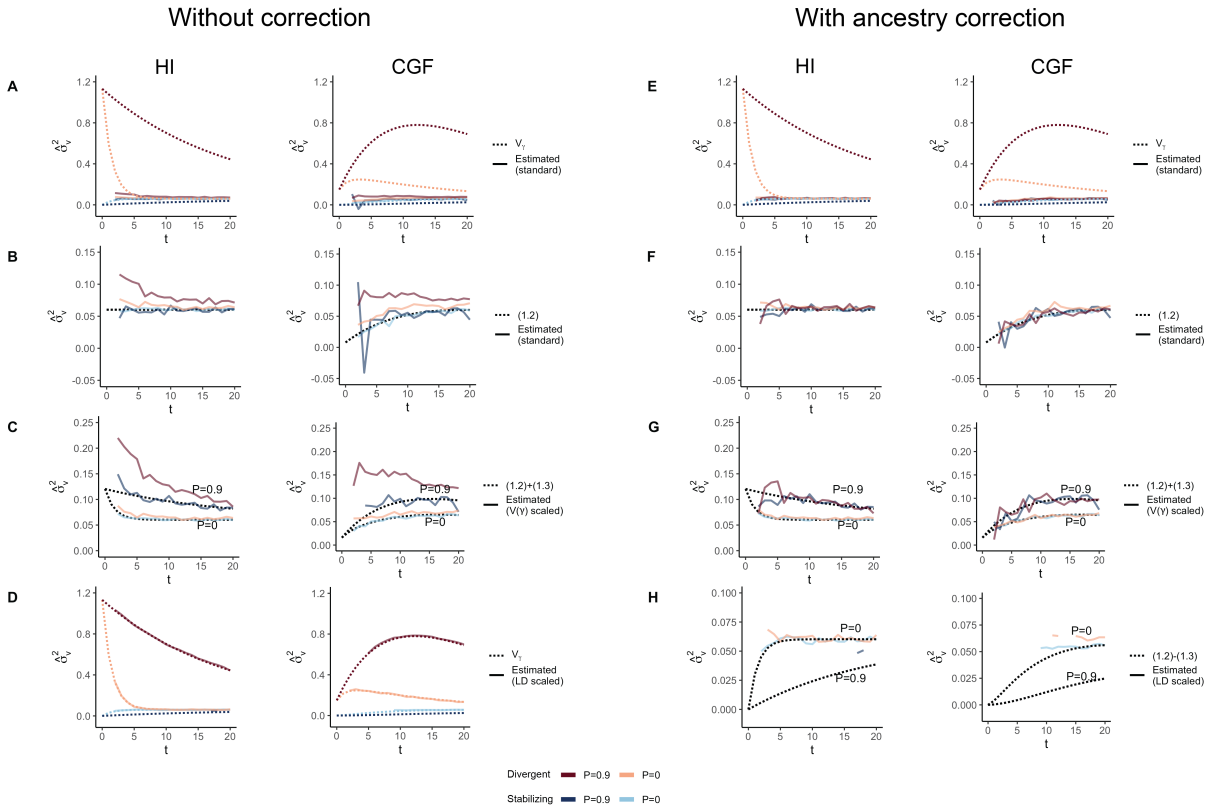

Figure 5: The behavior of GREML estimates of the variance due to local ancestry ($\hat{\sigma}_v^2$) in admixed populations under the HI (left column) and CGF (right column) models either without (A-D) or with (E-H) individual ancestry included as a fixed effect. The solid lines represent estimates from simulated data averaged across ten replicates with red and blue colors representing estimates for traits under divergent and stabilizing selection, respectively. P indicates the strength of assortative mating. The dotted lines either represent the expected variance in the population (A & B) or the expected estimate for three different ways of scaling local ancestry (B-D & F-H). (A-B & E-F) show the behavior of $\hat{\sigma}_v^2$ for the default scaling, (C, G) shows $\hat{\sigma}_v^2$ when local ancestry is scaled by the sample variance, and (D, H) when it is scaled by the sample covariance. Some runs in (D & H) failed to converge as seen by the missing segments of the solid lines because the expected variance in such cases was too small.

i.e. term (1.4), is positive and accounts for $\approx 40 - 50\%$ of the total variance explained. This is because of large allele frequency differences between Africans and Europeans that are correlated across skin pigmentation loci, consistent with strong polygenic selection favoring alleles for darker pigmentation in regions with high UV exposure and vice versa [35, 39–42]. But for most other traits, LD contributes relatively little, explaining a modest, but non-negligible proportion of the genetic variance in height, LDL and HDL cholestrol, mean corpuscular hemoglobin (MCH), neutrophil count (NEU), and white blood cell count (WBC) (Fig. 6). Because we selected independent associations for this exercise (Methods), the LD contribution is driven entirely due to population structure among ASW. The contribution of population structure to the genic variance, i.e., term (1.3) is also small even for traits like skin pigmentation and neutrophil count with large effect alleles that are highly diverged in frequency between Africans and Europeans [40, 41, 43–45]. Overall, this suggests that population structure contributes relatively little, as least to the variance explained by GWAS SNPs.

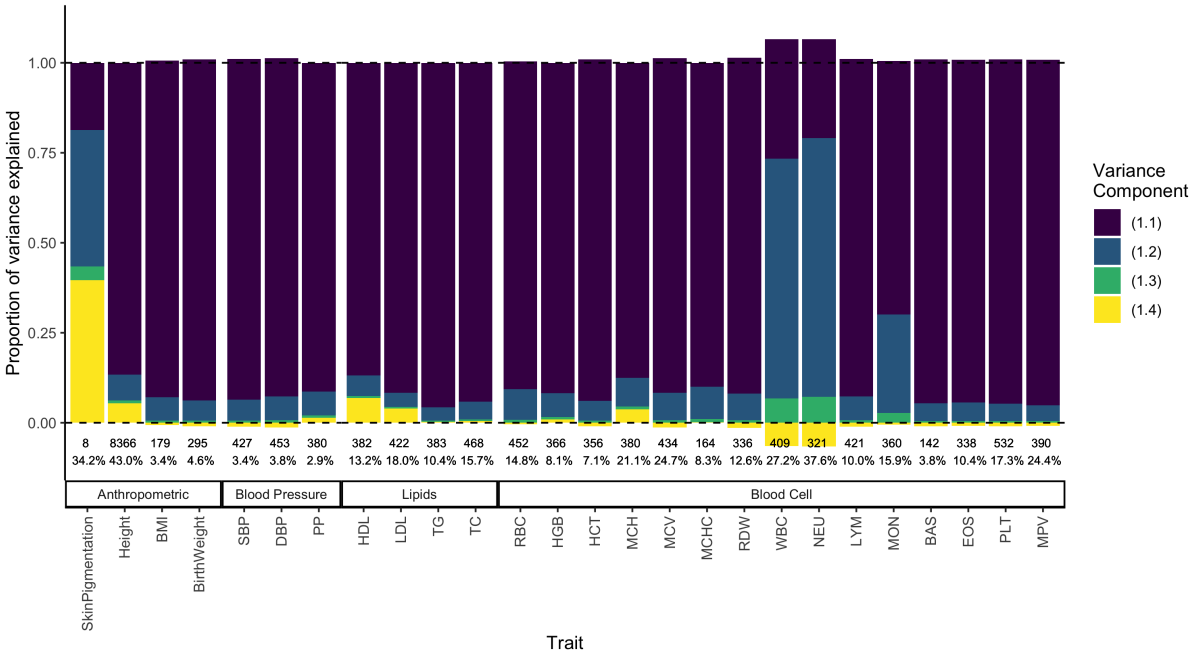

Figure 6: Decomposing the genetic variance explained by GWAS SNPs in the 1000 Genomes ASW (African Americans from Southwest). We calculated the four variance components listed in Eq. 1, their values shown on the y-axis as a fraction of the total variance explained (shown as percentage at the bottom). The number of variants used to calculate variance components for each trait is also shown at the bottom.

Discussion

Despite the growing size of GWAS and discovery of thousands of variants for hundreds of traits [38], the heritability explained by GWAS SNPs remains a fraction of twin-based heritability estimates. Yang *et al.* (2010) introduced the concept of SNP heritability (h^2_{snp}) that does not depend on the discovery of causal variants but assumes that they are numerous and are more or less uniformly distributed across the

genome (the infinitesimal model), their contributions to the genetic variance 'tagged' by genotyped SNPs [3]. h^2_{snp} is now routinely estimated in most genomic studies and at least for some traits (e.g. height and BMI), these estimates now approach twin-based heritability [6]. But despite the widespread use of $\hat{h}^2_{snp}$, its interpretation remains unclear, particularly in the presence of admixture and population structure. It is generally accepted that $\hat{h}^2_{snp}$ can be biased in structured populations because of confounding effects of unobserved environmental factors and LD between causal variants [4, 7, 9–11, 46]. But $\hat{h}^2_{snp}$ may be biased even in the absence of confounding because of misspecification of the GREML model, i.e., if the model does not represent the genetic architecture from which the trait is sampled [14–16, 47, 48].

Under the standard GREML model, SNP effects are assumed to be uncorrelated and the total genetic variance can be represented as the sum of the variance explained by individual loci, i.e. the genic variance [14–16]. In admixed populations, there is substantial LD, which can contribute to the genetic variance, and can persist for a number of generations, despite recombination, due to continued gene flow and/or ancestry-based assortative mating. GREML does not capture this LD contribution [12, 15], and therefore, may lead to biased estimates of h^2_{snp}. The LD contribution can be negative for traits under stabilizing selection, and positive for traits under divergent selection between the source populations, leading to over- or under-estimates, respectively. Thus, GREML estimates of h^2_{snp}, assuming genotypes are scaled properly (see below), is better interpreted as the proportion of phenotypic variance explained by the *genic* variance. Estimates of local ancestry heritability ($\hat{h}^2_{\gamma}$) [34, 49] should be interpreted similarly.

We show that $\hat{h}^2_{snp}$ can be biased even when the LD contribution is zero if the genotypes are scaled by $\sqrt{2f(1-f)}$ – the standard approach, which implicitly assumes a randomly mating population. In the presence of population structure, the variance in genotypes can be higher and $\hat{h}^2_{snp}$ does not capture this additional variance, which we show can be recovered by scaling genotypes by the SNP variance ($\sqrt{Var(x)}$). In principle, the LD contribution can also be recovered by scaling genotypes by the SNP covariance, i.e., the LD matrix, as previously suggested [32, 33]. But this approach is limited to situations where the sample size is much larger than the number of markers.

