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)

August 4, 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 estimates of h^2_{snp} are not guaranteed to be equal to h^2 in admixed populations, even in the absence of confounding and even if the causal variants are known. We interpret this discrepancy arising not because the estimate is biased, but because the estimand itself as defined under the random effects model may not be equal to h^2. The model assumes that SNP effects are uncorrelated which may not be true, even for unlinked loci in admixed and structured populations, leading to over- or under-estimates of h^2_{snp} relative to h^2. For the same reason, local ancestry heritability (h^2_γ) may also not be equal to the variance explained by local ancestry in admixed populations. We describe the quantitative behavior of h^2_{snp} and h^2_γ 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 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 [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 also [14, 15]). 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. But 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 [16–22]. 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 h^2_{snp} in admixed populations in relation to h^2. Should we expect the two to be equal in the ideal situation where causal variants are known? If not, how should we interpret h^2_{snp}? 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 test if h^2_{snp} estimated with genome-wide restricted maximum likelihood (GREML) [3, 5] – arguably still the most widely used method – is equal to h^2 and explore this equivalence as a function of admixture history and genetic architecture. We show that GREML-based $\hat{h}^2_{snp}$ is not guaranteed to be equal to h^2 in admixed and other structured populations, even in the absence of confounding and when all causal variants are known. We explain this discrepancy in terms of the generative model underlying GREML, which assumes that (i) the effect of causal variants are random and uncorrelated, and (ii) the population is mating randomly. As a result, the GREML estimand (h^2_{snp}) by definition equals h^2 only if these conditions are met. We discuss the implications of this 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 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 first three terms represent the sum of 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 in the admixed populations is a function of the mean, $\mathbb{E}(\theta)$, and variance, $\mathbb{V}(\theta)$, of individual ancestry. 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. 2A). 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. 2A). For brevity, we refer to these as the hybrid-isolation (HI) and continuous gene flow (CGF) models, respectively, following Pfaff *et al.* (2001) [23]. $\mathbb{V}(\theta)$ is also affected by assortative mating based on ancestry (hereafter referred to as assortative mating for bervity) and we model this following Zaitlen *et al.* (2017) using a parameter $P \in (0, 1)$, which represents the correlation in the ancestry of individuals in a mating pair [24].

Under these conditions, the behavior of $\mathbb{E}(\theta)$ and $\mathbb{V}(\theta)$ has been described previously [24, 25] (Fig. 1A and B). 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 all individuals either carry chromosomes 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. 1). 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.

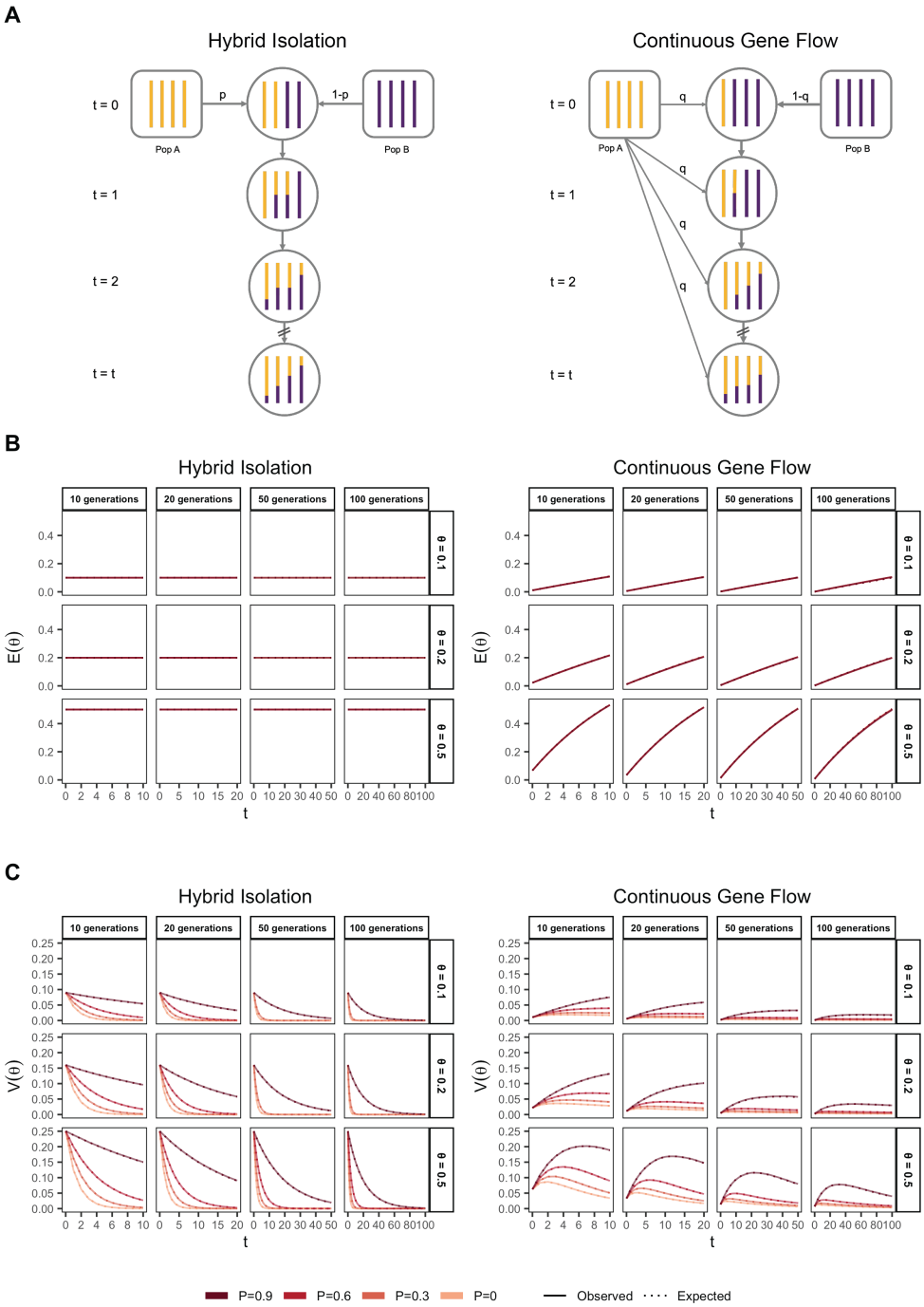

Figure 1: The behavior of mean and variance of individual ancestry as a function of admixture history. (A) Shows the demographic models under which simulations were carried out. Admixture might occur once (Hybrid Isolation, HI, left column) or continuously (Continuous Gene Flow, CGF, right column). (B) 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 $t = 0$, 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 $\mathbb{V}(\theta)$.

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 $V_{gb} = \frac{(\bar{g_1}-\bar{g_2})^2}{2}$ is positive only if there is a difference in the mean genotypic values (Fig. 2B). This variance increases with the degree of genetic drift 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 to increase monotonically with increasing genetic drift, $(f_{\mathrm{i}}^{\mathrm{A}} - f_{\mathrm{i}}^{\mathrm{B}})(f_{\mathrm{j}}^{\mathrm{A}} - f_{\mathrm{j}}^{\mathrm{B}})$ 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.

