



Testing for differences in polygenic scores in the presence of confounding

Jennifer Blanc , * Jeremy J. Berg *

Department of Human Genetics, University of Chicago, 920 E 58th St CLSC, Chicago, IL 60637, USA

*Corresponding author: Department of Human Genetics, University of Chicago, 920 E 58th St CLSC, Chicago, IL 60637, USA. Email: jgblanc@uchicago.edu;

*Corresponding author: Department of Human Genetics, University of Chicago, 920 E 58th St CLSC, Chicago, IL 60637, USA. Email: jjberg@uchicago.edu

Polygenic scores have become an important tool in human genetics, enabling the prediction of individuals' phenotypes from their genotypes. Understanding how the pattern of differences in polygenic score predictions across individuals intersects with variation in ancestry can provide insights into the evolutionary forces acting on the trait in question and is important for understanding health disparities. However, because most polygenic scores are computed using effect estimates from population samples, they are susceptible to confounding by both genetic and environmental effects that are correlated with ancestry. The extent to which this confounding drives patterns in the distribution of polygenic scores depends on the patterns of population structure in both the original estimation panel and in the prediction/test panel. Here, we use theory from population and statistical genetics, together with simulations, to study the procedure of testing for an association between polygenic scores and axes of ancestry variation in the presence of confounding. We use a general model of genetic relatedness to describe how confounding in the estimation panel biases the distribution of polygenic scores in ways that depends on the degree of overlap in population structure between panels. We then show how this confounding can bias tests for associations between polygenic scores and important axes of ancestry variation in the test panel. Specifically, for any given test, there exists a single axis of population structure in the genome-wide association study (GWAS) panel that needs to be controlled for in order to protect the test. In the context of this result, we study the behavior of multiple approaches to control for stratification along this axis, including standard methods such using principal components as fixed covariates in the GWAS, linear mixed models, and a novel approach for directly estimating the axis using the test panel genotypes. Our analyses highlight the role of estimation noise in the models of population structure as a plausible source of residual confounding in polygenic score analyses.

Keywords: polygenic scores; confounding; population structure

Introduction

The calculation of polygenic scores (Purcell *et al.* 2009) has become a routine procedure in many areas of human genetics. The promise of polygenic scores is that they provide a means for phenotypic prediction from genotype data alone. By measuring the association between a genetic variant and phenotype in a genome-wide association study (GWAS), we obtain an estimate of its effect on the phenotype, averaged over the environments experienced by the individuals in that sample. These effect estimates can then be combined into polygenic scores in a separate prediction panel by summing the genotypes of individuals in that panel, weighted by the estimated effects. Under the relatively strict assumptions that genetic and environmental effects combine additively, that variation in the phenotype is not correlated with variation in ancestry within the GWAS panel, and that the individuals in the prediction panel experience a similar distribution of environments to those in the GWAS panel, these scores can be viewed as an estimate of each individual's expected phenotype, given their genotypes at the included sites. If these assumptions are met, polygenic scores would seem to provide a means of isolating at least some of the genetic effects on a given phenotype.

However, this promise of polygenic scores is also one of their main pitfalls. The effects of individual variants are typically estimated from population samples in which the environments that individuals experience vary as a function of their social, cultural, economic, and political contexts. Differences in these factors are often correlated with differences in ancestry within population samples, and these ancestry-environment correlations can induce systematic biases in the estimated effects of individual variants. Similar biases can also arise if genetic effects on the phenotype vary as a function of ancestry within the GWAS sample. Ancestry stratification has long been recognized as a problem in GWAS study design (Lander and Schork 1994), and many steps have been taken to guard against its effects. These include bias avoidance approaches, like the sampling of GWAS panels that are relatively homogeneous with respect to ancestry, and statistical bias correction approaches, such as the inclusion of genetic principal components (PCs) as covariates (Price *et al.* 2006), linear mixed models (Kang *et al.* 2010; Loh *et al.* 2015), and linkage disequilibrium (LD) score regression (Bulik-Sullivan *et al.* 2015). The abundance of biological signal among GWAS datasets (Visscher *et al.* 2017) suggests that these approaches have been broadly successful at limiting individual false positive

Received on 16 January 2025; accepted on 28 March 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of The Genetics Society of America. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

associations (Lawson et al. 2020). However, even if this is the case, effect size estimates can still exhibit slight stratification biases that are not significant enough to alter false discovery rates for individual variants. These biases can compound when aggregated across loci, leading to confounded predictions in which ancestry-associated effects are mistaken for genetic effects.

Separation of direct genetic effects from correlations between ancestry and either the environment or the genetic background is important to all applications of polygenic scores. Empirically, polygenic scores exhibit geographic clustering even in relatively homogeneous samples and after strict control for population stratification (Abdellaoui et al. 2019; Haworth et al. 2019; Kerminen et al. 2019; Trochet et al. 2021). It is natural to ask whether these observed differences reflect a real difference in the average genetic effect on the trait. From a population biology perspective, these patterns may be signals of natural selection (Berg and Coop 2014) or phenotype-biased migration (Abdellaoui et al. 2019). Medically, it is important to know if polygenic score differences or gradients represent real underlying gradients in the average genetic effect (Rosenberg et al. 2019), whether those gradients are caused by nonneutral evolutionary mechanisms or not. However, observed patterns of polygenic scores may also be driven by residual bias in effect size estimates, and stratification biases remain a persistent issue (Tan et al. 2024).