One limitation of this research is that it focuses on the GREML estimator of h^2_{snp} because of its widespread use. Estimators of h^2_{snp} can be broadly grouped into random- and fixed effect estimators based on how they treat SNP effects [50]. Fixed effect estimators make fewer distributional assumptions but they are not as widely used because they require conditional estimates of all variants – a high-dimensional problem where the number of markers is often far larger than the sample size [51]. This is one reason why random effect estimators, such as GREML, are popular – because they reduce the number of parameters that need to be estimated by assuming that the effects are drawn from some distribution where the variance is the only parameter of interest. Fixed effects estimators, in principle, should be able to capture the LD contribution but this is not obvious in practice since the simulations used to evaluate the accuracy of such estimators still assume uncorrelated effects [50–52]. Further research is needed to clarify the interpretation of the different estimators of h^2_{snp} in structured populations under a range of genetic architectures.

Does the LD contribution to the genetic variance have practical implications? The answer to this depends on the context in which SNP heritability is used. $\hat{h}^2_{snp}$ can be useful in quantifying the power to detect variants in GWAS where the quantity of interest is the genic variance. But $\hat{h}^2_{snp}$ can lead to

misleading conclusions if used to measure the extent to which genetic variation contributes to phenotypic variation, in predicting the response to selection, or in defining the upper limit of polygenic prediction accuracy [2] – applications where the LD contribution is important.

Ultimately, the discrepancy between $\hat{h}^2_{snp}$ and h^2 in practice is an empirical question, the answer to which depends on the degree of population structure (which we can measure) and the genetic architecture of the trait (which we do not know *a priori*). We show that for most traits, the contribution of population structure to the variance explained by GWAS SNPs is modest among African Americans. Thus, if we assume that the genetic architecture of GWAS SNPs represents that of all causal variants, then despite incorrect assumptions, the discrepancy between $\hat{h}^2_{snp}$ and h^2 should be fairly modest. But this assumption is unrealistic given that GWAS SNPs are common variants that in most cases cumulatively explain a fraction of trait heritability. What is the LD contribution of the rest of the genome, particularly rare variants? This is not obvious and will become clearer in the near future through large sequence-based studies [53]. While these are underway, theoretical studies are needed to understand how different selection regimes influence the LD patterns between causal variants – clearly an important aspect of the genetic architecture of complex traits.

Methods

Simulating genetic architecture

We first drew the allele frequency (f^0) of 1,000 biallelic causal loci in the ancestor of populations A and B from a uniform distribution, $U(0.001, 0.999)$. Then, we simulated their frequency in populations A and B (f^{A} and f^{B}) under the Balding-Nichols model [31], such that $f^{\mathrm{A}}, f^{\mathrm{B}} \sim Beta(\frac{f^0(1-F)}{F}, \frac{(1-f^0)(1-F)}{F})$ where $F = 0.2$ is the inbreeding coefficient. We implemented this using code adapted from [54]. To avoid drawing extremely rare alleles, we continued to draw f^{A} and f^{B} until we had 1,000 loci with $f^{\mathrm{A}}, f^{\mathrm{B}} \in (0.01, 0.99)$.

We generated the effect size (β) of each locus by sampling from $\mathcal{N}(0, \frac{1}{2m\bar{f}(1-\bar{f})})$, where m is the number of loci and $\bar{f}$ is the mean allele frequency across populations A and B. Thus, rare variants have larger effects than common variants and the total genetic variance sums to 1. Given these effects, we simulated two different traits, one with a large difference in means between populations A and B (Trait 1) and the other with roughly no difference (Trait 2). This was achieved by permuting the signs of the effects 100 times to get a distribution of V_{gb} – the genetic variance between populations. This has the effect of varying the LD contribution without changing the F_{ST} at causal loci. We selected the maximum and minimum of V_{gb} to represent Traits 1 and 2.

Simulating admixture

We simulated the genotypes, local ancestry, and phenotype for 10,000 admixed individuals per generation under the hybrid isolation (HI) and continuous gene flow (CGF) models by adapting the code from Zaitlen *et al.* (2017) [25]. We denote the ancestry of a randomly selected individual k with θ, the fraction of their genome from population A. At $t = 0$ under the HI model, we set θ to 1 for individuals from

population A and 0 if they were from population B such that $\mathbb{E}(\theta) = p \in \{0.1, 0.2, 0.5\}$ with no further gene flow from either source population. In the CGF model, population B receives a constant amount q from population A in every generation starting at $t = 0$. The mean overall proportion of ancestry in the population is kept the same as the HI model by setting $q = 1 - (1-p)^{\frac{1}{t}}$ where t is the number of generations of gene flow from A. In every generation, we simulated ancestry-based assortative mating by selecting mates such that the correlation between their ancestries is $P \in \{0, 0.3, 0.6, 0.9\}$ in every generation. We do this by repeatedly permuting individuals with respect to each other until P falls within ± 0.01 of the desired value. It becomes difficult to meet this criterion when $\mathbb{V}(\theta)$ is small (Fig.1C). To overcome this, we relaxed the threshold up to 0.04 for some conditions, i.e., when $\theta \in \{0.1, 0.2\}$ and $t \geq 50$. We generated expected variance in individual ancestry using the expression in [25]. At time t since admixture, $\mathbb{V}(\theta_t) = \mathbb{V}(\theta_{t-1})\frac{(1+P)}{2}$ under the HI model where P measures the strength of assortative mating, i.e, the correlation between the ancestry between individuals in a mating pair. Under the CGF model, $\mathbb{V}(\theta_t) = q(1-q)\,\mathbb{E}(\theta_{t-1})^2 + q(1-q)\{1 - 2\,\mathbb{E}(\theta_{t-1})\} + (1-q)\,\mathbb{V}(\theta_{t-1})\frac{(1+P)}{2}$ (Appendix).

We sampled the local ancestry at each ith locus as $\gamma_i = \gamma_{if} + \gamma_{im}$ where $\gamma_{im} \sim Bin(1, \theta_m), \gamma_{if} \sim Bin(1, \theta_f)$ and θ_m and θ_f represent the ancestry of the maternal and paternal chromosome, respectively. The global ancestry of the individual is then calculated as $\theta_k = \sum_{i=1}^{m} \frac{\gamma_{im}+\gamma_{if}}{2m}$, where m is the number of loci. We sample the genotypes x_{im} and x_{if} from a binomial distribution conditioning on local ancestry. For example, the genotype on the maternal chromosome is $x_{im} \sim Bin(1, f_{\mathrm{i}}^{\mathrm{A}})$ if $\gamma_{im} = 1$ and $x_{im} \sim Bin(1, f_{\mathrm{i}}^{\mathrm{B}})$ if $\gamma_{im} = 0$ where $f_{\mathrm{i}}^{\mathrm{A}}$ and $f_{\mathrm{i}}^{\mathrm{B}}$ represent the allele frequency in populations A and B, respectively. Then, the genotype can be obtained as the sum of the maternal and paternal genotypes: $x_i = x_{im} + x_{ip}$. We calculate the genetic value of each individual as $g = \sum_{i=1}^{m} \beta_i x_i$ and the genetic variance as $\mathbb{V}(g)$.

Heritability estimation with GCTA

We used the *--reml* and *--reml-no-constrain* flags in GCTA [5] to estimate σ_u^2 and σ_v^2, the genetic variance due to genotypes and local ancestry, respectively. We could not run GCTA without noise in the genetic values so we simulated individual phenotypes with a heritability of $h^2 = 0.8$ by adding random noise $e \sim \mathcal{N}(0, V_g \frac{1-h^2}{h^2})$. We computed three different GRMs, which correspond to different transformations of the genotypes: (i) standard, (i) Variance or $V(x)$ scaled, and (ii) LD-scaled.

For the standard GRM, the genotypes at the ith SNP are standardized such that $z_i = \frac{x_i - 2f_i}{\sqrt{2f_i(1-f_i)}}$. For the variance scaled GRM, we computed $z_i = \frac{x_i - 2f_i}{\sqrt{\mathbb{V}(x_i)}}$ where $\mathbb{V}(x_i)$ is the sample variance of the genotypes at the ith SNP. The LD-scaled GRM conceptually corresponds to standardizing the genotypes by the SNP covariance, rather than its variance. Let $\boldsymbol{X}$ represent the $n \times m$ *unstandardized* matrix of genotypes and $\boldsymbol{P}$ represent an $n \times m$ matrix where the ith column contains the allele frequency of that SNP. Let $\boldsymbol{L}$ be the lower diagonal 'square root' matrix of the $m \times m$ SNP covariance matrix $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L'}$. Then, the standardized genotypes are computed as $\boldsymbol{Z} = \boldsymbol{L'}(\boldsymbol{X} - 2\boldsymbol{P})$ and the GRM becomes $(\boldsymbol{X} - 2\boldsymbol{P})\boldsymbol{\Sigma^{-1}}(\boldsymbol{X} - 2\boldsymbol{P})'$ [32]. Similarly, the three GRMs for local ancestry were computed by scaling local ancestry with (i) $\sqrt{2\bar{\gamma}_i(1-\bar{\gamma}_i)}$ where we denote $\bar{\gamma}_i$ as the mean local ancestry at the ith SNP, or with the (ii) variance, or (iii) covariance of local ancestry, respectively.

We estimated σ_u^2 and σ_v^2 with and without individual ancestry as a fixed effect to correct for any

422 confounding due to genetic stratification. This was done by using the *--qcovar* flag.

423 Estimating variance explained by GWAS SNPs

424 To decompose the variance explained by GWAS SNPs in African Americans, we needed four quantities:
425 (i) effect sizes of GWAS SNPs, (ii) their allele frequencies in Africans and Europeans, and (iii) the mean
426 and variance of global ancestry in African Americans (Equation 1).

427 We retrieved the summary statistics of 26 traits from GWAS catalog [38]. Full list of traits and the
428 source papers [42, 55–62] are listed in Table S1. To maximize the number of variants discovered, we chose
429 summary statistics from studies that were conducted in both European and multi-ancestry samples and
430 that reported the following information: effect allele, effect size, p-value, and genomic position. For birth
431 weight, we downloaded the data from the Early Growth Genetics (EGG) consortium website [59] since the
432 version reported on the GWAS catalog is incomplete. For skin pigmentation, we chose summary statistics
433 from the UKB [63] released by the Neale Lab (http://www.nealelab.is/uk-biobank) and processed by Ju
434 and Mathieson [42] to represent effect sizes estimated among individuals of European ancestry. We
435 also selected summary statistics from Lona-Durazo *et al.* (2019) where effect sizes were meta-analyzed
436 across four admixed cohorts [55]. Lona-Durazo *et al.* provide summary statistics separately with and
437 without conditioning on rs1426654 and rs35397 – two large effect variants in *SCL24A5* and *SLC45A2*.
438 We used the 'conditioned' effect sizes and added in the effects of rs1426654 and rs35397 to estimate
439 genetic variance.