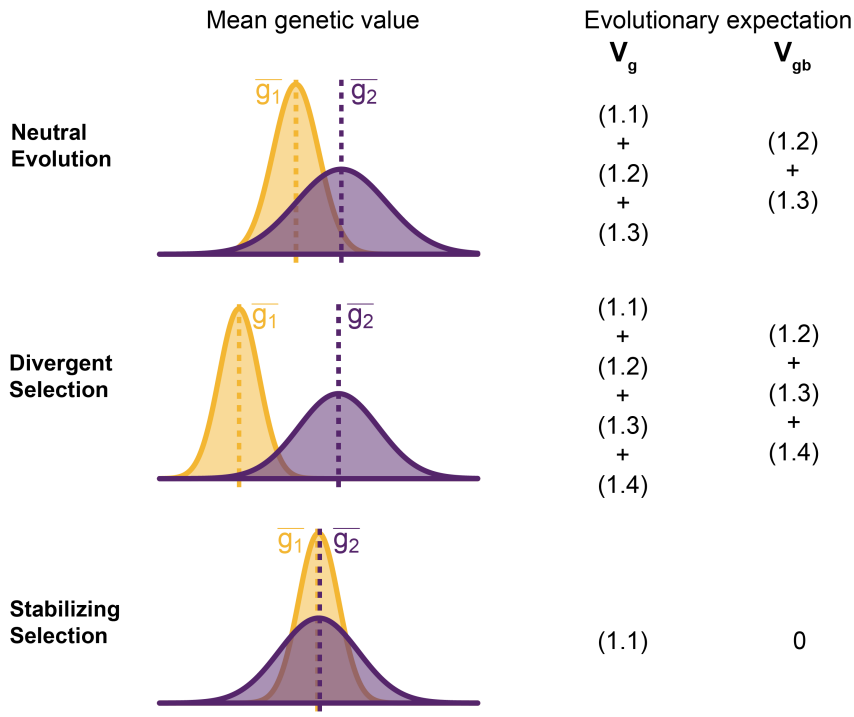

Figure 2: Decomposing genetic variance in a two-population system. The plot illustrates the expected distribution of genetic values in two populations under different selective pressures and the terms on the right list the total (V_g) and between-population genetic variance (V_{gb}) expected over the evolutionary process. For neutrally evolving traits (top row), we expect there to be an absolute difference in the mean genetic values ($|\bar{g_1} - \bar{g_2}|$) that is proportional to F_{ST}. For traits under divergent selection (middle), $|\bar{g_1} - \bar{g_2}|$ is expected to be greater than that expected under genetic drift. For traits under stabilizing selection, $|\bar{g_1} - \bar{g_2}|$ 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, increasing the LD contribution, i.e., term (1.4). Stabilizing selection, on the other hand, induces negative LD, reducing (1.4) [26, 27]. 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 [28]. We drew their effects (β) from $\mathcal{N}(0, \frac{1}{2\bar{f}(1-\bar{f})})$ where $\bar{f}$ is the mean allele frequency between the two populations. 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. 1) 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.

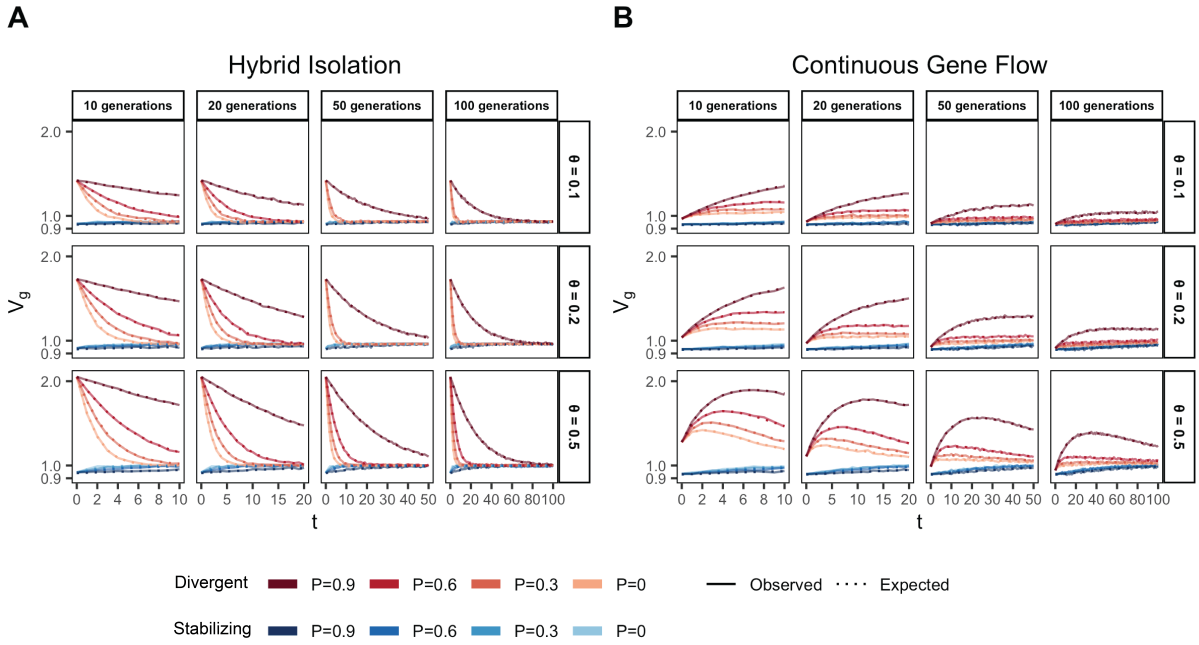

Figure 3: Genetic variance in admixed populations under the (A) HI and (B) CGF models. Solid lines represent the expected genetic variance based on Eq. (1) and dashed lines represent results of simulations averaged over ten replicates. Red and blue lines represent traits under divergent and stabilizing selection, respectively.

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 discrepancy between h^2_{snp} and h^2, 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 allele frequency. The SNP effects are assumed to be random and independent such that $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{I}\frac{\sigma^2_u}{m})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2_\epsilon)$ is random environmental error. Then, the phenotypic variance can be decomposed as:

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

where $\frac{\boldsymbol{Z}\boldsymbol{Z}'}{m}$ is the genetic relationship matrix (GRM), the variance components $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_\epsilon$ are estimated using restricted maximum likelihood, and $\hat{h}^2_{snp}$ is calculated as $\frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_u+\hat{\sigma}^2_\epsilon}$. Really, the key estimate is $\hat{\sigma}^2_u$ since $\hat{\sigma}^2_\epsilon = 1 - \hat{\sigma}^2_u$, and we are interested in asking whether $\hat{\sigma}^2_u$ is equal to V_g. To answer this, we constructed the GRM with causal variants and estimated $\hat{\sigma}^2_u$ using GCTA, including individual ancestry in the model as a fixed effect to correct for any confounding due to genetic stratification [3, 4].

We show that GCTA under- and over-estimates the genetic variance in admixed populations for traits under divergent (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, which, 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, $\hat{\sigma}^2_u$ does not capture the LD contribution, i.e. term (1.4) (Fig. S3A). In our simulations, the LD across loci is entirely due to population structure since they are unlinked. Thus, the LD contribution, and therefore, the bias in $\hat{\sigma}^2_u$ is larger in the presence of population structure.

But $\hat{\sigma}^2_u$ can be biased, even if the effects are uncorrelated and the LD contribution is zero. It is standard in GREML to scale genotypes with $\sqrt{2f_i(1-f_i)}$ 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 $\mathbb{V}(x_i) = 2f_i(1-f_i) + 2\,\mathbb{V}(\theta)(f^A_i - f^B_i)^2$, where f_i, f^A_i, and f^B_i correspond to frequency in the admixed population, and source populations, A and B, respectively. We show that this assumption biases $\hat{\sigma}^2_u$ downwards by a factor of $2\,\mathbb{V}(\theta)(f^A_i - f^B_i)^2$ – term (1.3) (Fig. S3B, Appendix). We confirm this by showing that we can fix this bias if we scale genotypes by their sample variance, i.e., $\sqrt{\mathbb{V}(x_i)}$ (Fig. S3C) (Appendix). Thus, with the standard scaling, $\hat{\sigma}^2_u$ is not even equal to the genic variance in the presence of population structure.