This issue has been particularly apparent in the detection of directional selection acting on complex traits. Polygenic scores are an ideal tool for this task, as studying the distribution of scores among individuals with differing ancestry allows us to aggregate the small changes in allele frequency induced by selection on a polygenic trait into a detectable signal (Latta R 1998; Latta RG 2004; Pritchard and Rienzo 2010; Pritchard et al. 2010). Several research groups have developed and applied methods to detect these signals (Turchin et al. 2012; Berg and Coop 2014; Field et al. 2016; Racimo et al. 2018; Edge and Coop 2019; Josephs et al. 2019; Uricchio et al. 2019; Stern et al. 2021). However, these efforts have been met with challenges, as several papers reported signals of recent directional selection on height in Europe using effects obtained from GWAS meta-analyses (Lango Allen et al. 2010; Turchin et al. 2012; Berg and Coop 2014; Wood et al. 2014; Mathieson et al. 2015; Robinson et al. 2015; Zoledziewska et al. 2015; Field et al. 2016; Berg et al. 2017; Guo et al. 2018; Racimo et al. 2018), only for these signals to weaken substantially or disappear entirely when reevaluated using effects estimated in the larger and more genetically homogeneous UK Biobank (Berg et al. 2019; Edge and Coop 2019; Sohail et al. 2019; Uricchio et al. 2019). Further analysis has suggested that much of the original signal could be attributed to spurious correlations between effect size estimates and patterns of frequency variation, presumably induced by uncorrected ancestry stratification in the original GWAS (Berg et al. 2019; Sohail et al. 2019).

Recently, in the context of selection tests, Chen et al. (2020) proposed a strategy to mitigate the impact of stratification by carefully choosing the GWAS panel such that, even if residual stratification biases in effect size estimates exist, they will be unlikely to confound the test (see also Le et al. 2022 for an example of this approach). They reasoned that because polygenic selection tests ask whether polygenic scores are associated with a particular axis of population structure in a given test panel, and because the bias induced by stratification in effect sizes depends on patterns of population structure in the GWAS panel (Robinson et al. 2015), then one should be able to guard against bias in polygenic selection tests by choosing GWAS and test panels where the patterns of population structure within the two panels are not expected to overlap.

However, this approach comes at the cost of reduced power: polygenic scores are generally less accurate when the effect sizes used to compute them are ported to genetically divergent samples (Martin et al. 2017; Wang et al. 2020; Carlson et al. 2022; Yair and Coop 2022; Ding et al. 2023). Less accurate polygenic scores are then less able to capture evolution of the mean polygenic score, all else being equal (Yair and Coop 2022). These decays in polygenic score accuracy also pose a significant challenge to their use in medicine. Scores that are predictive for some and not for others may exacerbate health inequities (Martin et al. 2019). Thus, realizing the potential of polygenic scores in both basic science and medical applications will require large, genetically diverse GWAS panels. Successfully deploying polygenic scores developed from these diverse panels will require that we have a precise understanding of how bias is produced in polygenic score predictions, and the development and evaluation of methods to protect against this bias.

In this article, we first model the covariance of genotypes in a GWAS and test panel in terms of an underlying population genetic model and provide expressions for the bias in the distribution of polygenic scores as a function of the underlying model. We then show how bias in the association between polygenic scores and a specific axis of ancestry variation in the test panel depends on the extent to which potential confounders in the GWAS lie along a specific axis of ancestry variation in the GWAS panel. Next, we evaluate ways to control for confounding along this axis, including the standard Principal Component Analysis (PCA)-based approach and linear mixed models, as well as a new approach inspired by our theoretical results, which uses test panel genotypes to estimate the axis directly. We find that the utility of each approach depends on several factors, including the number of independent single nucleotide polymorphisms (SNPs) used to compute the correction, the number of samples in the GWAS panel, and the amount of variance in the GWAS panel explained by the target axis.

Model

To model the distribution of genotypes in both panels, we assume that each individual's expected genotype at each site can be modeled as a linear combination of contributions from a potentially large number of ancestral populations, which are themselves related via an arbitrary demographic model. Natural selection, genetic drift, and random sampling each independently contribute to the distribution of genotypes across panels, and we make the approximation that these three effects can be combined linearly. In [Supplementary Section S1](#), we develop the full population model which we then extend to individuals. In the main text, we present only the individual genotype model, along with our model of the phenotype.

Genotypes

We consider two samples of individuals, one to compose the GWAS panel and one to compose the test panel. Individuals in each panel are created as mixtures of an arbitrary number of K underlying populations. These populations are related via an arbitrary demographic model (see [Supplementary Sections S1.1](#) and [S1.2](#)), where a_ℓ is the ancestral allele frequency at site ℓ . There are N test panel individuals and the vector of deviations of their genotypes from the mean genotype in the ancestral population ($2a_\ell$) is

$$X_\ell = X_{\ell,D} + X_{\ell,S} + X_{\ell,B}, \quad (1)$$

where $X_{\ell,D}$ and $X_{\ell,S}$ are the deviations due to drift and natural selection, respectively. We can think of the quantity $2a_\ell + X_{\ell,D} +$

$X_{\ell,S}$ as giving a set of expected genotypes given the evolutionary history of the populations from which the test panel individuals were sampled, and $X_{\ell,B}$ contains the binomial sampling deviations across individuals, given these expected genotypes.

Similarly, for the M individuals in the GWAS panel, the deviations of their genotypes can be decomposed as

$$G_{\ell} = G_{\ell,D} + G_{\ell,S} + G_{\ell,B}, \quad (2)$$

where $G_{\ell,D}$ and $G_{\ell,S}$ are the deviations due to drift and selection, and $G_{\ell,B}$ captures the binomial sampling variance, given the expected genotypes of the GWAS panel individuals.