440 We selected independent hits for each trait by pruning and thresholding with PLINK v1.90b6.21 [64]
441 in two steps as in Ju *et al.* (2020) [42]. We used the genotype data of GBR from the 1000 genome
442 project [37] as the LD reference panel. We kept only SNPs (indels were removed) that passed the
443 genome-wide significant threshold (*--clump-p1 5e-8*) with a pairwise LD cutoff of 0.05 (*--clump-r2 0.05*)
444 and a physical distance threshold of 250Kb (*--clump-kb 250*) for clumping. Second, we applied a second
445 round of clumping (*--clump-kb 100*) to remove SNPs within 100kb.

446 When GWAS was carried out separately in different ancestry cohorts in the same study, we used
447 inverse-variance weighting to meta-analyze effect sizes for variants that were genome-wide significant
448 (p-value $< 5 \times 10^{-8}$) in at least one cohort. This allowed us to maximize the discovery of variants such
449 as the Duffy null allele that are absent among individuals of European ancestry but polymorphic in other
450 populations [45].

451 We used allele frequencies from the 1000 Genomes CEU and YRI to represent the allele frequencies
452 of GWAS SNPs in Europeans and Africans, respectively, making sure that the alleles reported in the
453 summary statistics matched the alleles reported in the 1000 Genomes. We estimated the global ances-
454 try of ASW individuals (N = 74) with CEU and YRI individuals from 1000 genome (phase 3) using
455 ADMIXTURE 1.3.0 [65] with k=2 and used it to calculate the mean (proportion of African ancestry =
456 0.767) and variance (0.018) of global ancestry in ASW. With the effect sizes, allele frequencies, and the
457 mean and variance in ancestry, we calculated the four components of genetic variance using Equation 1
458 and expressed them as a fraction of the total genetic variance.

459 Initially, the multi-ancestry summary statistics for a few traits (NEU, WBC, MON, MCH, BAS)

yielded values > 1 for the proportion of variance explained. This is likely because, despite LD pruning, some of the variants in the model are not independent and tag large effect variants under divergent selection such as the Duffy null allele, leading to an inflated contribution of LD. We checked this by calculating the pairwise contribution , i.e., $\beta_i\beta_j(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}})$, of all SNPs in the model and show long-range positive LD between variants on chromosome 1 for NEU, WBC, and MON, especially with the Duffy null allele (Fig. S3A-C). A similar pattern was observed on chromosome 16 for MCH, confirming our suspicion. This also suggests that for certain traits, pruning and thresholding approaches are not guaranteed to yield independent hits. To get around this problem, we retained only one association with the lowest p-value, each from chromosome 1 (rs2814778 for NEU, WBC, and MON) and chromosome 16 (rs13331259 for MCH) (Fig. S3D). For BAS, we observed that the variance explained was driven by a rare variant (rs188411703, MAF $= 0.0024$) of large effect ($\beta = -2.27$). We believe this effect estimate to be inflated and therefore, we removed it from our calculation.

As a sanity check, we independently estimated the genetic variance as the variance in polygenic scores, calculated using *--score* flag in PLINK, [64] in ASW individuals. We compared the first estimate of the genetic variance to the second (Fig. S4) to confirm two things: (i) the allele frequencies, and mean and variance in ancestry are estimated correctly, and (ii) the variants are more or less independent in that they do not absorb the effects of other variants in the model. We show that the two estimates of the genetic variance are strongly correlated ($r \sim 0.85$, Fig. S4).

Code availability

We carried out all analyses in R version 4.2.3 [66], PLINK v1.90b6.21 and PLINK 2.0 [64, 67], and GCTA version 1.94.1 [5]. All code is freely available on https://github.com/jinguohuang/admix_heritability.git.

Acknowledgements

We thank Iain Mathieson and Doc Edge for helpful comments on the manuscript. This study was funded by National Institute of General Medical Sciences award R00GM137076 to A.A.Z. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits* 1–980 (Sinauer Associates, Inc, 1998).

2. Visscher, P. M., Hill, W. G., *et al.* Heritability in the genomics era–concepts and misconceptions. *Nature reviews. Genetics* **9,** 255–66 (4 2008).

3. Yang, J., Benyamin, B., *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42,** 565–569 (7 2010).

4. Yang, J., Manolio, T. A., *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics 2011 43:6* **43,** 519–525 (6 2011).

5. Yang, J., Lee, S. H., *et al.* GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88,** 76–82 (1 2011).

6. Wainschtein, P., Jain, D., *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics 2022 54:3* **54,** 263–273 (3 2022).

7. Browning, S. R. & Browning, B. L. Population structure can inflate SNP-based heritability estimates. *American Journal of Human Genetics* **89,** 191–193 (1 2011).

8. Goddard, M. E., Lee, S. H., *et al.* Response to browning and browning. *American Journal of Human Genetics* **89,** 193–195 (1 2011).

9. Kumar, S. K., Feldman, M. W., *et al.* Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E61–E70 (1 2016).

10. Yang, J., Leec, S. H., *et al.* GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *Proceedings of the National Academy of Sciences* **113,** E4579–E4580 (32 2016).

11. Lin, Z., Seal, S., *et al.* Estimating SNP heritability in presence of population substructure in biobank-scale datasets. *Genetics* **220** (4 2022).

12. Border, R., O'Rourke, S., *et al.* Assortative mating biases marker-based heritability estimators. *Nature Communications 2022 13:1* **13,** 1–10 (1 2022).

13. Visscher, P. M., Yang, J., *et al.* A Commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010). *Twin Research and Human Genetics* **13,** 517–524 (6 2010).

14. De los Campos, G., Sorensen, D., *et al.* Genomic Heritability: What Is It? *PLOS Genetics* **11,** e1005048 (5 2015).

15. Rawlik, K., Canela-Xandri, O., *et al.* SNP heritability: What are we estimating? *bioRxiv,* 2020.09.15.276121 (2020).

16. Lara, L. A. C., Pocrnic, I., *et al.* Temporal and genomic analysis of additive genetic variance in breeding programmes. *Heredity 2021 128:1* **128,** 21–32 (1 2021).

17. Wojcik, G. L., Graff, M., *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570,** 514–518 (7762 2019).

18. Ben-Eghan, C., Sun, R., *et al.* Don't ignore genetic data from minority populations. *Nature 2021 585:7824* **585,** 184–186 (7824 2020).

19. Verma, A., Damrauer, S. M., *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *Journal of Personalized Medicine* **12,** 1974 (12 2022).

20. Fatumo, S., Chikowore, T., *et al.* A roadmap to increase diversity in genomic studies. *Nature Medicine 2022 28:2* **28,** 243–250 (2 2022).

21. Of Us Research Program Investigators, T. A. The "All of Us" Research Program. *New England Journal of Medicine* **381,** 668–676 (7 2019).

22. Sohail, M., Chong, A. Y., *et al.* Nationwide genomic biobank in Mexico unravels demographic history and complex trait architecture from 6,057 individuals. *bioRxiv,* 2022.07.11.499652 (2022).

23. Johnson, R., Ding, Y., *et al.* The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. *Cell Genomics* **3,** 100243 (1 2023).

24. Pfaff, C. L., Parra, E. J., *et al.* Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *American journal of human genetics* **68,** 198–207 (1 2001).

25. Zaitlen, N., Huntsman, S., *et al.* The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium. *Genetics* **205,** 375–383 (1 2017).

26. Verdu, P. & Rosenberg, N. A. A General Mechanistic Model for Admixture Histories of Hybrid Populations. *Genetics* **189,** 1413–1426 (4 2011).

27. Corre, V. L. & Kremer, A. Genetic Variability at Neutral Markers, Quantitative Trait Loci and Trait in a Subdivided Population Under Selection. *Genetics* **164,** 1205–1219 (3 2003).

28. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics* **10,** e1004412 (8 2014).

29. Bulmer, M. G. The Effect of Selection on Genetic Variability. *The American Naturalist* **105,** 201–211 (943 1971).

30. Yair, S. & Coop, G. Population differentiation of polygenic score predictions under stabilizing selection. *Philosophical Transactions of the Royal Society B* **377** (1852 2022).

31. Balding, D. J. & Nichols, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96,** 3–12 (1-2 1995).

32. Mathew, B., Léon, J., *et al.* A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. *Heredity 2017 120:4* **120,** 356–368 (4 2017).

33. Ma, R. & Dicker, L. H. The Mahalanobis kernel for heritability estimation in genome-wide association studies: fixed-effects and random-effects methods (2019).

34. Zaitlen, N., Pasaniuc, B., *et al.* Leveraging population admixture to characterize the heritability of complex traits. *Nature Genetics* **46,** 1356–1362 (12 2014).

35. Zaidi, A. A., Mattern, B. C., *et al.* Investigating the case of human nose shape and climate adaptation. *PLoS Genetics* **13,** 2017 (3 2017).

36. Schraiber, J. G. & Edge, M. D. Heritability within groups is uninformative about differences among groups: cases from behavioral, evolutionary, and statistical genetics. *bioRxiv,* 2023.11.06.565864 (2023).

37. Auton, A., Abecasis, G. R., *et al.* A global reference for human genetic variation. *Nature 2015 526:7571* **526,** 68–74 (7571 2015).

38. Sollis, E., Mosaku, A., *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51,** D977–D985 (D1 2023).

39. Jablonski, N. G. the Evolution of Human Skin and Skin Color. *Annual Review of Anthropology* **33,** 585–623 (1 2004).

40. Lamason, R. L., Mohideen, M. A. P., *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310,** 1782–1786 (5755 2005).

41. Beleza, S., Johnson, N. A., *et al.* Genetic architecture of skin and eye color in an African-European admixed population. *PLoS genetics* **9** (ed Spritz, R. A.) e1003372 (3 2013).

42. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation in West Eurasia. *Proceedings of the National Academy of Sciences of the United States of America* **118,** e2009227118 (1 2020).

43. Nalls, M. A., Wilson, J. G., *et al.* Admixture Mapping of White Cell Count: Genetic Locus Responsible for Lower White Blood Cell Count in the Health ABC and Jackson Heart Studies. *The American Journal of Human Genetics* **82,** 81–87 (1 2008).

44. Reich, D., Nalls, M. A., *et al.* Reduced Neutrophil Count in People of African Descent Is Due To a Regulatory Variant in the Duffy Antigen Receptor for Chemokines Gene. *PLoS Genetics* **5** (ed Visscher, P. M.) e1000360 (1 2009).