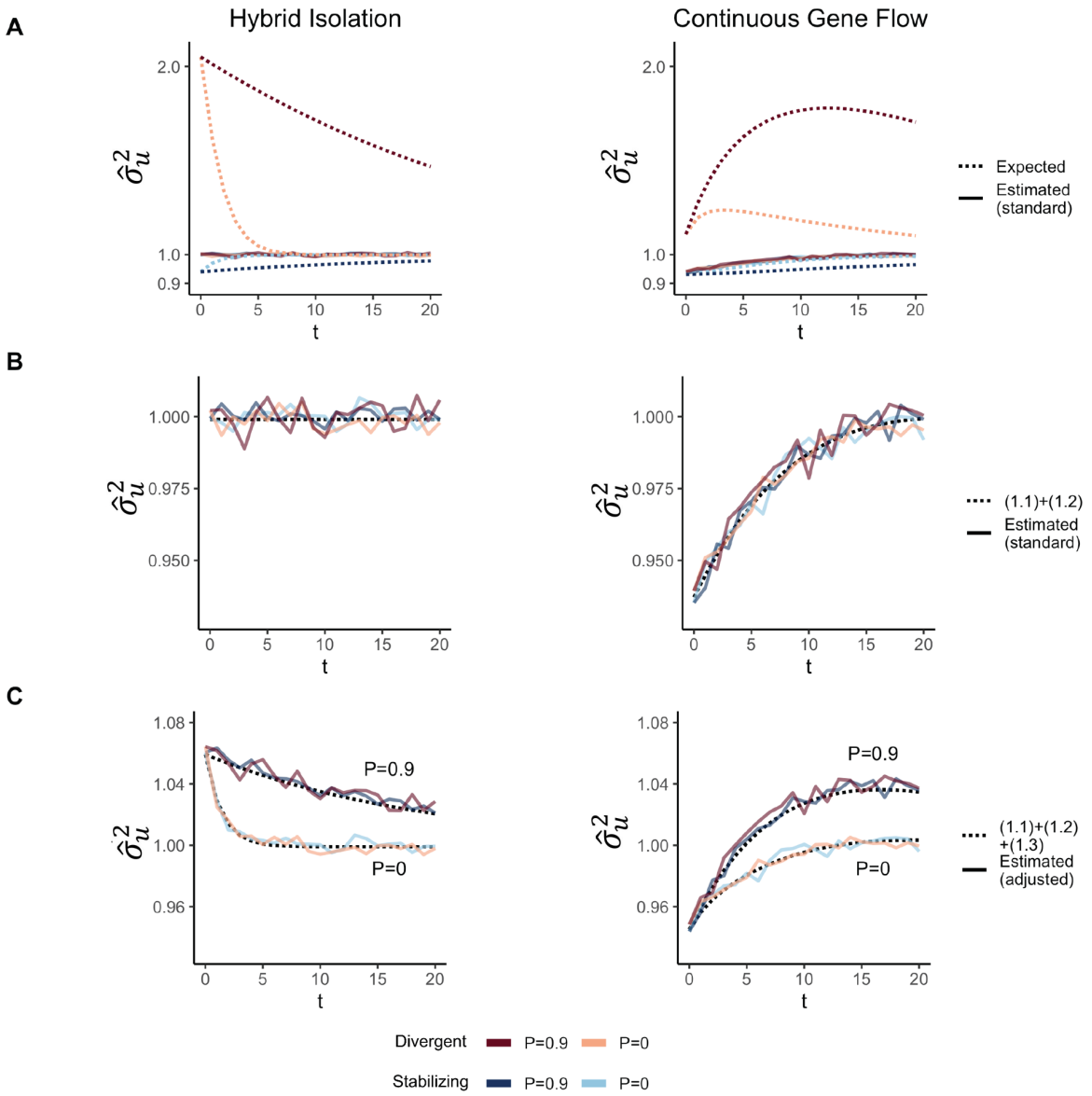

Figure 4: The behavior of GREML estimates of the genetic variance ($\hat{\sigma}_u^2$) using global ancestry as a covariate in admixed populations under the HI (left column) and CGF (right column) models. The solid lines represent values observed in simulations averaged across ten replicates and the dotted lines represent the expected values based on Eq. 1. Red and blue lines represent values for traits under divergent and stabilizing selection, respectively. P indicates the strength of assortative mating. (A) and (B) shows the behavior of $\hat{\sigma}_u^2$ using the standard $\sqrt{2f(1-f)}$ genotype scaling. In (C), we show $\hat{\sigma}_u^2$ when genotypes are scaled by the square root of their sample variance, i.e., $\sqrt{\mathbb{V}(x)}$.

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. If there is no more gene flow, $\mathbb{V}(\theta)$ will ultimately go to zero and V_g will converge towards $\hat{\sigma}_u^2$. But note that even though $\hat{\sigma}_u^2$ may be biased relative to V_g, it is not biased relative to the estimand under the model assumed by GREML (σ_u^2), which is more accurately interpreted as the genetic variance expected if there were no correlation in effect sizes and if the population were mating randomly. In other words, $\mathbb{E}(\hat{\sigma}_u^2) = \sigma_u^2 = (1.1) + (1.2) \neq V_g$ (Fig. S3B).

Local ancestry heritability

A related quantity of interest in admixed populations is local ancestry heritability (h_γ^2), which is defined as the proportion of phenotypic variance that can be explained by local ancestry. Zaitlen *et al.* (2014) [29] showed that this quantity is related to, and can be used to estimate, h_{snp}^2 in admixed populations. The advantage of this 'indirect' 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 and structural variants [29]. Here, we show that under the random effects model, h_γ^2 may not be equal to the local ancestry ancestry heritability in the population because of the same reasons that h_{snp}^2 is not equal to the heritability.

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 $\mathbb{E}(\gamma_{ik}) = 2\theta_k$. Similar to the genetic value of an individual, we define 'ancestry value' 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:

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

and heritability explained by local ancestry is simply the ratio of V_γ and the phenotypic variance. Note that $V_\gamma = (1.2) + (1.3) + (1.4)$ – the genetic variance between the source populations – 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).

GREML estimation of h_γ^2 is similar to the estimation of h_{snp}^2, the key difference being that the former involves constructing the GRM using local ancestry instead of genotypes [29]. 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, $\boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{I}\frac{\sigma_u^2}{m})$ are local

ancestry effects, and $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}\sigma_\xi^2)$. 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 here that this equivalence is not guaranteed in the presence of population structure and/or correlated effects.

We calculated the GRM from local ancestry at causal variants with our simulated data, and estimated σ_v^2 with individual ancestry as a fixed effect to correct for genetic stratification [29]. 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 (Fig. 5A). In addition, if local ancestry is scaled by its expectation under random mating rather than the square root of the sample variance, $\hat{\sigma}_v^2$ will be underestimated even if the effects are uncorrelated (Fig. 5B). The overall bias is equal to the terms weighted by $\mathbb{V}(\theta)$ – (1.3) + (1.4). As a result, $\hat{h}_\gamma^2$ is not guaranteed to be equal to the heritability explained by local ancestry, even in the absence of confounding and even if local ancestry at causal variants is known without error. This is not because of an inherent bias in the estimation procedure since $\mathbb{E}(\hat{\sigma}_v^2) = \sigma_v^2 =$ (1.2), but because the estimand itself defined under the random-effects model does not capture the variance due to LD and population structure.