Individuals in the two panels may draw ancestry from the same populations, or from related populations, which induces the following joint covariance structure

$$\text{Var}\left(\begin{bmatrix} X_{\ell,D} \\ G_{\ell,D} \end{bmatrix}\right) = 4a_{\ell}(1 - a_{\ell})\mathbf{F}, \quad (3)$$

where the matrix

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{XX} & \mathbf{F}_{XG} \\ \mathbf{F}_{GX} & \mathbf{F}_{GG} \end{bmatrix} \quad (4)$$

contains the within and between panel relatedness coefficients. The entries of \mathbf{F} give the relatedness between pairs of individuals given the underlying demographic model and the fraction of ancestry each individual draws from each population. Therefore, the entries of \mathbf{F} are directly related to the expected pairwise coalescent times between pairs of samples, given the demographic model (McVean 2009).

Phenotypes

We assume that individuals in the GWAS panel are phenotyped and that the trait includes a contribution from S causal variants, which make additive genetic contributions, as well as an independent environmental effect. The vector of mean-centered phenotypes for the M individuals in the GWAS panel can then be written

$$\begin{aligned} y &= \sum_{\ell} \beta_{\ell} G_{\ell} + e \\ &= u + e, \end{aligned} \quad (5)$$

where $u = \sum_{\ell} \beta_{\ell} G_{\ell}$ is the combined genetic effect of all S causal variants, and e represents the combination of all environmental effects.

We assume that the environmental effect on each individual is an independent Normally distributed random variable with variance σ_e^2 , but that the expected environmental effect can differ in some arbitrary but unknown way across individuals. We write the distribution of environmental effects as multivariate Normal, $e \sim \text{MVN}(c, \sigma_e^2 \mathbf{I})$, where c is the vector of expected environmental effects.

Similar to our decomposition in Equation 2, the genetic effect, u , can be broken down into the contributions from drift, selection, and binomial sampling such that $u = u_D + u_S + u_B$. Here, $u_S = \sum_{\ell} \beta_{\ell} G_{\ell,S}$ contains fixed effects reflecting the expected genetic contributions to the phenotype, given the history of selection acting on the phenotype, and given the ancestries of the individuals in the GWAS panels (see Supplementary Section S1.4). Both u_D and u_B have expectation zero, so $\mathbb{E}[u] = u_S$. The vector of individuals' expected phenotypes, given their ancestry and socio-environmental contexts, is therefore given by $u_S + c$. We assume that these are not known.

Results

Now, given these modeling assumptions, we describe how the relationship between the GWAS and test panels impacts the distribution of polygenic scores and the association between the polygenic scores and a given axis of population structure that is observed only in the test panel. We first consider the case in which no attempt is made to correct for population structure. Motivated by these results, we then outline the conditions that need to be met in order to ensure an unbiased association test. Finally, we explore how various correction strategies—including the standard PCA approach, a linear mixed model, and a novel approach that uses the test panel genotypes—perform in practice.

The impact of stratification bias on polygenic scores

We consider a vector of mean-centered polygenic scores, computed in the test panel. If the causal effects (β_{ℓ}) were known, then the polygenic scores would be given by

$$Z = \sum_{\ell} \beta_{\ell} X_{\ell}. \quad (6)$$

Of course, the causal effects are not known, and must be estimated in the GWAS panel. Conditional on the genetic and environmental effects on the phenotypes of individuals in the GWAS panel (i.e. u and e), and the genotypes at the focal site (G_{ℓ}), the marginal effect size estimate for site ℓ is given by

$$\hat{\beta}_{\ell} | G_{\ell}, u, e = \frac{y^T G_{\ell}}{G_{\ell}^T G_{\ell}} = \beta_{\ell} + \frac{u^T G_{\ell}}{G_{\ell}^T G_{\ell}} + \frac{e^T G_{\ell}}{G_{\ell}^T G_{\ell}}, \quad (7)$$

where we have decomposed the genetic effect into the causal contribution from the focal site and the contribution from the background, i.e. $u = \beta_{\ell} G_{\ell} + u_{-\ell}$. This allows us to further decompose the marginal association in Equation 7 into the causal effect (β_{ℓ}), the association between the focal site and the background genetic contribution from all other sites ($u_{-\ell}^T G_{\ell} / G_{\ell}^T G_{\ell}$), and the association with the environment ($e^T G_{\ell} / G_{\ell}^T G_{\ell}$).

The deviation of an allele's estimated effect size from its expectation depends in part on $G_{\ell,D}$, the component of variation in the GWAS panel genotypes due to genetic drift (see Supplementary Section S2). Because $G_{\ell,D}$ can be correlated with $X_{\ell,D}$ (deviations due to drift in the test panel genotypes) due to shared ancestry, the estimated effect sizes can become correlated with the pattern of genotypic variation in the test panel for reasons unrelated to the actual genetic effect of the variant. This leads to a bias in the polygenic scores,

$$\mathbb{E}[\hat{Z} - Z]^T = \mathbb{E}\left[\sum_{\ell=1}^S \frac{u^T G_{\ell}}{G_{\ell}^T G_{\ell}} X_{\ell}^T + \sum_{\ell=1}^S \frac{e^T G_{\ell}}{G_{\ell}^T G_{\ell}} X_{\ell}^T\right] \quad (8)$$

$$\approx \frac{S}{M} (\mu_S^T + c^T) \bar{\mathbf{F}}_{GX}, \quad (9)$$

(see Supplementary Section S3), where μ_S is the vector of expected genetic backgrounds, c is the vector of expected environmental effects, and

$$\begin{aligned} \bar{\mathbf{F}}_{GX} &= \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^T}{(G_{\ell,D} + G_{\ell,B})^T (G_{\ell,D} + G_{\ell,B}) / M}\right] \\ &\approx \frac{\mathbf{F}_{GX}}{1 + \bar{F}_G}. \end{aligned} \quad (10)$$