45. McManus, K. F., Taravella, A. M., *et al.* Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLOS Genetics* **13,** e1006560 (3 2017).

46. Kumar, S. K., Feldman, M. W., *et al.* Reply to Yang et al.: GCTA produces unreliable heritability estimates. *Proceedings of the National Academy of Sciences* **113,** E4581–E4581 (32 2016).

47. Speed, D., Hemani, G., *et al.* Improved heritability estimation from genome-wide SNPs. *American journal of human genetics* **91,** 1011–21 (6 2012).

48. Speed, D., Cai, N., *et al.* Reevaluation of SNP heritability in complex human traits. *Nature Genetics* **49,** 986–992 (7 2017).

49. Chan, T. F., Rui, X., *et al.* Estimating heritability explained by local ancestry and evaluating stratification bias in admixture mapping from summary statistics. *bioRxiv,* 2023.04.10.536252 (2023).

50. Min, A., Thompson, E., *et al.* Comparing heritability estimators under alternative structures of linkage disequilibrium. *G3 Genes/Genomes/Genetics* **12** (8 2022).

51. Schwartzman, A., Schork, A. J., *et al.* A simple, consistent estimator of SNP heritability from genome-wide association studies. *https://doi.org/10.1214/19-AOAS1291* **13,** 2509–2538 (4 2019).

52. Hou, K., Ding, Y., *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nature Genetics 2023 55:4* **55,** 549–558 (4 2023).

53. Backman, J. D., Li, A. H., *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature 2021 599:7886* **599,** 628–634 (7886 2021).

54. Lin, M., Park, D. S., *et al.* Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies. *Frontiers in Genetics* **12,** 673167 (2021).

55. Lona-Durazo, F., Hernandez-Pacheco, N., *et al.* Meta-analysis of GWA studies provides new insights on the genetic architecture of skin pigmentation in recently admixed populations. *BMC Genetics* **20,** 1–16 (1 2019).

56. Yengo, L., Vedantam, S., *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610,** 704–712 (7933 2022).

57. Hoffmann, T. J., Choquet, H., *et al.* A large multiethnic genome-wide association study of adult body mass index identifies novel loci. *Genetics* **210,** 499–515 (2 2018).

58. Pulit, S. L., Stoneman, C., *et al.* Meta-Analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Human Molecular Genetics* **28,** 166–174 (1 2019).

59. Warrington, N. M., Beaumont, R. N., *et al.* Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nature Genetics 2019 51:5* **51,** 804–814 (5 2019).

60. Surendran, P., Feofanova, E. V., *et al.* Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. *Nature Genetics* **52,** 1314–1332 (12 2020).

61. Graham, S. E., Clarke, S. L., *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600,** 675–679 (7890 2021).

62. Chen, M. H., Raffield, L. M., *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182,** 1198–1213.e14 (5 2020).

63. Bycroft, C., Freeman, C., *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (7726 2018).

64. Chang, C. C., Chow, C. C., *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** s13742–015–0047–8 (1 2015).

65. Alexander, D. H., Novembre, J., *et al.* Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19,** 1655–64 (9 2009).

66. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2023).

67. Purcell, S., Neale, B., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81,** 559–75 (3 2007).

68. Chen, G. B. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Frontiers in Genetics* **5,** 72296 (APR 2014).

Appendix

Table of contents

1 Variance in ancestry . . . 25
2 Genetic variance - V_g . . . 25
2.1 $\mathbb{V}(x_i)$ . . . 26
2.2 $\mathbb{C}(x_i, x_j)$ . . . 27
3 The effect of genotype scaling on $\hat{\sigma}_u^2$ . . . 28
4 Genetic variance after correction for individual ancestry . . . 30
5 Effect size of local ancestry . . . 31
6 Genetic variance due to local ancestry . . . 32

1 Variance in ancestry

We denote variance and covariance with $\mathbb{V}(.)$ and $\mathbb{C}(.)$ and used the expressions in [25] to generate the expected value for the variance in ancestry, i.e., $\mathbb{V}(\theta)$. This is straightforward for the HI model, where at time t $\mathbb{V}(\theta_t) = \mathbb{V}(\theta_{t-1})\frac{(1+P_{t-1})}{2}$. $P_t = Cor(\theta_m, \theta_f)$ measures the strength of assortative mating, i.e, the correlation between the ancestry across mating pairs (θ_m, θ_f) at time t. For simplicity, we assumed this to be constant in every generation, i.e. $P_t = P_{t-1} = P$ following [25]. Since our notation slightly differs from [25], we re-derived the expression for $V(\theta_t)$ for the CGF model where population B receives a constant amount q of gene flow from population A in every generation. Note, that $\mathbb{E}(\theta_t) = q + (1-q)\,\mathbb{E}(\theta_{t-1})$. Then,

$$
\begin{aligned}
\mathbb{V}(\theta_t) =& \mathbb{E}(\theta_t^2) - \mathbb{E}(\theta_t)^2 \\
=& q + (1-q)\,\mathbb{E}\left[\left(\frac{\theta_{t-1}^m + \theta_{t-1}^f}{2}\right)\left(\frac{\theta_{t-1}^m + \theta_{t-1}^f}{2}\right)\right] - \{q + (1-q)\,\mathbb{E}(\theta_{t-1})\}^2 \\
=& q + \frac{(1-q)}{4}\{2\,\mathbb{E}(\theta_{t-1}^2) + 2\,\mathbb{E}(\theta_{t-1}^m\theta_{t-1}^f)\} - \{q^2 + 2q(1-q)\,\mathbb{E}(\theta_{t-1}) + (1-q)^2\,\mathbb{E}(\theta_{t-1})^2\} \\
=& q + \frac{1-q}{2}\,\mathbb{E}(\theta_{t-1}^2) + \frac{1-q}{2}\,\mathbb{E}(\theta_{t-1}^m\theta_{t-1}^f) - q^2 - 2q(1-q)\,\mathbb{E}(\theta_{t-1}) - (1-q)^2\,\mathbb{E}(\theta_{t-1})^2 \\
=& q(1-q) + \frac{1-q}{2}\{\mathbb{V}(\theta_{t-1}) + \mathbb{E}(\theta_{t-1})^2\} + \frac{1-q}{2}\{\mathbb{C}(\theta_{t-1}^m, \theta_{t-1}^f) + \mathbb{E}(\theta_{t-1})^2\} - 2q(1-q)\,\mathbb{E}(\theta_{t-1}) - \mathbb{E}(\theta_{t-1})^2 \\
=& q(1-q) + \frac{1-q}{2}\,\mathbb{V}(\theta_{t-1}) + \frac{1-q}{2}\,\mathbb{E}(\theta_{t-1})^2 + \frac{1-q}{2}P_{t-1}\,\mathbb{V}(\theta_{t-1}) + \frac{1-q}{2}\,\mathbb{E}(\theta_{t-1})^2 - 2q(1-q)\,\mathbb{E}(\theta_{t-1}) - \mathbb{E}(\theta_{t-1})^2 \\
=& q(1-q) + \frac{1-q}{2}\,\mathbb{V}(\theta_{t-1})\{1 + P_{t-1}\} + (1-q)\,\mathbb{E}(\theta_{t-1})^2 - 2q(1-q)\,\mathbb{E}(\theta_{t-1}) - (1-q)^2\,\mathbb{E}(\theta_{t-1})^2 \\
=& q(1-q)\,\mathbb{E}(\theta_{t-1})^2 + q(1-q)\{1 - 2\,\mathbb{E}(\theta_{t-1})\} + \frac{1-q}{2}\,\mathbb{V}(\theta_{t-1})\{1 + P_{t-1}\}
\end{aligned}
$$

2 Genetic variance - V_g

Let $y = g + e$, where y is the phenotypic value of an individual, g is the genotypic value, and e is random error. We assume additive effects such that $g = \sum_{i=1}^{m} \beta_i x_i$ where β_i is the effect size of the ith biallelic

locus and $x_i \in \{0, 1, 2\}$ is the number of copies of the trait-increasing allele. Then, the genetic variance V_g is:

$$V_g = \mathbb{V}(\sum_{i=1}^{m} \beta_i x_i) = \sum_{i=1}^{m} \beta_i^2 \, \mathbb{V}(x_i) + \sum_{j \neq i} \beta_i \beta_j \, \mathbb{C}(x_i, x_j)$$

In the following sections, we decompose $\mathbb{V}(x_i)$ and $\mathbb{C}(x_i, x_j)$ further as functions of ancestry.

2.1 $\mathbb{V}(x_i)$

We first derive $\mathbb{V}(x_i)$ as a function of ancestry (θ) using the law of total variance:

$$\mathbb{V}(x_i) = \underset{\theta}{\mathbb{E}}\{\mathbb{V}(x_i|\theta)\} + \mathbb{V}\{\underset{\theta}{\mathbb{E}}(x_i|\theta)\}$$

where $\mathbb{E}_\theta$ represents the expectation taken over θ.

2.1.1 $\mathbb{E}_\theta\{\mathbb{V}(x_i|\theta)\}$

We derive $\mathbb{V}(x_i|\theta)$ by further conditioning on the local ancestry at each locus.

$$\mathbb{V}(x_i|\theta) = \underset{\gamma}{\mathbb{E}}\{\mathbb{V}(x_i|\gamma, \theta)\} + \mathbb{V}\{\underset{\gamma}{\mathbb{E}}(x_i|\gamma, \theta)\}$$

where $\mathbb{E}_\gamma$ represents expectation taken over local ancestry. Since we are interested in the variance at a single locus, we will ignore the subscript i and denote the frequency of the trait-increasing allele in populations A and B with f^{A} and f^{B}, respectively.

$$\begin{aligned}
\underset{\gamma}{\mathbb{E}}\{\mathbb{V}(x_i|\gamma, \theta)\} =& \mathbb{V}(x_i|\gamma = 0, \theta)\, \mathbb{P}(\gamma = 0|\theta) + \mathbb{V}(x_i|\gamma = 1, \theta)\, \mathbb{P}(\gamma = 1|\theta) + \mathbb{V}(x_i|\gamma = 2, \theta)\, \mathbb{P}(\gamma = 2|\theta) \\
=& 2\, f^{\mathrm{B}}(1 - f^{\mathrm{B}})(1 - \theta)^2 + \{f^{\mathrm{A}}(1 - f^{\mathrm{A}}) + f^{\mathrm{B}}(1 - f^{\mathrm{B}})\}2\theta(1 - \theta) + 2\, f^{\mathrm{A}}(1 - f^{\mathrm{A}})\theta^2 \\
=& (2\, f^{\mathrm{B}} - 2f^{A^2})(1 - 2\theta + \theta^2) + (f^{\mathrm{A}} - f^{A^2} + f^{\mathrm{B}} - f^{B^2})(2\theta - 2\theta^2) + (2\, f^{\mathrm{A}} - 2f^{A^2})\theta^2 \\
=& 2\, f^{\mathrm{B}} - 2\theta\, f^{\mathrm{B}} - 2f^{B^2} + 2\theta f^{B^2} + 2\theta\, f^{\mathrm{A}} - 2\theta f^{A^2} \\
=& 2\, f^{\mathrm{B}}(1 - \theta) - 2f^{B^2}(1 - \theta) + 2\theta\, f^{\mathrm{A}}(1 - f^{\mathrm{A}}) \\
=& 2\, f^{\mathrm{B}}(1 - f^{\mathrm{B}})(1 - \theta) + 2\theta\, f^{\mathrm{A}}(1 - f^{\mathrm{A}})
\end{aligned}$$