We note that if individual ancestry is not included as a covariate, $\hat{\sigma}_v^2$ tends to be biased even relative to σ_v^2 due to confounding effects of genetic stratification (Fig. S4). We did not see the same level of confounding when individual ancestry was not included in the model estimating σ_u^2 (Fig. S3). We do not fully understand the reason for this but we think it might be because genetic stratification leads to more inflation in the effect size of local ancestry compared to the effect size of genotype.

How much does population structure contribute in practice?

In the previous sections, we showed theoretically that h_{snp}^2 is not equal to h^2 in admixed populations even if the causal variants are known. 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 in African Americans? To answer this, we used independent genome-wide significant SNPs for 26 quantitative traits from the GWAS catalog [30]. We calculated the total genetic variance explained and decomposed it into the four components in Equation 1 using allele frequencies (f^A and f^B) from the 1000 Genomes YRI and CEU [31], and the mean ($\mathbb{E}(\theta) \approx 0.77$) and variance ($\mathbb{V}(\theta) \approx 0.02$) of individual ancestry from the 1000 Genomes ASW (Methods).

We show that for skin pigmentation – a trait under strong divergent selection – the LD contribution, i.e. term (1.4), is positive and accounts for ≈ 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 due to strong selection favoring alleles for darker pigmentation in regions with high UV exposure and vice versa [32–35]. 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).

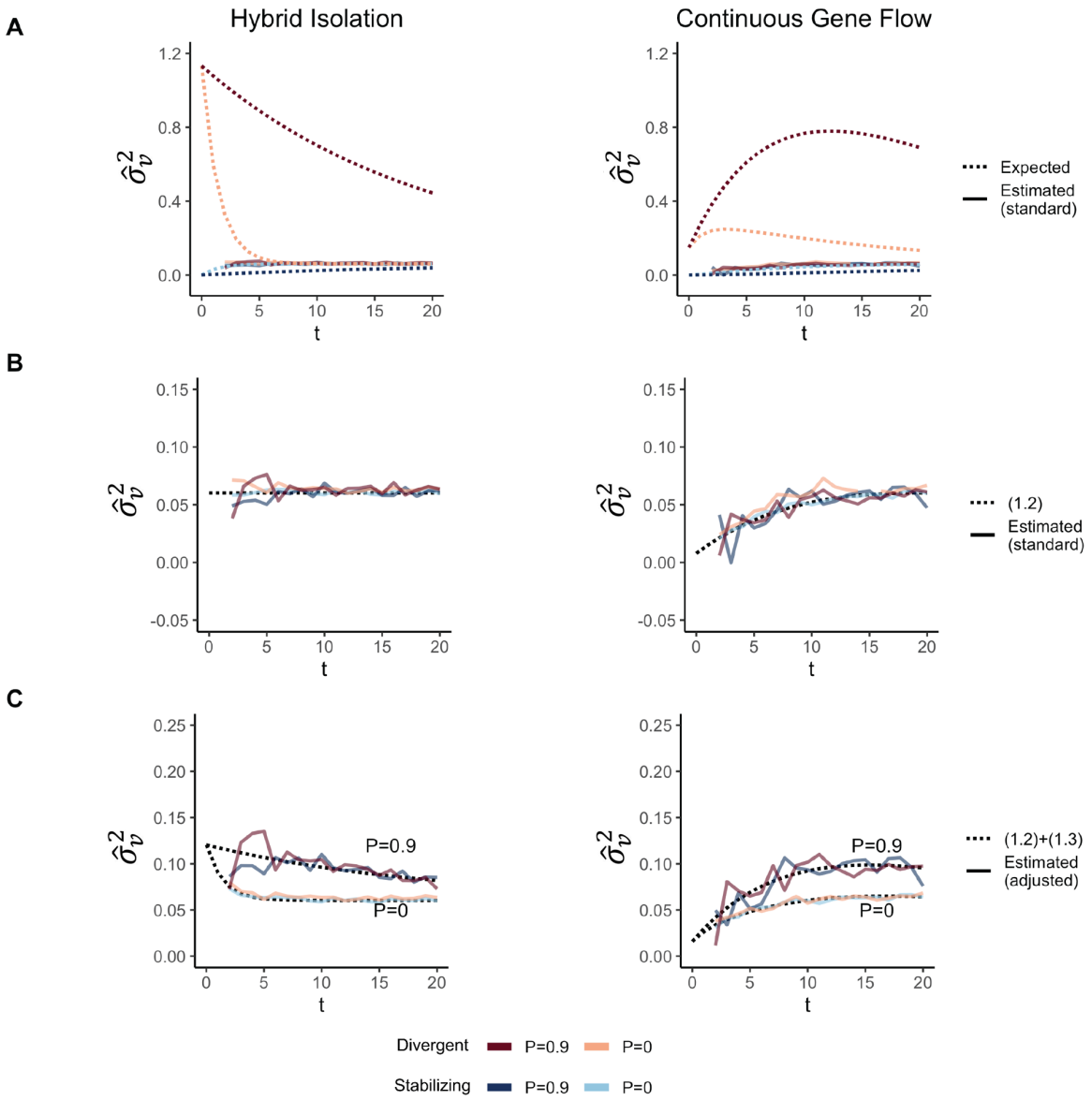

Figure 5: The behavior of GREML estimates of the variance due to local ancestry ($\hat{\sigma}_v^2$) using global ancestry as a covariate in admixed populations under the HI (left column) and CGF (right column) models. The solid lines represent values observed in simulations averaged across ten replicates and the dotted lines represent the expected values based on Eq. 1. Red and blue lines represent values for Traits 1 and 2, respectively. P indicates the strength of assortative mating. (A) and (B) shows the behavior of $\hat{\sigma}_v^2$ when the default scaling of local ancestry is used. In (C), we show $\hat{\sigma}_v^2$ when local ancestry is scaled with square root of the sampling variance, i.e., $\sqrt{\mathbb{V}(\gamma_i)}$.

Because we selected independent associations for this exercise (Methods), the LD contribution is driven entirely due to population structure among African Americans. 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 [33, 34, 36–38]. Overall, this suggests that population structure contributes relatively little, as least to the variance explained by GWAS SNPs.