Here, $\bar{F}_G = \frac{1}{M} \sum_{m=1}^M F_{mm}$ is the average level of self-relatedness in the GWAS panel and $\bar{\mathbf{F}}_{GX}$ is the expected cross-panel genetic relatedness

matrix computed on standardized genotypes. This matrix is approximately equal to $\frac{F_{GX}}{(1+F_G)}$ if F_G is small. We also note that in Equation 9 we make the assumption that $X_{\ell,S}$ and $G_{\ell,S}$ are small relative to $X_{\ell,D}$ and $G_{\ell,D}$, a common assumption supported by the observation that for highly polygenic traits (like the type we are considering here) the effect of selection on any individual site will be small (Bulmer 1971; Latta R 1998; Le Corre and Kremer 2003; Kremer and Le Corre 2012; Berg and Coop 2014).

If the GWAS and test panels do not overlap in population structure, then $\tilde{F}_{XG} = \mathbf{0}$, and the polygenic scores are unbiased with respect to ancestry (i.e. $\mathbb{E}[\hat{Z} - Z] = 0$), regardless of the confounders μ_S and c (Purcell et al. 2009; Chen et al. 2020; Le et al. 2022). Stratification may still bias individual effects; but these residual biases are indistinguishable from noise from the perspective of the polygenic scores, as they are uncorrelated with all axes of population structure present in the test panel.

Bias in polygenic scores leads to biased polygenic score associations

We want to test the hypothesis that the polygenic scores are associated with some test vector, T . We assume that T is measured only in the test panel and might represent an ecogeographic variable of interest (e.g. latitude Berg and Coop 2014 or an encoding of whether one lives in a particular geographic region or not Abdellaoui et al 2019; Abdellaoui et al. 2022, the fraction of an individual's genome assigned to a particular "ancestry group" Turchin et al. 2012; Racimo et al. 2018, or one of the top genetic PCs of the test panel genotype matrix Josephs et al. 2019).

To test for association of polygenic scores with the test vector, we take our test statistic the as slope of the regression of the polygenic scores against the test vector, which we denote q . Assuming T is standardized, this slope is given by

$$q = \frac{1}{N} Z^T T. \quad (11)$$

A more powerful test is available by modeling the neutral correlation structure among individuals due to relatedness (see Supplementary Section S9), but the simpler i.i.d. model presented here is sufficient for our purposes. Under the null model where selection has not perturbed allele frequencies in the test panel, $\mathbb{E}[q] = 0$, reflecting the fact that genetic drift is directionless.

In practice, an estimate of q is obtained using the polygenic scores computed from estimated effect sizes, i.e. $\hat{q} = \frac{1}{N} \hat{Z}^T T$. The bias in the polygenic score association test statistic (\hat{q}) then follows straightforwardly from the bias in the polygenic scores (see Supplementary Section S4 for additional details),

$$\begin{aligned} \mathbb{E}[\hat{q} - q] &= \mathbb{E}[\hat{Z} - Z]^T T \\ &\approx \frac{S}{NM} (\mu_S^T + c^T) \tilde{F}_{GX} T. \end{aligned} \quad (12)$$

Therefore, we expect the polygenic score association test to be biased when the test vector (T) aligns with the vector of expected phenotypes ($\mu_S + c$) in a space defined by the cross-panel genetic relatedness matrix (\tilde{F}_{XG}). The conditions for an unbiased polygenic score association test are therefore narrower than the conditions needed to ensure unbiased polygenic scores in general. Rather than requiring that $\tilde{F}_{XG} = \mathbf{0}$, we need only to ensure that

a certain linear combination of the entries of \tilde{F}_{XG} are equal to zero, i.e. that $\tilde{F}_{GX} T = 0$.

We can gain further intuition by expressing the association statistic, q , in a different way. Specifically, we can reframe this test as a statement about the association between the effect sizes and a set of genotype contrasts, $r_\ell = \frac{1}{N} X_\ell^T T$, which measure the association between the test vector and the genotypes at each site (Berg and Coop 2014). Writing β and r for the vectors of effect sizes and genotype contrasts across loci, the association test statistic can be rewritten as

$$q = \beta^T r. \quad (13)$$

This allows us to rewrite the bias in the estimator, \hat{q} , as

$$\begin{aligned} \mathbb{E}[\hat{q} - q] &= \frac{S}{M} \mathbb{E}[(\hat{\beta}^T - \beta^T) r] \\ &\approx \frac{S}{M} (\mu_S^T + c^T) \tilde{F}_{Gr}, \end{aligned} \quad (14)$$

where

$$\begin{aligned} \tilde{F}_{Gr} &= \mathbb{E} \left[\frac{G_{\ell,D} r_{\ell,D}^T}{(G_{\ell,D} + G_{\ell,B})^T (G_{\ell,D} + G_{\ell,B}) / M} \right] \\ &= \tilde{F}_{GX} T. \end{aligned} \quad (15)$$

Here, Equation 14 expresses the bias entirely in terms of vectors that belong to the GWAS panel: for each GWAS panel individual m , $\tilde{F}_{Gr,m}$ measures the covariance between individual m 's genotype and the genotype contrasts of the test, standardized at each site by the variance of genotypes across individuals in the GWAS panel (Equation 15). Thus, \hat{q} is biased when the vector of expected phenotypes ($\mu_S + c$) aligns with this vector of standardized covariances (\tilde{F}_{Gr}). Confounders that are orthogonal to this axis do not generate bias in the association test, even if they bias the polygenic scores along other axes.