To derive $\mathbb{V}\{\mathbb{E}_\gamma(x|\gamma, \theta)\}$, note that

$$\begin{aligned}
\underset{\gamma}{\mathbb{E}}(x|\gamma, \theta) =& \underset{\gamma}{\mathbb{E}}\{\mathbb{E}(x|\theta)\} \\
=& \mathbb{E}(x|\gamma = 0, \theta)\, \mathbb{P}(\gamma = 0|\theta) + \mathbb{E}(x|\gamma = 1, \theta)\, \mathbb{P}(\gamma = 1|\theta) + \mathbb{E}(x|\gamma = 2, \theta)\, \mathbb{P}(\gamma = 2|\theta) \\
=& 2\theta\, f^{\mathrm{A}} + 2(1 - \theta)\, f^{\mathrm{B}}
\end{aligned}$$

And,

$$
\begin{aligned}
\mathbb{V}\{\underset{\gamma}{\mathbb{E}}(x|\gamma,\theta)\} =& \left[\mathbb{E}(x|\gamma=0,\theta)-\mathbb{E}(x|\theta)\right]^2 \mathbb{P}(\gamma=0|\theta) \\
&+\left[\mathbb{E}(x|\gamma=1,\theta)-\mathbb{E}(x|\theta)\right]^2 \mathbb{P}(\gamma=1|\theta) \\
&+\left[\mathbb{E}(x|\gamma=2,\theta)-\mathbb{E}(x|\theta)\right]^2 \mathbb{P}(\gamma=2|\theta) \\
=&\theta^2\left[2\,f^{\mathrm{A}}-\{2\theta\,f^{\mathrm{A}}+2(1-\theta)\,f^{\mathrm{B}}\}\right]^2 \\
&+2\theta(1-\theta)\left[f^{\mathrm{A}}+f^{\mathrm{B}}-\{2\theta\,f^{\mathrm{A}}+2(1-\theta)\,f^{\mathrm{B}}\}\right]^2 \\
&+(1-\theta)^2\left[2\,f^{\mathrm{B}}-\{2\theta\,f^{\mathrm{A}}+2(1-\theta)\,f^{\mathrm{B}}\}\right]^2 \\
=&2\theta(1-\theta)(f^{\mathrm{A}}-f^{\mathrm{B}})^2
\end{aligned}
$$

Putting this together,

$$
\begin{aligned}
\underset{\theta}{\mathbb{E}}\{\mathbb{V}(x_i|\theta)\} =& \underset{\theta}{\mathbb{E}}\{2\,f^{\mathrm{B}}(1-f^{\mathrm{B}})(1-\theta)+2\theta\,f^{\mathrm{A}}(1-f^{\mathrm{A}})+2\theta(1-\theta)(f^{\mathrm{A}}-f^{\mathrm{B}})^2\} \\
=&2\,f^{\mathrm{B}}(1-f^{\mathrm{B}})\{1-\underset{\theta}{\mathbb{E}}(\theta)\}+2\,\underset{\theta}{\mathbb{E}}(\theta)\,f^{\mathrm{A}}(1-f^{\mathrm{A}})+2\,\underset{\theta}{\mathbb{E}}(\theta-\theta^2)(f^{\mathrm{A}}-f^{\mathrm{B}})^2 \\
=&2\,f^{\mathrm{B}}(1-f^{\mathrm{B}})\{1-\underset{\theta}{\mathbb{E}}(\theta)\}+2\,\underset{\theta}{\mathbb{E}}(\theta)\,f^{\mathrm{A}}(1-f^{\mathrm{A}})+2\{\underset{\theta}{\mathbb{E}}(\theta)-\underset{\theta}{\mathbb{E}}(\theta^2)\}(f^{\mathrm{A}}-f^{\mathrm{B}})^2 \\
=&2\,f^{\mathrm{B}}(1-f^{\mathrm{B}})\{1-\underset{\theta}{\mathbb{E}}(\theta)\}+2\,\underset{\theta}{\mathbb{E}}(\theta)\,f^{\mathrm{A}}(1-f^{\mathrm{A}})+2\{\underset{\theta}{\mathbb{E}}(\theta)-\mathbb{V}(\theta)-\underset{\theta}{\mathbb{E}}(\theta)^2\}(f^{\mathrm{A}}-f^{\mathrm{B}})^2 \\
=&2\,f^{\mathrm{B}}(1-f^{\mathrm{B}})\{1-\underset{\theta}{\mathbb{E}}(\theta)\}+2\,\underset{\theta}{\mathbb{E}}(\theta)\,f^{\mathrm{A}}(1-f^{\mathrm{A}})+2\,\underset{\theta}{\mathbb{E}}(\theta)(1-\underset{\theta}{\mathbb{E}}(\theta))(f^{\mathrm{A}}-f^{\mathrm{B}})^2-2\,\mathbb{V}(\theta)(f^{\mathrm{A}}-f^{\mathrm{B}})^2
\end{aligned}
$$

2.1.2 $\mathbb{V}\{\mathbb{E}_\theta(x_i|\theta)\}$

Recall from the previous section that $\mathbb{E}_\theta(x_i|\theta)=2\theta\,f^{\mathrm{A}}+2(1-\theta)\,f^{\mathrm{B}}$. Then,

$$
\begin{aligned}
\mathbb{V}\{\underset{\theta}{\mathbb{E}}(x_i|\theta)\} =& \mathbb{V}\{2\theta\,f^{\mathrm{A}}+2(1-\theta)\,f^{\mathrm{B}}\} \\
=&4\,\mathbb{V}(\theta)f^{A^2}+4\,\mathbb{V}(1-\theta)f^{B^2}+2\,\mathbb{C}(2\theta\,f^{\mathrm{A}},2(1-\theta\,f^{\mathrm{B}}) \\
=&4\,\mathbb{V}(\theta)f^{A^2}+4\,\mathbb{V}(1-\theta)f^{B^2}-8\,f^{\mathrm{A}}\,f^{\mathrm{B}}\,\mathbb{V}(\theta) \\
=&4\,\mathbb{V}(\theta)(f^{\mathrm{A}}-f^{\mathrm{B}})^2
\end{aligned}
$$

We are now ready to express $\mathbb{V}(x_i)$:

$$
\begin{aligned}
\mathbb{V}(x_i) =& 2\,f^{\mathrm{B}}(1-f^{\mathrm{B}})\{1-\underset{\theta}{\mathbb{E}}(\theta)\}+2\,\underset{\theta}{\mathbb{E}}(\theta)\,f^{\mathrm{A}}(1-f^{\mathrm{A}})+2\,\underset{\theta}{\mathbb{E}}(\theta)(1-\underset{\theta}{\mathbb{E}}(\theta))(f^{\mathrm{A}}-f^{\mathrm{B}})^2 \\
&-2\,\mathbb{V}(\theta)(f^{\mathrm{A}}-f^{\mathrm{B}})^2+4\,\mathbb{V}(\theta)(f^{\mathrm{A}}-f^{\mathrm{B}})^2 \\
=&2\,\underset{\theta}{\mathbb{E}}(\theta)\,f_{\mathrm{i}}^{\mathrm{A}}(1-f_{\mathrm{i}}^{\mathrm{A}})+2\{1-\underset{\theta}{\mathbb{E}}(\theta)\}\,f_{\mathrm{i}}^{\mathrm{B}}(1-f_{\mathrm{i}}^{\mathrm{B}}) \\
&+2\,\underset{\theta}{\mathbb{E}}(\theta)\{1-\underset{\theta}{\mathbb{E}}(\theta)\}(f_{\mathrm{i}}^{\mathrm{A}}-f_{\mathrm{i}}^{\mathrm{B}})^2-2\,\mathbb{V}(\theta)(f_{\mathrm{i}}^{\mathrm{A}}-f_{\mathrm{i}}^{\mathrm{B}})^2
\end{aligned}
$$

2.2 $\mathbb{C}(x_i,x_j)$

We can derive $\mathbb{C}(x_i,x_j)$ using the law of total covariance:

$$
\begin{aligned}
\mathbb{C}(x_i, x_j) =& \underset{\theta}{\mathbb{E}}\{\mathbb{C}(x_i, x_j|\theta)\} + \mathbb{C}\{\underset{\theta}{\mathbb{E}}(x_i|\theta), \underset{\theta}{\mathbb{E}}(x_j|\theta)\} \\
=& 0 + \mathbb{C}\{2\, f_{\mathrm{i}}^{\mathrm{A}}\, \theta + 2\, f_{\mathrm{i}}^{\mathrm{B}}(1-\theta), 2\, f_{\mathrm{j}}^{\mathrm{A}}\, \theta + 2\, f_{\mathrm{j}}^{\mathrm{B}}(1-\theta)\} \\
=& \mathbb{C}(2\, f_{\mathrm{i}}^{\mathrm{A}}\, \theta, 2\, f_{\mathrm{j}}^{\mathrm{A}}\, \theta) + \mathbb{C}(2\, f_{\mathrm{i}}^{\mathrm{A}}\, \theta, 2\, f_{\mathrm{j}}^{\mathrm{B}}(1-\theta) + \\
& \mathbb{C}(2\, f_{\mathrm{i}}^{\mathrm{B}}(1-\theta), 2\, f_{\mathrm{j}}^{\mathrm{A}}\, \theta) + \mathbb{C}(2\, f_{\mathrm{i}}^{\mathrm{B}}(1-\theta), 2\, f_{\mathrm{j}}^{\mathrm{B}}(1-\theta)) \\
=& 4\, \mathbb{V}(\theta)(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}})
\end{aligned}
$$