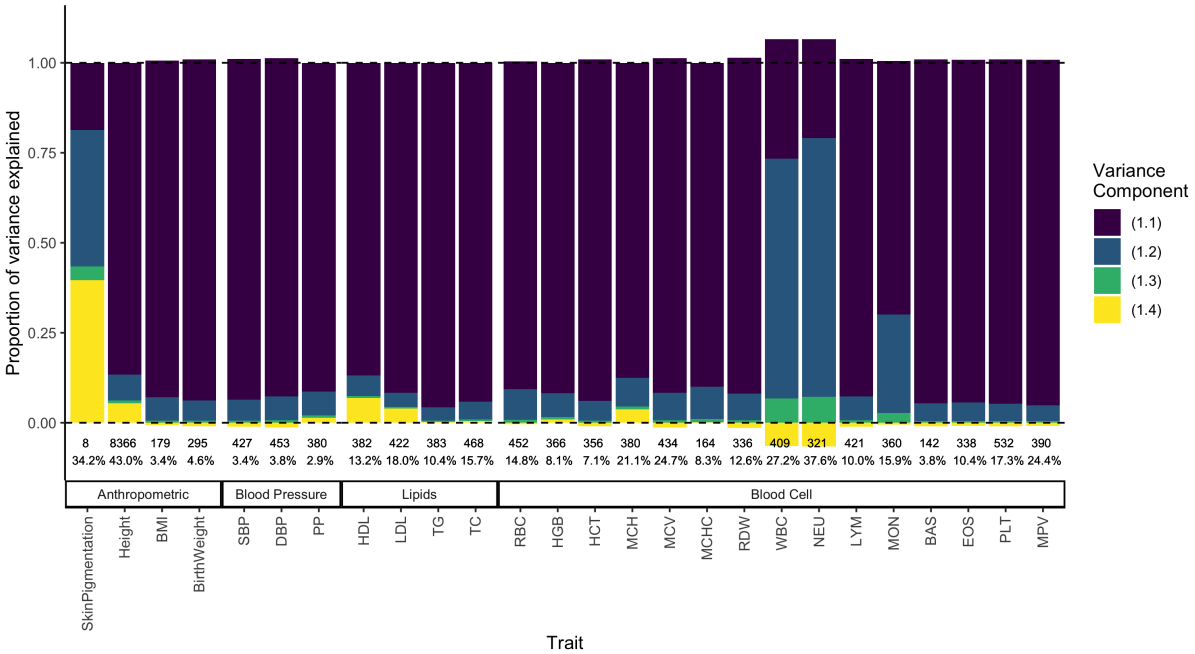

Figure 6: Decomposing the genetic variance explained by GWAS SNPs in African Americans. We calculated the four variance components listed in Equation 1, their values shown on the y-axis as a fraction of the total variance explained (shown as percentage at the bottom). The number of variants used to calculate 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 [30], 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 h^2_{snp}, its interpretation remains unclear, particularly its equivalence to heritability in the presence of population structure. It is generally accepted that $\hat{h}^2_{snp}$ – the estimator – can be biased in structured populations [4, 7, 9–11, 39]. Here, we show how $\hat{h}^2_{snp}$ may not be equal to h^2 in admixed populations even in the absence of confounding and even if causal variants are known.

GREML assumes that SNP effects are random and independent – an assumption that may not be true, especially in the presence of admixture and population structure, which create LD across unlinked loci. This LD contributes to the genetic variance and can persist, despite recombination, for a number of generations due to continued gene flow and/or assortative mating. Because the LD contribution can be positive or negative, $\hat{h}^2_{snp}$ can under- or over-estimate h^2. But $\hat{h}^2_{snp}$ can be biased even when effects are uncorrelated 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. For these reasons, there is no guarantee that $\hat{h}^2_{snp}$ that will be equal to h^2, even if the causal variants are known. But technically this is not because the estimate is biased, but because the estimand itself as defined under the random effects model is not equal to the heritability [14, 15]. h^2_{snp}, assuming the genotypes are scaled properly, is better interpreted as the proportion of phenotypic variance explained by the *genic* variance. We show that h^2_γ as defined under random effects models [29, 40] should be interpreted similarly.

Does the LD contribution to the genetic variance have practical implications? The answer to this depends on how one intends to use SNP heritability. h^2_{snp} can be useful in qauntifying the power to detect variants in GWAS where the quantity of interest is the genic variance. But if one is interested in using h^2_{snp} 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 – then h^2_{snp} is technically not the relevant quantity.

Ultimately, the discrepancy between 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 h^2_{snp} and h^2 should be fairly modest. But this assumption is obviously unrealistic given that GWAS SNPs are common variants that in most cases cumulatively explain a small fraction of trait heritability. We know that rare variants contribute disproportionately to the genic variance because of negative selection [41]. What is their LD contribution? This will become clearer in the near future with the discovery of rare variants through large sequence-based studies [42]. While these are underway, theoretical studies are needed to understand how different selection regimes influence the LD patterns between causal variants – an important aspect of the genetic architecture of complex traits.

One limitation of this research is that we studied only the GREML estimator of h^2_{snp} because of its widespread use. There are many estimators, which can be broadly grouped into random- and fixed effect estimators based on how they treat SNP effects [43]. 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 [44]. This is one reason why random effect estimators, such as GREML, are popular – because they reduce the dimensionality by assuming that the effects are drawn from some distribution where the variance is the only parameter of interest. Fixed effects estimators should be able to capture the LD

contribution, in principle, but this is not obvious in practice since the simulations used to evaluate the accuracy of such estimators still assume uncorrelated effects [43–45]. 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.

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 [28], 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 [46]. 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) [24]. 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.2C). 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 [24]. 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 genotype x_i from a binomial distribution conditioning on local ancestry. Thus, $x_{im} \sim Bin(1, f_{\mathrm{i}}^{\mathrm{A}})$ if $\gamma_{im} = 1$ and $x_{im} \sim Bin(1, f_{\mathrm{i}}^{\mathrm{B}})$ if $\gamma_{im} = 0$ and similarly for x_{if}. 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 GCTA to estimate σ_u^2 and σ_v^2 using the *--reml* and *--reml-no-constrain* flags. We could not do this without any error in the genetic values so we simulated individual phenotypes with a heritability of 0.8 by adding random noise $e \sim \mathcal{N}(0, V_g/4)$ to the genetic value. We used GCTA [5] to construct the standard GRM with the *--make-grm* flag. We also constructed an 'adjusted' GRM, where instead of standardizing the genotype of the ith SNP with $\sqrt{2f_i(1-f_i)}$, we used $\sqrt{\mathbb{V}(x_i)}$. Similarly, the 'adjusted' local ancestry GRM was constructed by scaling local ancestry with $\sqrt{\mathbb{V}(\gamma_i)}$. For both σ_u^2 and σ_v^2, we included individual ancestry as a covariate to correct for any confounding due to genetic stratification.

Estimating variance explained by GWAS SNPs

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

We retrieved the summary statistics of 26 traits from GWAS catalog [30]. Full list of traits and the source papers [47–55] are listed in Table S1. To maximize the number of variants discovered, we chose summary statistics from studies that were conducted in both European and multi-ancestry samples and that reported the following information: effect allele, effect size, p-value, and genomic position. For birth weight, we downloaded the data from the Early Growth Genetics (EGG) consortium website [52] since the version reported on the GWAS catalog is incomplete. For skin pigmentation, we chose summary statistics from the UKB [56] released by the Neale Lab (http://www.nealelab.is/uk-biobank) and processed by Ju and Mathieson [47] to represent effect sizes estimated among individuals of European ancestry. We also selected summary statistics from Lona-Durazo *et al.* (2019) where effect sizes were meta-analyzed across four admixed cohorts [48]. Lona-Durazo *et al.* provide summary statistics separately with and without conditioning on rs1426654 and rs35397 – two large effect variants in *SCL24A5* and *SLC45A2*. We used the 'conditioned' effect sizes and added in the effects of rs1426654 and rs35397 to estimate genetic variance.

We selected independent hits for each trait by pruning and thresholding with PLINK v1.90b6.21 [57] in two steps as in Ju *et al.* (2020) [47]. We used the genotype data of GBR from the 1000 genome project [31] as the LD reference panel. We kept only SNPs (indels were removed) that passed the

genome-wide significant threshold (*--clump-p1 5e-8*) with a pairwise LD cutoff of 0.05 (*--clump-r2 0.05*) and a physical distance threshold of 250Kb (*--clump-kb 250*) for clumping. Second, we applied a second round of clumping (*--clump-kb 100*) to remove SNPs within 100kb.

When GWAS was carried out separately in different ancestry cohorts in the same study, we used inverse-variance weighting to meta-analyze effect sizes for variants that were genome-wide significant (p-value $< 5 \times 10^{-8}$) in at least one cohort. This allowed us to maximize the discovery of variants such as the Duffy null allele that are absent among individuals of European ancestry but polymorphic in other populations [38].