Controlling for stratification bias in polygenic association tests

Given the above results, how can we ensure that patterns we observe in the distribution of polygenic scores are not the result of stratification bias? As discussed above, a conservative solution is to prevent bias by choosing a GWAS panel that does not have any overlap in population structure with the test panel, but this is not ideal due to the well-documented portability issues that plague polygenic scores (Martin et al. 2017; Mostafavi et al. 2020; Ding et al. 2023), and because it limits which GWAS datasets can be used to test a given hypothesis. Another obvious solution is to include the vectors of expected genetic and environmental effects, u_S and c respectively, as covariates in the GWAS. Doing so would remove all ancestry-associated bias from the estimated effects and thus ensure that any polygenic score association test carried out using these effects would be unbiased. However, u_S and c are typically not measurable, so this is generally not feasible. Alternatively, our analysis above suggests that including \tilde{F}_{Gr} as a covariate in the GWAS model is the sufficient condition for an unbiased test no matter what pattern of confounding exists in the GWAS panel.

Including \tilde{F}_{Gr} removes stratification bias

If we include \tilde{F}_{Gr} as a single fixed-effect covariate in the GWAS model, variation along \tilde{F}_{Gr} can no longer be used to estimate effect

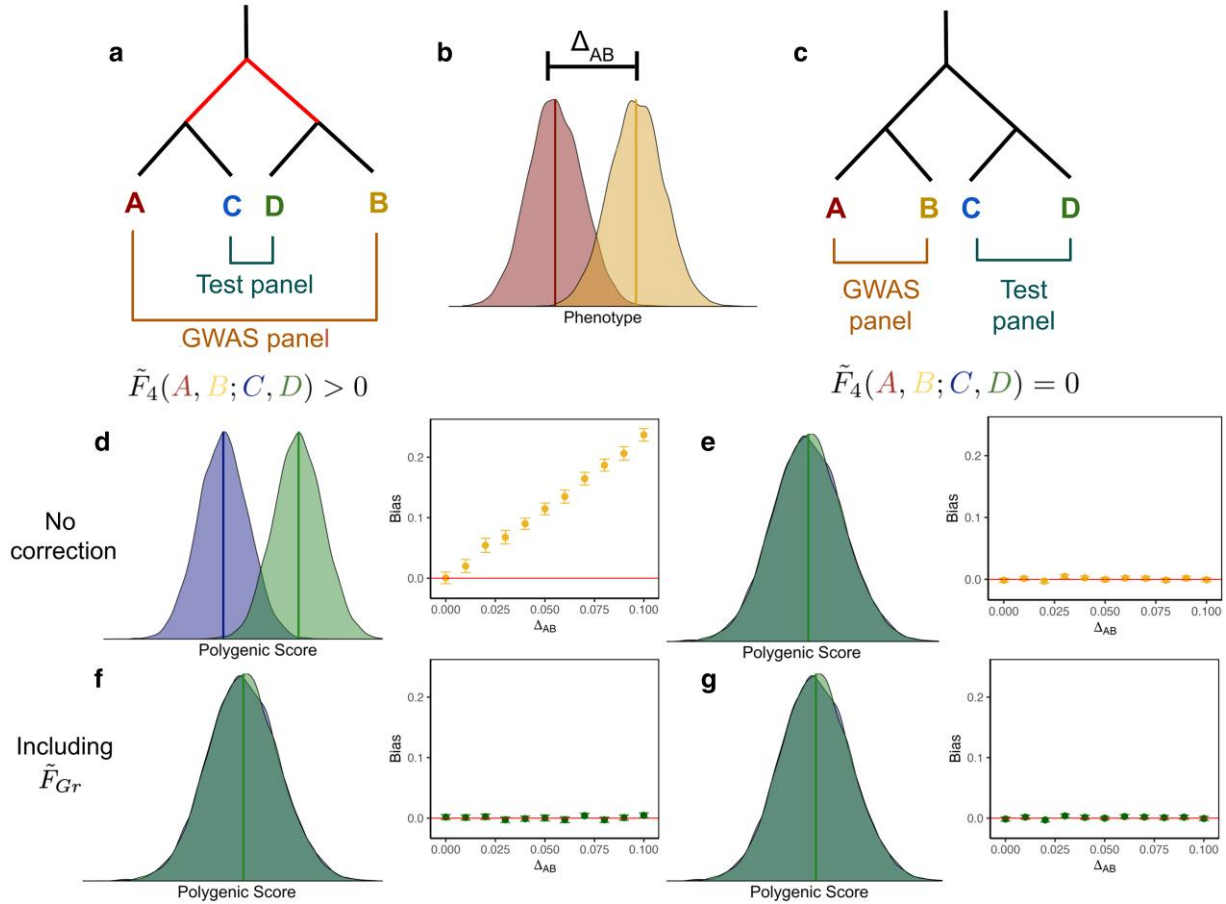


Fig. 1. Schematic of two different panel configurations. The effect of stratification depends on the overlapping structure between the GWAS and test panels. a, c) Two different topologies were used to create the GWAS and test panels. b) Stratification in the GWAS panel was modeled by drawing an individual's phenotype $y \sim N(0, 1)$ and adding Δ_{AB} if they originated from population B. d) When there is overlapping structure between GWAS and test panels, there is an expected mean difference between polygenic scores in populations C and D. Additionally, the bias in \hat{q} increases with the magnitude of stratification in the GWAS. e) However, when there is no overlapping structure between panels, there is no expected difference in mean polygenic scores between C and D and \hat{q} remains unbiased regardless of the magnitude of stratification. f, g) Including \tilde{F}_{Gr} as a covariate in the GWAS controls for stratification, eliminating bias in \hat{q} regardless of Δ_{AB} or the overlapping structure between GWAS and test panels.