$\mathbb{E}_\theta\{\mathbb{C}(x_i, x_j|\theta)\} = 0$ because we assume that the loci are unlinked and therefore, x_i and x_j are conditionally independent. Putting this all together, we get the genetic variance in admixed populations as presented in the main text:

$$
\begin{aligned}
V_g =& \sum_{i=1}^{m} \beta_i^2\, \mathbb{V}(x_i) + \sum_{j \neq i} \beta_i \beta_j\, \mathbb{C}(x_i, x_j) \\
=& \sum_{i=1}^{m} \beta_i^2 2\, \underset{\theta}{\mathbb{E}}(\theta)\, f_{\mathrm{i}}^{\mathrm{A}}(1 - f_{\mathrm{i}}^{\mathrm{A}}) + \sum_{i=1}^{m} \beta_i^2 2\{1 - \underset{\theta}{\mathbb{E}}(\theta)\}\, f_{\mathrm{i}}^{\mathrm{B}}(1 - f_{\mathrm{i}}^{\mathrm{B}}) \\
&+ \sum_{i=1}^{m} \beta_i^2 2\, \underset{\theta}{\mathbb{E}}(\theta)\{1 - \underset{\theta}{\mathbb{E}}(\theta)\}(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2 + \\
&+ \sum_{i=1}^{m} \beta_i^2 2\, \mathbb{V}(\theta)(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2] \\
&+ \sum_{j \neq i} \beta_i \beta_j 4\, \mathbb{V}(\theta)(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}})
\end{aligned}
$$

The only difference being that in the main text we use $\mathbb{E}$ instead of $\mathbb{E}_\theta$ for simplicity. With two 'unadmixed' source populations with equal number of individuals, $\mathbb{E}(\theta) = 0.5$ and $\mathbb{V}(\theta) = \mathbb{E}(\theta)\{1 - \mathbb{E}(\theta)\} = 0.25$ and V_g reduces to:

$$
\begin{aligned}
V_g =& \mathbb{V}\left(\sum_{i=1}^{m} \beta_i x_i\right) = \sum_{i=1}^{m} \beta_i^2\, \mathbb{V}(x_i) + \sum_{j \neq i} \beta_i \beta_j\, \mathbb{C}(x_i, x_j) \\
=& \sum_{i=1}^{m} \beta_i^2 \left[f_{\mathrm{i}}^{\mathrm{A}}(1 - f_{\mathrm{i}}^{\mathrm{A}}) + f_{\mathrm{i}}^{\mathrm{B}}(1 - f_{\mathrm{i}}^{\mathrm{B}})\right] \\
&+ \sum_{i=1}^{m} \beta_i^2 (f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2 \\
&+ \sum_{i \neq j} \beta_i \beta_j (f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}})
\end{aligned}
$$

3 The effect of genotype scaling on $\hat{\sigma}_u^2$

In the main text we showed that when we scale genotypes by $\sqrt{2f_i(1 - f_i)}$ in calculating the GRM, we get biased GREML estimates even if the effects are uncorrelated. We can fix this issue by scaling with the square root of sample variance, $\sqrt{\mathbb{V}(x_i)}$. We provide an explanation of this behavior using the Haseman-Elston (HE) regression estimator, which is asymptotically equivalent to the GREML estimator

if effects are uncorrelated [68] but which, unlike GREML, has a closed-form solution. We assume random effects corresponding to the unscaled genotypes are $\beta_i \sim \mathcal{N}(0, \frac{\sigma_u^2}{2mf_i(1-f_i)})$.

$$V_g = \mathbb{V}\left(\sum_{i=1}^{m} \beta_i x_i\right) = \sum_{i=1}^{m} \beta_i^2 \, \mathbb{V}(x_i)$$

In expectation,

$$\begin{aligned}
\mathbb{V}(x_i) =& 2\,\mathbb{E}(\theta)\, f_{\mathrm{i}}^{\mathrm{A}}(1-f_{\mathrm{i}}^{\mathrm{A}}) + 2\{1-\mathbb{E}(\theta)\}\, f_{\mathrm{i}}^{\mathrm{B}}(1-f_{\mathrm{i}}^{\mathrm{B}}) + \\
& 2\,\mathbb{E}(\theta)\{1-\mathbb{E}(\theta)\}(f_{\mathrm{i}}^{\mathrm{A}}-f_{\mathrm{i}}^{\mathrm{B}})^2 + 2\,\mathbb{V}(\theta)(f_{\mathrm{i}}^{\mathrm{A}}-f_{\mathrm{i}}^{\mathrm{B}})^2 \\
=& 2f_i(1-f_i) + 2\,\mathbb{V}(\theta)(f_i^A - f_i^B)^2
\end{aligned}$$

And

$$\begin{aligned}
V_g =& \sum_{i=1}^{m} \beta_i^2 \{2f_i(1-f_i) + 2\,\mathbb{V}(\theta)(f_i^A - f_i^B)^2\} \\
=& \sum_{i=1}^{m} \frac{\sigma_u^2}{2mf_i(1-f_i)} \{2f_i(1-f_i) + 2\,\mathbb{V}(\theta)(f_i^A - f_i^B)^2\} \\
=& \frac{\sigma_u^2}{m} \sum_{i=1}^{m} \{1 + \mathbb{V}(\theta)\frac{(f_{\mathrm{i}}^{\mathrm{A}}-f_{\mathrm{i}}^{\mathrm{B}})^2}{f_i(1-f_i)}\} \\
=& \sigma_u^2 + \underbrace{\mathbb{V}(\theta)\frac{\sigma_u^2}{m}\sum_{i=1}^{m}\frac{(f_{\mathrm{i}}^{\mathrm{A}}-f_{\mathrm{i}}^{\mathrm{B}})^2}{f_i(1-f_i)}}_{\substack{\text{contribution of population structure} \\ \text{to the genic variance}}}
\end{aligned}$$

Thus, the genetic variance can be decomposed into two terms, one that depends on the degree of population structure and one that does not.

The HE estimator of V_g is based on the regression of products of (centered) phenotypes $y_k y_l$ for all pairs of individuals $k \neq l$ on the corresponding entries of the GRM ($\boldsymbol{\psi}$) where $\psi_{kl} = \frac{\sum_{i=1}^{m} z_{ik} z_{il}}{m}$ and z_{ik} is the centered and scaled genotype of individual k for locus i:

$$\begin{aligned}
\hat{V}_g =& \frac{\mathbb{C}(y_k y_l, \psi_{kl})}{\mathbb{V}(\psi_{kl})} \\
=& \frac{\mathbb{E}(y_k y_l \psi_{kl}) - \mathbb{E}(y_k y_l)\,\mathbb{E}(\psi_{kl})}{\mathbb{E}(\psi_{kl}^2) - \mathbb{E}(\psi_{kl})^2} \\
=& \frac{\mathbb{E}(y_k y_l \psi_{kl})}{\mathbb{E}(\psi_{kl}^2)}
\end{aligned}$$

With the standard scaling, $z_{ik} = \frac{x_{ik}-2f_i}{\sqrt{2f_i(1-f_i)}}$ and the corresponding effects are $u_i \sim \mathcal{N}(0, \sigma_u^2)$. In this case, the HE estimator is:

$$
\begin{aligned}
\hat{V}_g =& \frac{\mathbb{E}(y_k y_l \psi_{kl})}{\mathbb{E}(\psi_{kl}^2)} = \frac{\mathbb{E}\left(\sum_{i=1}^m u_i z_{ik} \sum_{i=1}^m u_i z_{il} \psi_{kl}\right)}{\mathbb{E}(\psi_{kl}^2)} \\
=& \frac{\mathbb{E}\left(\sum_{i=1}^m u_i^2 z_{ik} z_{il} \psi_{kl}\right)}{\mathbb{E}(\psi_{kl}^2)} \\
=& \frac{\mathbb{E}(u_i^2)\,\mathbb{E}\left(\sum_{i=1}^m z_{ik} z_{il} \psi_{kl}\right)}{\mathbb{E}(\psi_{kl}^2)} = \frac{\frac{\sigma_u^2}{m}\,\mathbb{E}(m\psi_{kl}^2)}{\mathbb{E}(\psi_{kl}^2)} \\
=& \frac{\sigma_u^2\,\mathbb{E}(\psi_{kl}^2)}{\mathbb{E}(\psi_{kl}^2)} = \sigma_u^2
\end{aligned}
$$

Which is not equal to V_g because it does not capture the contribution of population structure. Next, we consider the case where the genotypes are standardized instead by the sample variance, i.e., $z_{kl} = \frac{x_{ik} - 2f_i}{\sqrt{\mathbb{V}(x_i)}}$ such that $\mathbb{V}(z_i) = 1$. We can derive $\mathbb{E}(u_i^2)$ corresponding to this scaling by noting that the genetic variance remains the same, i.e., the genetic variance is invariant under linear transformations of the genotype [14]:

$$
\begin{aligned}
\sum_{i=1}^m \beta_i^2\,\mathbb{V}(x_i) &= \sum_{i=1}^m u_i^2\,\mathbb{V}(z_i) \\
m\,\mathbb{E}(u^2) &= \sigma_u^2 + \mathbb{V}(\theta)\frac{\sigma_u^2}{m}\sum_{i=1}^m \frac{(f_\mathrm{i}^\mathrm{A} - f_\mathrm{i}^\mathrm{B})^2}{f_i(1-f_i)} \\
\mathbb{E}(u^2) &= \frac{\sigma_u^2}{m} + \mathbb{V}(\theta)\frac{\sigma_u^2}{m^2}\sum_{i=1}^m \frac{(f_\mathrm{i}^\mathrm{A} - f_\mathrm{i}^\mathrm{B})^2}{f_i(1-f_i)}
\end{aligned}
$$

Note that because the genotypes are scaled properly, $\mathbb{E}(\psi_{kl}) = 0$ and $\mathbb{V}(\psi_{kl}) = 1$. Then, the HE estimator becomes:

$$
\begin{aligned}
\hat{V}_g =& \frac{\mathbb{E}(y_k y_l \psi_{kl})}{\mathbb{E}(\psi_{kl}^2)} \\
=& \frac{\mathbb{E}(u_i^2)\,\mathbb{E}\left(\sum_{i=1}^m z_{ik} z_{il}\right)}{\mathbb{E}(\psi_{kl}^2)} \\
=& \left(\frac{\sigma_u^2}{m} + \mathbb{V}(\theta)\frac{\sigma_u^2}{m^2}\sum_{i=1}^m \frac{(f_\mathrm{i}^\mathrm{A} - f_\mathrm{i}^\mathrm{B})^2}{f_i(1-f_i)}\right)\frac{m\,\mathbb{E}(\psi_{kl}^2)}{\mathbb{E}(\psi_{kl}^2)} \\
=& \sigma_u^2 + \mathbb{V}(\theta)\frac{\sigma_u^2}{m}\sum_{i=1}^m \frac{(f_\mathrm{i}^\mathrm{A} - f_\mathrm{i}^\mathrm{B})^2}{f_i(1-f_i)}
\end{aligned}
$$

Which provides an unbiased estimate of the genic variance.