We used allele frequencies from the 1000 Genomes CEU and YRI to represent the allele frequencies of GWAS SNPs in Europeans and Africans, respectively, making sure that the alleles reported in the summary statistics matched the alleles reported in the 1000 Genomes. We estimated the global ancestry of ASW individuals (N = 74) with CEU and YRI individuals from 1000 genome (phase 3) using ADMIXTURE 1.3.0 [58] with k=2 and used it to calculate the mean (proportion of African ancestry = 0.767) and variance (0.018) of global ancestry in ASW. With the effect sizes, allele frequencies, and the mean and variance in ancestry, we calculated the four components of genetic variance using Equation 1 and expressed them as a fraction of the total genetic variance.

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. S5). 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. S5). 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-sum* flag in PLINK, [57] in ASW individuals. We compared the first estimate of the genetic variance to the second (Fig. S6) 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. S6).

Code availability

We carried out all analyses in R version 4.2.3 [59], PLINK v1.90b6.21 and PLINK 2.0 [57, 60], 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 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. Wojcik, G. L., Graff, M., *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570,** 514–518 (7762 2019).

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

18. 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).

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

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

21. 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).

22. 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).

23. 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).

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

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

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

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

28. 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).

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

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

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

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

33. 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).

34. 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).

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. 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).

37. 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).

38. 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).

39. 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).

40. 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).

41. Schoech, A. P., Jordan, D. M., *et al.* Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nature Communications* **10,** 790 (1 2019).

42. 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).

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

44. 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).

45. 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).

46. 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).

47. 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).

48. 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).

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

50. 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).

51. 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).

52. 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).

53. 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).

54. 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).

55. 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).

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

57. 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).

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

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

60. 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).

61. 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

Variance in ancestry 24

Genetic variance 24

The effect of genotype scaling on $\hat{sigma}_u^2$ 26

Effect size of local ancestry 28

Genetic variance due to local ancestry 29

Variance in ancestry

We denote variance and covariance with $\mathbb{V}(.)$ and $\mathbb{C}(.)$ and used the expressions in [24] 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 $\mathbb{V}(\theta_t) = \mathbb{V}(\theta_{t-1})\frac{(1+P_{t-1})}{2}$. $P = Cor(\theta_m, \theta_f)$ measures the strength of assortative mating, i.e, the correlation between the ancestry in a mating pair (θ_m, θ_f). Since our notation slightly differs from [24], 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,

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

Genetic variance

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:

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

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

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

where θ represents global ancestry. Let $f_{\mathrm{i}}^{\mathrm{A}}$ and $f_{\mathrm{i}}^{\mathrm{B}}$ represent the allele frequency at the ith locus. Then,

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

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

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

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

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

$\mathbb{E}\{\mathbb{C}(x_i, x_j|\theta)\} = 0$ 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:

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

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:

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

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, $\sqrt{\mathbb{V}(x_i)}$ (Fig. S3). 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 [61] 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 $\alpha_i \sim \mathcal{N}(0, \sigma_u^2)$. In this case, the HE estimator is:

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

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 $\mathbb{E}(\alpha_i^2)$ corresponding to this scaling by noting that the genetic variance remains the same:

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

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

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

Which provides an unbiased estimate of the genic variance.

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

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 $\mathbb{C}(\gamma_i, \gamma_j)$:

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

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

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

Supplement

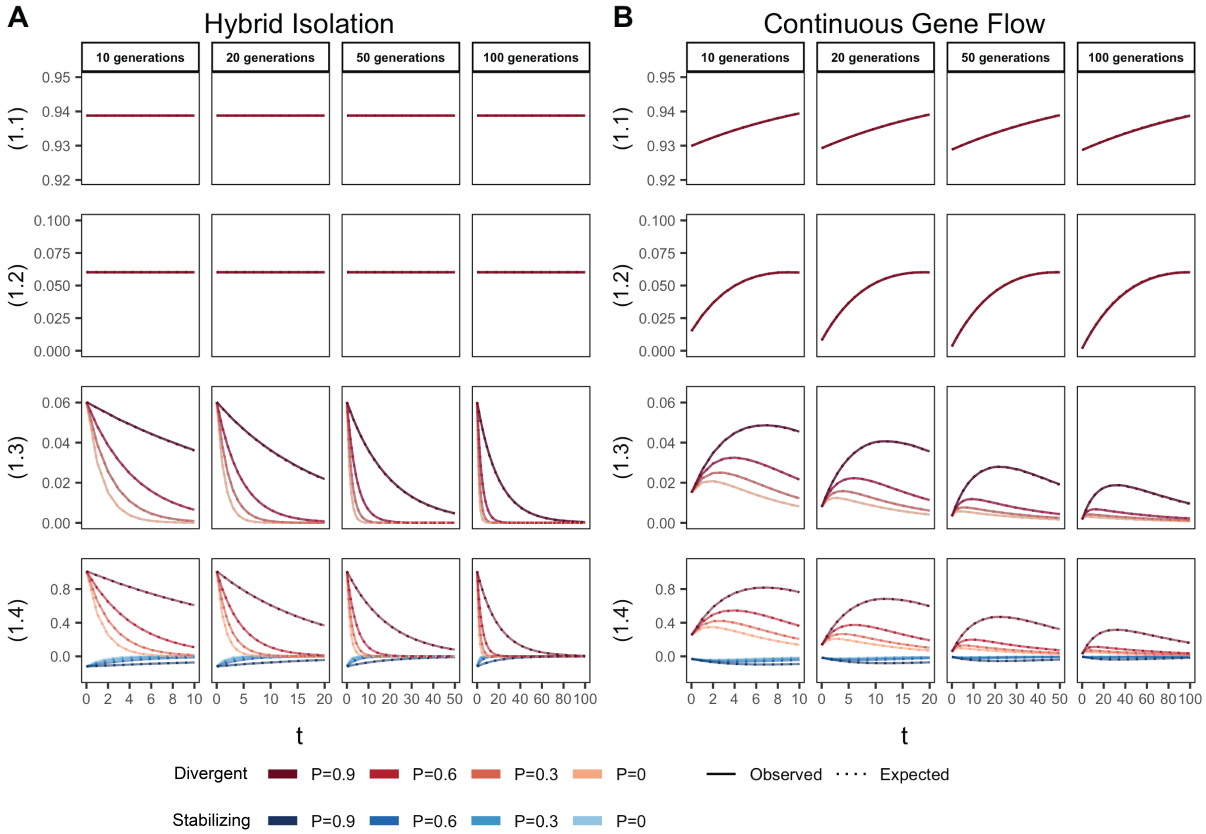

Figure S1: The behavior of the four components of genetic variance in admixed populations under the (A) HI and (B) CGF models. We assume that the mean ancestry proportion in the population is 0.5. The solid lines represent values observed in simulations averaged across ten replicates and the dotted lines 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.