sizes. As a result $\hat{\beta}$ is uncorrelated with genotype contrasts r under the null. If there is confounding along other shared axes of ancestry variation, the polygenic scores may still be biased along other axes, as

$$\mathbb{E}[\hat{Z} - Z]^T \approx \frac{S}{M} (\mu_S^T + c^T) \tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}}, \quad (16)$$

where

$$\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \approx \mathbf{P} \tilde{\mathbf{F}}_{GX} \quad (17)$$

and $\mathbf{P} = (\mathbf{I} - \frac{1}{\|\tilde{\mathbf{F}}_{Gr}\|^2} \tilde{\mathbf{F}}_{Gr} \tilde{\mathbf{F}}_{Gr}^T)$. $\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}}$ therefore captures cross-panel relatedness along all axes of variation other than that specified by \tilde{F}_{Gr} . Controlling for variation aligned with \tilde{F}_{Gr} ensures that $\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr} T} = 0$, and it follows that

$$\mathbb{E}[\hat{q} - q] \approx \frac{S}{NM} (\mu_S^T + c^T) \tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr} T} \approx 0 \quad (18)$$

and the polygenic score association test will be unbiased (see [Supplementary Sections S5 and S6](#)).

Relationship between \tilde{F}_{Gr} and PCA

A standard approach to controlling for population stratification in polygenic scores is to include the top J PCs of the GWAS panel genotype matrix as covariates in the GWAS, for some suitably large value of J ([Price et al. 2006](#)). In our model, how does this approach relate to including \tilde{F}_{Gr} as a covariate in the GWAS?

As outlined in Genotypes section, \mathbf{F}_{GG} contains the expected within panel relatedness for the individuals in the GWAS panel, the structure of which is determined by the demographic model. If we could take the eigendecomposition of \mathbf{F}_{GG} directly, the resulting PCs are what we refer to as “theoretical” PCs. The number of theoretical PCs that correspond to structure is entirely dependent on the population model. For example, in the [Toy model](#) section below we simulate under a 4 population sequential split model ([Fig. 1](#)), in which case there are three theoretical PCs that reflect real underlying structure. Later, in the [Grid simulations](#) section, we simulate under a symmetric equilibrium migration model on a six-by-six lattice grid ([Fig. 4](#)), in which case there are 35 theoretical PCs reflecting underlying population structure. Including these theoretical PCs as covariates in the GWAS would be sufficient to remove all ancestry-associated bias in effect size estimates and render the resulting polygenic scores uncorrelated with any axis of ancestry variation under the null hypothesis.

To see how the PCA correction approach works in the context of our theory, we can write \tilde{F}_{Gr} as a linear combination of GWAS panel population PCs,

$$\tilde{F}_{Gr} = \sum_i \eta_i U_i, \quad (19)$$

where U_i is the i th PC of \mathbf{F}_{GG} and the weights are given by $\eta_i = \text{Cov}(U_i, \tilde{F}_{Gr})$. Estimating the marginal associations with \tilde{F}_{Gr} as a covariate can therefore be understood as fitting a model in which all theoretical PCs are included as covariates, but the relative magnitude of the contributions from different PCs is fixed, and we estimate only a single slope that scales the contributions of all of the PCs jointly, i.e.

$$y = G_\ell \beta_\ell + \left(\sum_i \eta_i U_i \right) \omega + e. \quad (20)$$

As a corollary, if we perform a polygenic score association test using GWAS effect size estimates in which the top J theoretical PCs of \mathbf{F}_{GG} are included as covariates, a sufficient condition for the included PCs to protect against bias from unmeasured confounders in a particular polygenic score association test is that \tilde{F}_{Gr} is captured by these J top PCs, i.e. that $\eta_i \approx 0$ for $i > J$.

A second interpretation of the PC correction approach is that it operates on the hypothesis that the major axes of confounding in a given GWAS panel (i.e. μ_S and c in our notation) can be captured by the included PCs (Vilhjálmsson and Nordborg 2013). If this condition is met, effect size estimates are unbiased with respect to all axes of ancestry variation, whether they exist within a given test panel or not, and therefore any polygenic score association test that uses these effect size estimates will be unbiased with respect to ancestry as well. Combining this interpretation with results from above, theoretical PCs should successfully eliminate bias in polygenic score association tests if the J PCs included in the GWAS either capture the confounding effects on the phenotype, eliminating all effect size bias, or if they capture \tilde{F}_{Gr} , ensuring that effect size bias relevant to the test is removed.