4 Genetic variance after correction for individual ancestry

In the main text, we stated that including individual ancestry as a fixed effect in GREML can lead to an underestimate of V_g in the presence of population structure. Mixed effect models deal with fixed effects (ancestry in our case) by projecting them out of the phenotypes, and estimating the residual variance. This is conceptually equivalent to measuring the residual variance of the regression between phenotype and ancestry. As a result, any variance in the phenotype that is explained by ancestry is removed. To understand this quantitatively, it is helpful to decompose V_g into components of variance explained by and variance orthogonal to ancestry:

$$\mathbb{V}(g) = \underbrace{\mathbb{V}\{\mathop{\mathbb{E}}_{\theta}(g|\theta)\}}_{\substack{\text{variance along}\\ \text{ancestry axis}}} + \underbrace{\mathop{\mathbb{E}}_{\theta}\{\mathbb{V}(g|\theta)\}}_{\substack{\text{variance orthogonal}\\ \text{to ancestry axis}}}$$

We can express the residual variance as:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\theta}\{\mathbb{V}(g|\theta)\} =& \mathop{\mathbb{E}}_{\theta}\{\mathbb{V}(\sum_{i=1}^{M} \beta_i^2 x_i|\theta)\} \\
=& \mathop{\mathbb{E}}_{\theta}\{\sum_{i=1}^{M} \beta_i^2\, \mathbb{V}(x_i|\theta)\} + \mathop{\mathbb{E}}_{\theta}\{\sum_{i\neq j} \beta_i\beta_j\, \mathbb{C}(x_i, x_j|\theta) \\
=& \sum_{i=1}^{M} \beta_i^2\, \mathop{\mathbb{E}}_{\theta}\{\mathbb{V}(x_i|\theta)\} + 0 \\
=& 2\, \mathop{\mathbb{E}}_{\theta}(\theta) \sum_{i=1}^{M} \beta^2\, f_{\mathrm{i}}^{\mathrm{A}}(1 - f_{\mathrm{i}}^{\mathrm{A}}) + 2\{1 - \mathop{\mathbb{E}}_{\theta}(\theta)\} \sum_{i=1}^{M} \beta^2\, f_{\mathrm{i}}^{\mathrm{B}}(1 - f_{\mathrm{i}}^{\mathrm{B}}) \\
&+ 2\, \mathop{\mathbb{E}}_{\theta}(\theta) \sum_{i=1}^{M} \beta^2 \{1 - \mathop{\mathbb{E}}_{\theta}(\theta)\}(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2 - 2\, \mathbb{V}(\theta) \sum_{i=1}^{M} \beta^2 (f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})^2
\end{aligned}$$

Note, that this represents the following components of V_g from the main text: (1.1) + (1.2) - (1.3). Note, that (1.3), which is subtracted out, is always positive and depends on $\mathbb{V}(\theta)$. Thus, the residual genetic variance will be underestimated, regardless of trait architecture, in the presence of population structure, i.e. when $\mathbb{V}(\theta) > 0$.

5 Effect size of local ancestry

We define local ancestry $\gamma_i \in \{0, 1, 2\}$ as the number of alleles at locus i that trace their ancestry to population A. Thus, the local ancestry at locus i in individual k is a Binomial random variable with $\mathbb{E}(\gamma_{i,k}) = 2\theta_k$. We define the ancestry value of an individual as the weighted sum of their local ancestry: $\sum_{i=1}^{m} \phi_i \gamma_i$ where $\phi_i = \beta_i(f_{\mathrm{i}}^{\mathrm{B}} - f_{\mathrm{i}}^{\mathrm{A}})$.

To show this, note that $\phi = \mathbb{E}(y|\gamma = 1) - \mathbb{E}(y|\gamma = 0)$ where $\mathbb{E}(y|\gamma = 1) = \int_{-\infty}^{\infty} y h(y|\gamma = 1)$ and h is a density function. Our goal is to express ϕ in terms of β, which is equal to $\mathbb{E}(y|x = 1) - \mathbb{E}(y|x = 0)$. Furthermore, $\mathbb{E}(y|x = 1) = \int_{-\infty}^{\infty} y h(y|x = 1)$. We can express $h(y|\gamma)$ in terms of $h(y|x)$ as follows:

$$\begin{aligned}
h(y|\gamma = 1) =\ & h(y|x = 0)\, \mathbb{P}(x = 0|\gamma = 1) + h(y|x = 1)\, \mathbb{P}(x = 1|\gamma = 1) + h(y|x = 2)\, \mathbb{P}(x = 2|\gamma = 1) \\
=\ & h(y|x = 0) 2(1 - f^{\mathrm{A}})(1 - f^{\mathrm{A}}) + h(y|x = 1)\{f^{\mathrm{A}}(1 - f^{\mathrm{B}}) + f^{\mathrm{B}}(1 - f^{\mathrm{A}})\} + h(y|x = 2) 2\, f^{\mathrm{A}}\, f^{\mathrm{B}}
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}(y|\gamma=1) = &\int_{-\infty}^{\infty} y h(y|\gamma=1) dy \\
=&(1-f^{\mathrm{A}})(1-f^{\mathrm{B}}) \int_{-\infty}^{\infty} y h(y|x=0) dy \\
&+ \{f^{\mathrm{A}}(1-f^{\mathrm{B}}) + f^{\mathrm{B}}(1-f^{\mathrm{A}})\} \int_{-\infty}^{\infty} y h(y|x=1) dy \\
&+ f^{\mathrm{A}} f^{\mathrm{B}} \int_{-\infty}^{\infty} y h(y|x=2) dy \\
=&(1-f^{\mathrm{A}})(1-f^{\mathrm{B}})\, \mathbb{E}(y|x=0) + \{f^{\mathrm{A}}(1-f^{\mathrm{B}}) + f^{\mathrm{B}}(1-f^{\mathrm{A}})\}\, \mathbb{E}(y|x=1) + f^{\mathrm{A}} f^{\mathrm{B}}\, \mathbb{E}(y|x=2) \\
=&0 + \{f^{\mathrm{A}}(1-f^{\mathrm{B}}) + f^{\mathrm{B}}(1-f^{\mathrm{A}})\}\beta + f^{\mathrm{A}} f^{\mathrm{B}}\, 2\beta \\
=&\beta\, f^{\mathrm{A}} + \beta\, f^{\mathrm{B}}
\end{aligned}
$$

Similary, $\mathbb{E}(y|\gamma=0) = 2\beta\, f^{\mathrm{B}}$ and $\phi = \mathbb{E}(y|\gamma=1) - \mathbb{E}(y|\gamma=0) = \beta(f^{\mathrm{B}} - f^{\mathrm{A}})$

6 Genetic variance due to local ancestry

$$
\begin{aligned}
V_\gamma = &\ \mathbb{V}(\sum_{i=1}^{m} \phi_i \gamma_i) \\
= &\sum_{i=1}^{m} \phi_i^2\, \mathbb{V}(\gamma_i) + \sum_{i=1}^{m} \sum_{j \neq i} \phi_i \phi_j\, \mathbb{C}(\gamma_i, \gamma_j)
\end{aligned} \tag{1}
$$

We use the law of total variance and covariance to derive $\mathbb{V}(\gamma_i)$ and $\mathbb{C}(\gamma_i, \gamma_j)$:

$$
\begin{aligned}
\mathbb{V}(\gamma_i) =& \mathbb{E}\{\mathbb{V}(\gamma_i|\theta)\} + \mathbb{V}\{\mathbb{E}(\gamma_i|\theta)\} \\
=& \mathbb{E}\{2\theta(1-\theta)\} + \mathbb{V}(2\theta) \\
=&2\,\mathbb{E}(\theta) - 2\,\mathbb{E}(\theta^2) + 4\,\mathbb{V}(\theta) \\
=&2\,\mathbb{E}(\theta) - 2\,\mathbb{V}(\theta) - 2\,\mathbb{E}(\theta)^2 + 4\,\mathbb{V}(\theta) \\
=&2\,\mathbb{E}(\theta)\{1-\mathbb{E}(\theta)\} + 2\,\mathbb{V}(\theta)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{C}(\gamma_i, \gamma_j) =& \mathbb{E}\{\mathbb{C}(\gamma_i, \gamma_j|\theta)\} + \mathbb{C}\{\mathbb{E}(\gamma_i, \gamma_j|\theta)\} \\
=&0 + \mathbb{C}(2\theta, 2\theta) = 4\,\mathbb{V}(\theta)
\end{aligned}
$$

$$
\begin{aligned}
V_\gamma =&2\,\mathbb{E}(\theta)\{1-\mathbb{E}(\theta)\} \sum_{i=1}^{m} \phi_i^2 + 2\,\mathbb{V}(\theta) \sum_{i=1}^{m} \phi_i^2 + 4\,\mathbb{V}(\theta) \sum_{i=1}^{m} \sum_{j \neq i} \phi_i \phi_j \\
=&2\,\mathbb{E}(\theta)\{1-\mathbb{E}(\theta)\} \sum_{i=1}^{m} \beta_i^2 (f_{\mathrm{i}}^{\mathrm{B}} - f_{\mathrm{i}}^{\mathrm{A}})^2 \\
&+ 2\,\mathbb{V}(\theta) \sum_{i=1}^{m} \beta_i^2 (f_{\mathrm{i}}^{\mathrm{B}} - f_{\mathrm{i}}^{\mathrm{A}})^2 \\
&+ 4\,\mathbb{V}(\theta) \sum_{i=1}^{m} \sum_{j \neq i} \beta_i \beta_j (f_{\mathrm{i}}^{\mathrm{B}} - f_{\mathrm{i}}^{\mathrm{A}})(f_{\mathrm{j}}^{\mathrm{B}} - f_{\mathrm{j}}^{\mathrm{A}})
\end{aligned}
$$