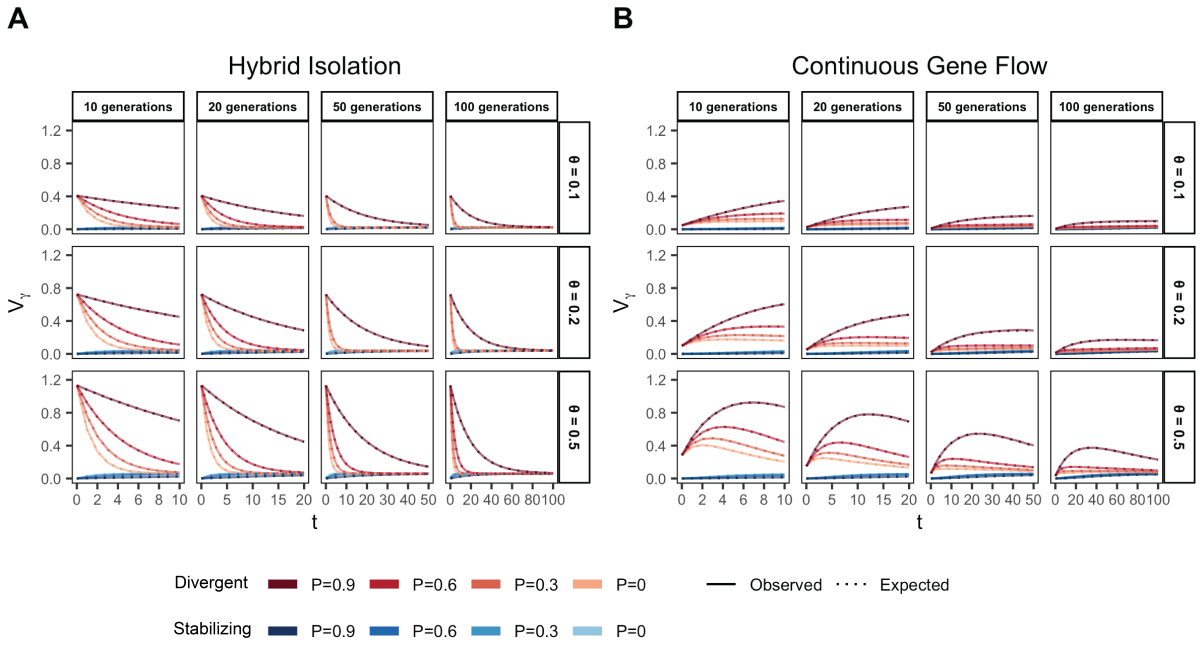

Figure S2: The behavior of the genetic variance due to local ancestry in admixed populations under the (A) HI and (B) CGF models. The solid lines represent values observed in simulations averaged across ten replicates and the dotted lines represent the expected values based on Eq. 1 of the main text. The red and blue lines represent values for traits 1 and 2, respectively. P indicates the strength of assortative mating.

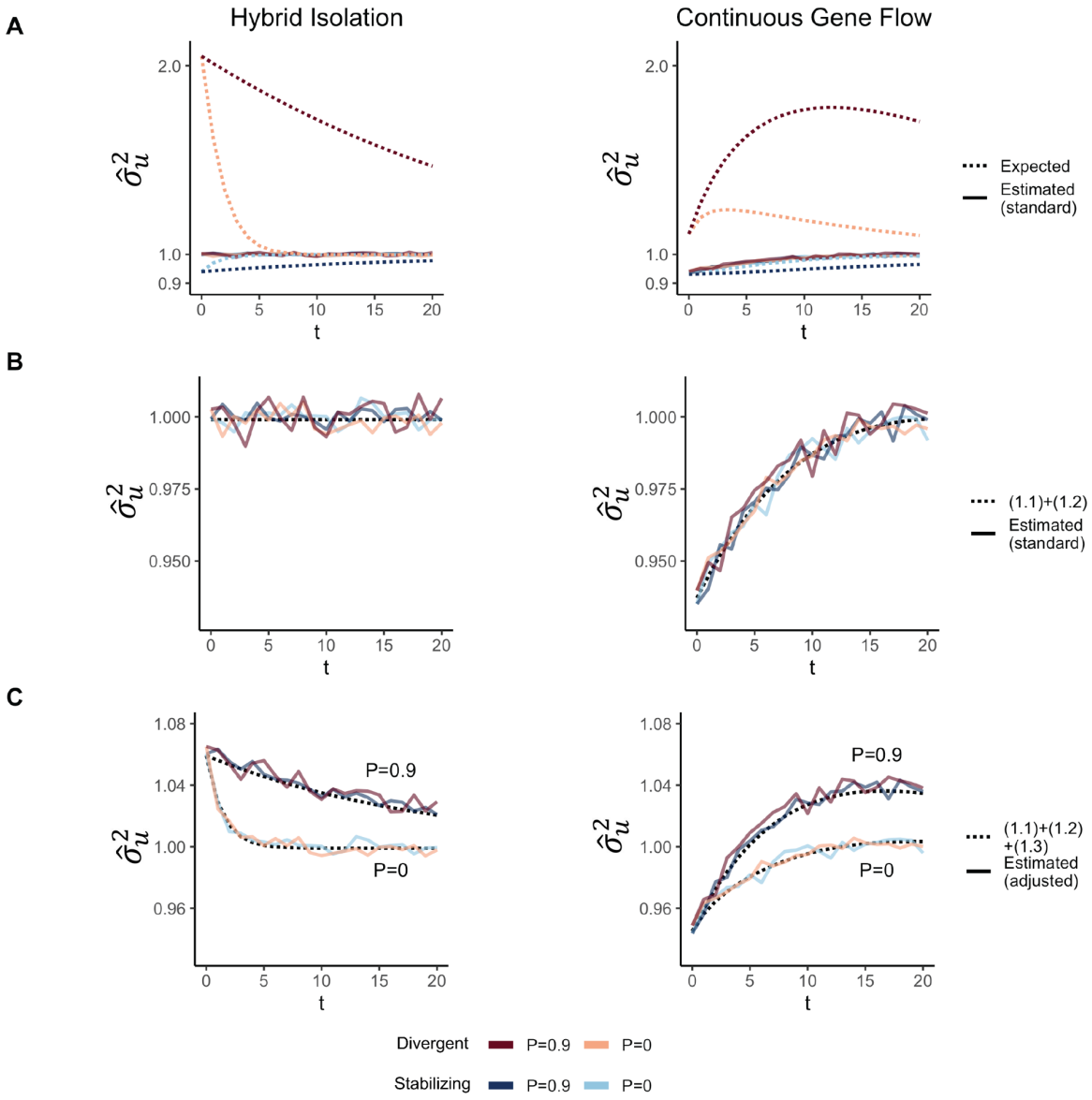

Figure S3: The behavior of the genetic variance estimated with GREML ($\hat{sigma}_u^2$) in admixed populations when individual ancestry is not included as a fixed effect in the model. (A) The expected (dotted lines) and estimated (solid lines) genetic variance (y-axes) in admixed populations under the HI (left column) and CGF (right columns) models for Trait 1 (red) and 2 (blue) shown as a function of time (x-axes). Estimates represent mean across ten replicates. (B) GCTA estimates based on the standard GRM are equal to the sum of terms (1.1) and (1.2), i.e. the genetic variance expected at equlibrium if the population were randomly mating. (C) We can recover the bias in estimates of the genic contribution if we scale the genotypes by $\sqrt{\mathbb{V}(x_i)}$ instead of $\sqrt{2f_i(1-f_i)}$

.