Controlling for bias in practice

So far we have shown the conditions under which including \tilde{F}_{Gr} or the top J theoretical PCs as fixed covariates removes stratification bias and leads to an unbiased association test. However, both \tilde{F}_{Gr} and U are theoretical quantities that depend on the population model, which we do not observe in practice. Instead, we must estimate these quantities, \tilde{F}_{Gr} and \hat{U} , with error, from sample genotype data. In this section, we investigate the role of estimation error in both of the quantities when controlling for bias in practice. Additionally, we outline the relationship between PCA and linear mixed models, demonstrating how estimation error and model constraint can also impact the ability of linear mixed models to unbiased the test statistic.

Sample principal components

The sample PCs, \hat{U} , can be computed by taking the eigendecomposition of the empirical genetic relatedness matrix (GRM), or the singular value decomposition of the genotype matrix. Existing results from random matrix theory allow us to obtain some understanding of the accuracy of \hat{U} as an estimator of U . Specifically, in many GWASs the number of individuals in the GWAS panel, M , is roughly on the same order as the number of SNPs, L . In this setting, the accuracy of the sample eigenvector \hat{U}_j depends on the

corresponding theoretical eigenvalue (λ_j) and the ratio of the number of individuals to the number of SNPs in the GWAS panel (M/L). As shown first by Patterson et al. (2006) in the context of genetics (see also Baik et al. 2005), PCA exhibits a phase change behavior in which a given sample PC is only expected to align with the theoretical PC if the corresponding theoretical eigenvalue is greater than a threshold value of $1 + \sqrt{\frac{M}{L}}$. Below this threshold, the sample PC is orthogonal to the theoretical PC.

However, even when the corresponding eigenvalue exceeds this threshold, the angle between the sample PC and the theoretical PC may still be substantially less than one, particularly if the relevant eigenvalue does not far exceed the detection threshold (Johnstone and Paul 2018; Bloemendal and Chen 2019). Specifically, the squared correlation between the theoretical PC and the sample PC is approximately

$$\left(U_j^T \hat{U}_j \right)^2 \approx \begin{cases} \frac{1 - \frac{M}{L} / (\lambda_j - 1)^2}{1 + \frac{M}{L} / (\lambda_j - 1)^2}, & \lambda_j > 1 + \sqrt{\frac{M}{L}} \\ 0, & \lambda_j \in \left[1, 1 + \sqrt{\frac{M}{L}} \right] \end{cases} \quad (21)$$

(see Johnstone and Paul 2018 for details). Thus, even in cases where \tilde{F}_{Gr} is fully captured by the top J theoretical PCs, either of these two related phenomena may make it difficult to accurately approximate \tilde{F}_{Gr} as a linear combination of the top J sample PCs, leading to a failure to fully account for stratification bias in polygenic score association tests.

Linear mixed models

Another common approach to controlling for stratification bias in practice is to use linear mixed models (LMMs) (Kang et al. 2010; Loh et al. 2015). In LMMs, phenotypic resemblance between related individuals is modeled via the inclusion of a random effect with a covariance structure given by the GWAS panel GRM, i.e.

$$\begin{aligned} y &= \beta_\ell G_\ell + u + e \\ u &\sim \text{MVN}(0, \sigma^2 \mathbf{F}_{GG}), \end{aligned} \quad (22)$$

where σ^2 is a free parameter that controls the total scale of the variance component that tracks relatedness. This can be justified as a model for the distribution of the genetic background under an assumption of neutrality (Schraiber et al. 2024), or by an assumption that environmental similarity should roughly track genetic relatedness (Vilhjálmsson and Nordborg 2013). Previous work has shown that fitting this standard GWAS LMM is equivalent to estimating the effect size under a model in which all PCs of the GRM are included as covariates, but the effect of each PC is constrained by a normal prior with a variance proportional to its corresponding eigenvalue (Hoffman 2013; Zhang and Pan 2015; Schraiber et al. 2024). In other words, a model in which

$$\begin{aligned} y &= \beta_\ell G_\ell + \sum_i \hat{U}_i \alpha_i + e \\ \alpha_i &\sim N(0, \hat{\lambda}_i \sigma^2), \end{aligned} \quad (23)$$

where \hat{U}_i is the i th sample PC, $\hat{\lambda}_i$ is the corresponding sample eigenvalue. This relationship suggests at least two possible ways that LMMs might fail to accurately model stratification along the \tilde{F}_{Gr} axis. First, because the relevant PCs in Equation 23 are the sample PCs, LMMs are potentially susceptible to the PC accuracy issues as

we described for PCA above. Second, even if the PCs are well estimated, the common scale of variation imposed by the Normal prior constrains the magnitude of confounding that can be captured by a given PC.

Estimating \hat{F}_{Gr} directly using test panel genotypes

Given the limitations of PCA and LMMs, it is natural to ask whether alternative estimators of \hat{F}_{Gr} might perform better. One option, suggested by our theoretical results, is a direct estimator that utilizes the relevant test panel genotype contrasts. Given the test panel genotype contrasts (r_ℓ) and GWAS panel genotypes (G_ℓ), we can obtain a direct estimator of \hat{F}_{Gr} as

$$\hat{F}_{Gr} = \frac{1}{L} \sum_{\ell=1}^L \frac{G_\ell r_\ell}{G_\ell^T G_\ell / M}. \quad (24)$$

Then, if \hat{F}_{Gr} is a sufficiently accurate estimator of \tilde{F}_{Gr} , we should be able to render a given polygenic score association test unbiased by estimating marginal effects under the model

$$y = G_\ell \beta_\ell + \hat{F}_{Gr} \omega + \varepsilon, \quad (25)$$

and ascertaining SNPs for inclusion in the polygenic scores via standard methods.