We note, that

Supplement

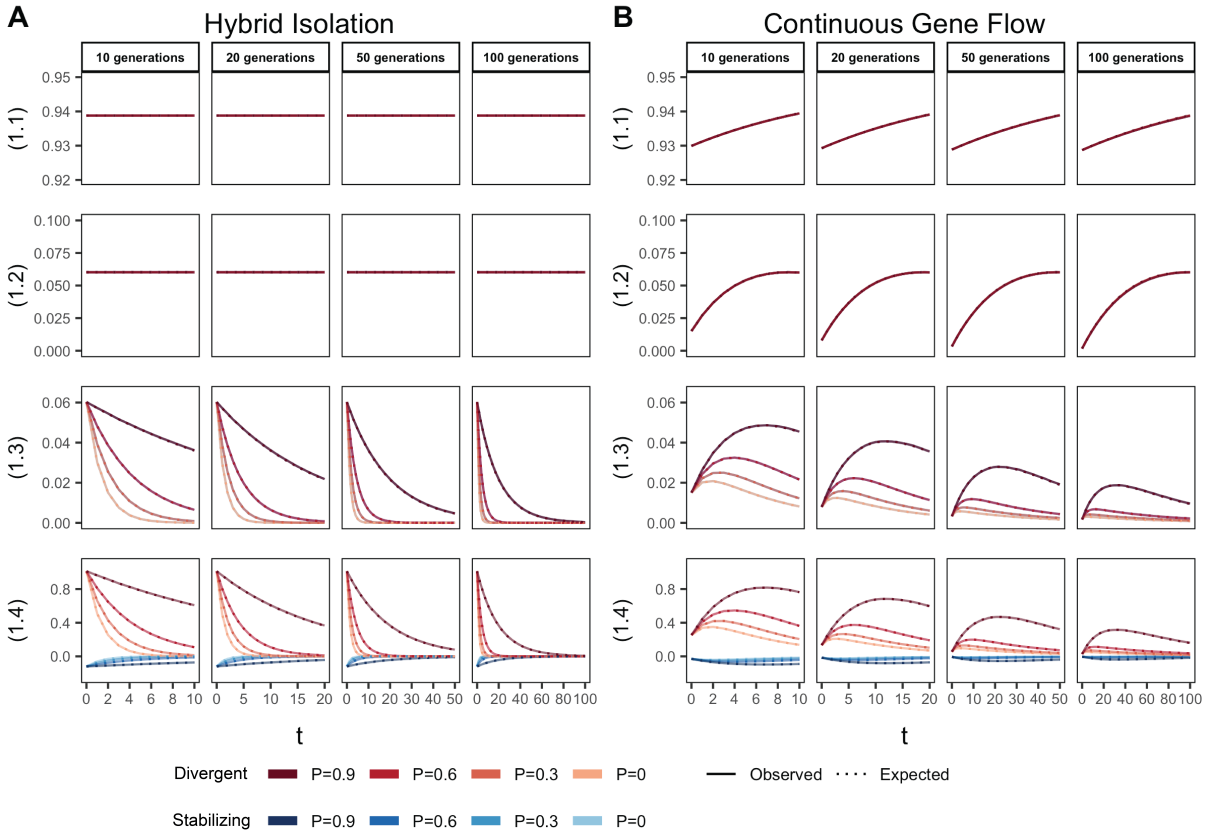

Figure S1: The behavior of the four components of genetic variance in admixed populations under the (A) HI and (B) CGF models. We assume that the mean ancestry proportion in the population is 0.5. The solid lines represent values observed in simulations averaged across ten replicates and the dotted lines represent the expected values based on Eq. 1 of the main text. The red and blue lines represent values for traits 1 and 2, respectively. P indicates the strength of assortative mating. P=0.6 is missing for simulations run for 50 and 100 generations and $\theta \in \{0.1, 0.2\}$ due to the difficulty in finding mate pairs (Methods).

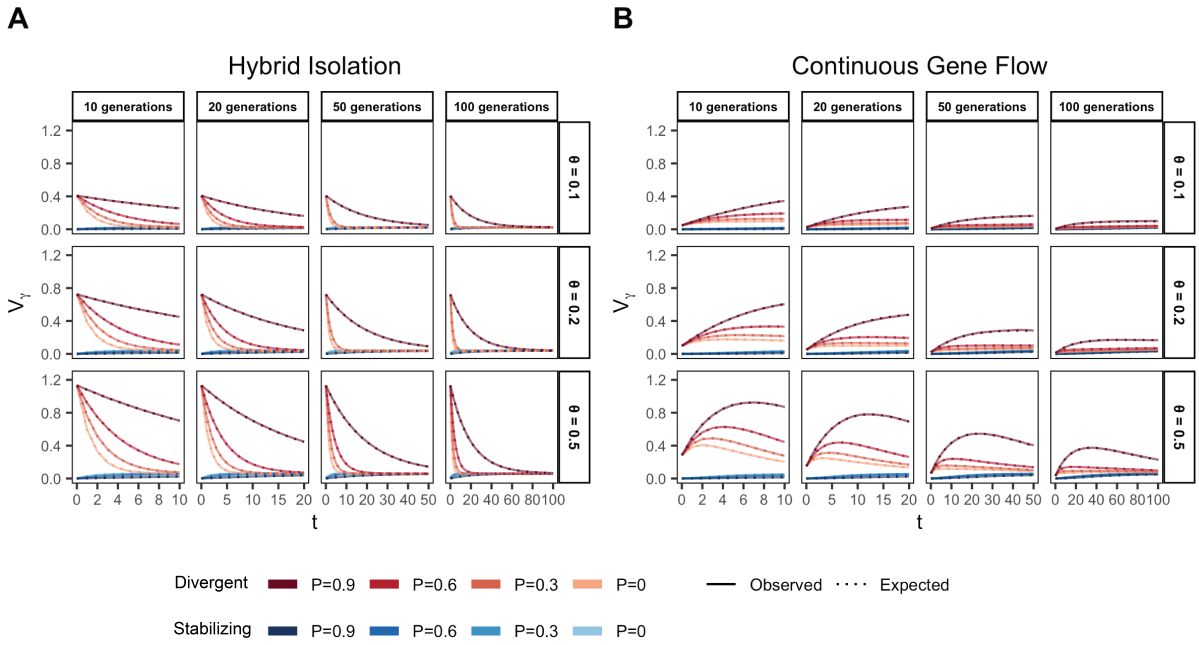

Figure S2: The behavior of the genetic variance due to local ancestry in admixed populations under the (A) HI and (B) CGF models. The solid lines represent values observed in simulations averaged across ten replicates and the dotted lines represent the expected values based on Eq. 1 of the main text. The red and blue lines represent values for traits 1 and 2, respectively. P indicates the strength of assortative mating. P=0.6 is missing for simulations run for 50 and 100 generations and $\theta \in \{0.1, 0.2\}$ due to the difficulty in finding mate pairs (Methods).

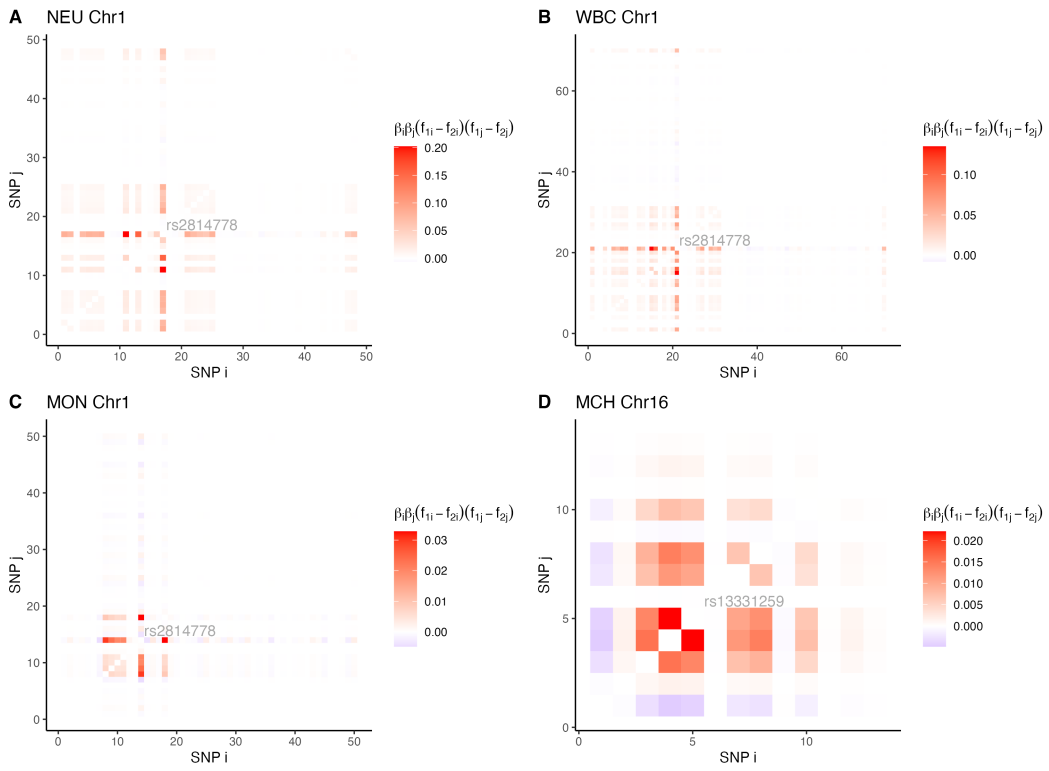

Figure S3: The LD contribution to the variance explained by variant pairs for (A) neutrophil counts (NEU), (B) white blood count (WBC), (C) monocyte count (MON), and (D) mean corpuscular hemoglobin (MCH). Only chromosomes where we suspected there was a disproportionate contribution to the variance explained are shown.

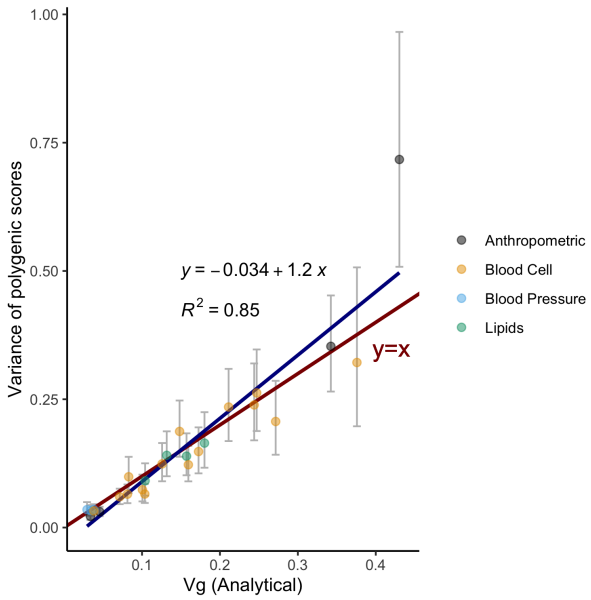

Figure S4: Expected variance explained estimated using Equation 1 (x-axis) vs the observed variance in polygenic scores (y-axis) in ASW are strongly correlated. Confidence intervals were generated by non-parametric bootstrap.