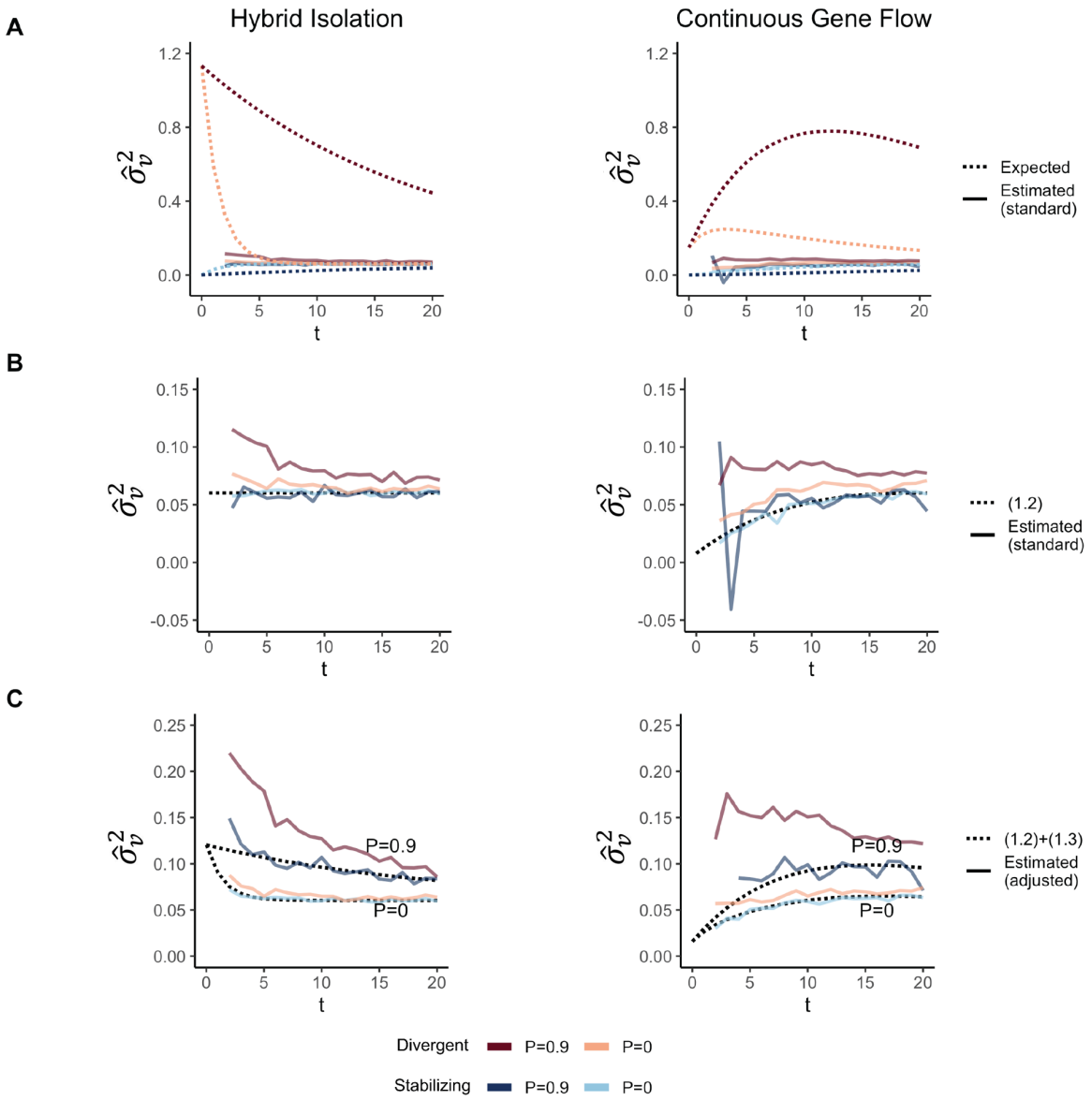

Figure S4: The behavior of the genetic variance due to local ancestry estimated with GREML ($\hat{sigma}_v^2$) when individual ancestry is not included as a fixed effect in the model. The solid lines represent values estimated from simulations averaged across ten replicates and the dotted lines represent 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.

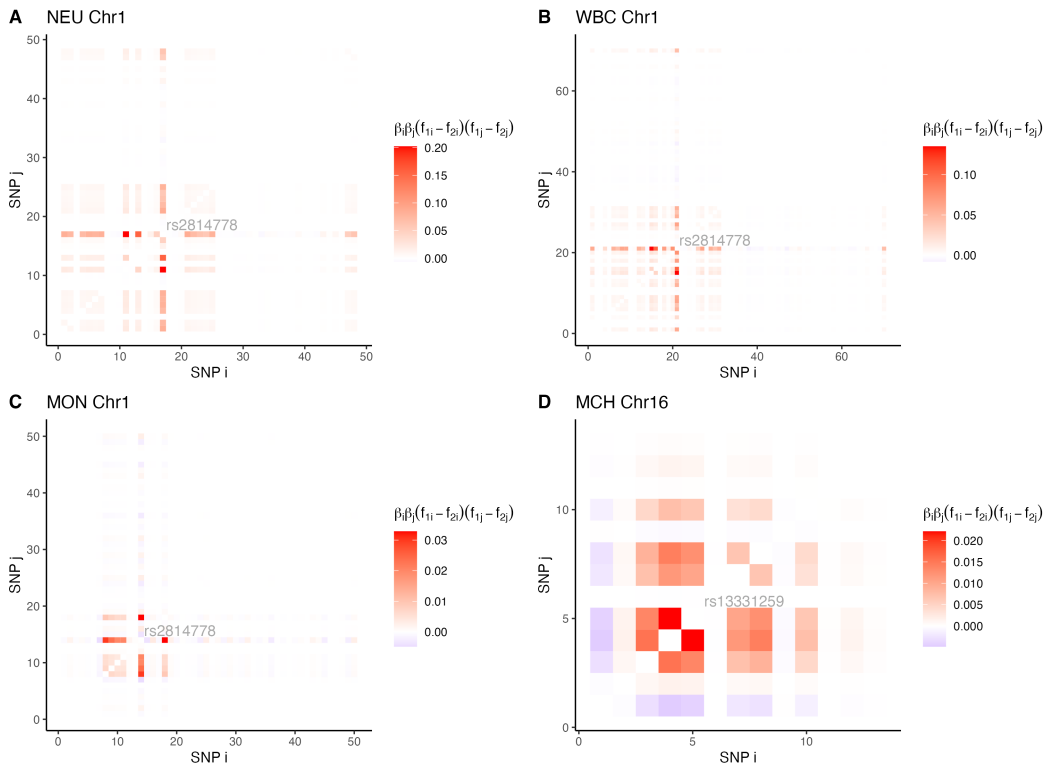

Figure S5: The LD contribution to the variance explained by variant pairs for (A) neutrophil counts (NEU), (B) white blood count (WBC), (C) monocyte count (MON), and (D) mean corpuscular hemoglobin (MCH). Only chromosomes where we suspected there was a disproportionate contribution to the variance explained are shown.

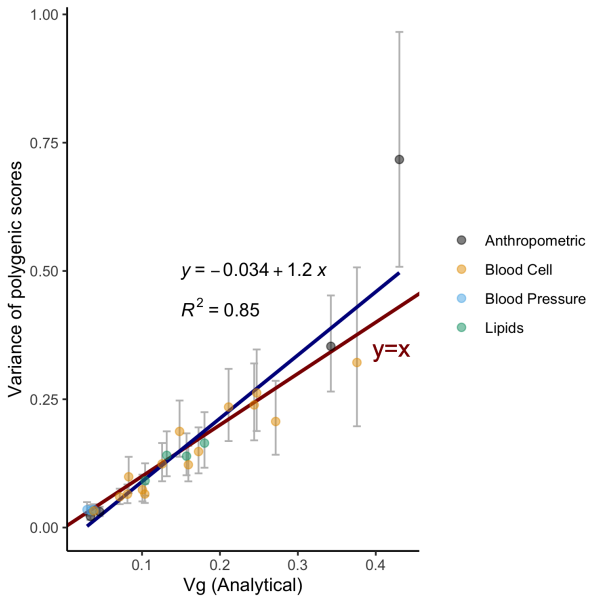

Figure S6: Expected variance explained estimated using Equation 1 (x-axis) vs the observed variance in polygenic scores (y-axis) in ASW are strongly correlated. Confidence intervals were generated by non-parametric bootstrap.