We expect this method to be successful when the variance of the error component in \hat{F}_{Gr} is small relative to the variance of the entries of \tilde{F}_{Gr} . The variance of \tilde{F}_{Gr} increases when there is a greater amount of overlap in population structure between the two panels along this specific axis. To understand the variance of the error component, we can consider a linear model that predicts the GWAS panel genotypes using the test panel genotype contrasts. Let \tilde{G}_i denote the vectors of genotypes for GWAS individual i and let \tilde{r} represent the test panel genotype contrasts, each standardized by the variance in the GWAS panel. We can then fit the linear model

$$\tilde{G}_i = \tilde{r} \tilde{F}_{Gr,i} + e. \quad (26)$$

The regression coefficient estimate from the fitted model is then the i th entry in our population structure estimator, \hat{F}_{Gr} . The error in \hat{F}_{Gr} therefore behaves like the error in a typical regression coefficient and should be minimized when the number of SNPs included, L , is large, and when the test panel sample size, N , is large, so that the \tilde{r} are well estimated.

This approach proposes using the test panel genotype data twice: once when controlling for stratification in the GWAS panel, and a second time when testing for an association between the polygenic scores and the test vector. One concern is that this procedure might remove the signal we are trying to detect. In [Supplementary Section S7.1](#), we show that while this is true for naive applications, the effect will be small so long as the number of SNPs used to compute the correction is large relative to the number included in the polygenic score (i.e. $S \ll L$). Notably, controlling for sample PCs of the GWAS panel genotype matrix will induce a similar effect if the sample PCs capture \tilde{F}_{Gr} . We confirm via simulations (see [Supplementary Section S7.2](#) and [Fig. S1](#)) that the downward bias in \hat{q} when including \hat{F}_{Gr} or sample PCs is minimal when $S \ll L$. Further concerns about downward biases in applications could likely be ameliorated via the “leave one chromosome out” scheme commonly implemented in the context of linear mixed models ([Listgarten et al. 2012](#); [Loh et al. 2015](#)) or via iterative

approaches that first aim to ascertain SNPs using a genome-wide estimate of \hat{F}_{Gr} before reestimating effects using an estimate of \hat{F}_{Gr} computed from sites not in strong LD with any of the ascertained sites.

Applications

In this section, using theory and simulations, we consider a number of concrete examples with varying degrees of alignment between the axis of stratification and the axis of population structure relevant to the polygenic score association test. We demonstrate how these biases play out in practice and how well PCs, LMMs, and \hat{F}_{Gr} capture bias in different circumstances.

Toy model. Stratification bias depends on $\tilde{F}_4(A, B; C, D)$

We first consider a toy model with four populations (labeled A, B, C, and D) that are related by an evenly balanced population phylogeny ([Fig. 1](#)). The GWAS panel consists of an equal mixture of individuals from populations A and B, and we test for a difference in the mean polygenic score between populations C and D under two different topologies: one where A and C are sister to one another ([Fig. 1a](#)), and another where A and B are sister ([Fig. 1c](#)).

For simplicity, we consider a purely environmental phenotype (i.e. $h^2 = 0$) with a mean difference between populations A and B equal to Δ_{AB} ([Fig. 1b](#)). Following from [Equation 7](#), the marginal effect size estimate for site ℓ is

$$\begin{aligned} \hat{\beta}_\ell | G_\ell, e &= \frac{G_\ell^T e}{G_\ell^T G_\ell} \\ &= \frac{1}{2} \frac{\Delta_{AB} (\hat{p}_{A,\ell} - \hat{p}_{B,\ell})}{G_\ell^T G_\ell / M} + \frac{G_\ell^T \varepsilon}{G_\ell^T G_\ell}, \end{aligned} \quad (27)$$

where $\hat{p}_{A,\ell}$ and $\hat{p}_{B,\ell}$ are the observed sample allele frequencies for populations A and B at site ℓ (see also [Equation 2.3](#) in the supplement of [Robinson et al. 2015](#)).

Then, using these effect sizes to test for a difference in mean polygenic score between populations C and D, the bias in our association test statistic is,

$$\begin{aligned} \mathbb{E}[\hat{q} - q] &= \Delta_{AB} \sum_{\ell=1}^S \mathbb{E} \left[\frac{(\hat{p}_{A,\ell} - \hat{p}_{B,\ell})(\hat{p}_{C,\ell} - \hat{p}_{D,\ell})}{G_\ell^T G_\ell / M} \right] \\ &= \Delta_{AB} S \tilde{F}_4(A, B; C, D), \end{aligned} \quad (28)$$

where $\tilde{F}_4(A, B; C, D)$ is a version of Patterson’s F_4 statistic ([Reich et al. 2009](#); [Patterson et al. 2012](#)), standardized by the genotypic variance in the GWAS panel, which measures the amount of genetic drift common to populations A and B that is also shared by populations C and D. Writing the bias in terms of this modified F_4 statistic helps illustrate the role of cross-panel population structure in driving stratification bias in polygenic scores. The effect estimate at site ℓ is a linear function of $\hat{p}_{A,\ell} - \hat{p}_{B,\ell}$, so the test will be biased if $\hat{p}_{A,\ell} - \hat{p}_{B,\ell}$ is correlated with $\hat{p}_{C,\ell} - \hat{p}_{D,\ell}$. This is true for the demographic model in [Fig. 1a](#), where shared drift on the internal branch generates such a correlation, yielding a positive value for $\tilde{F}_4(A, B; C, D)$. In contrast, for the model in [Fig. 1c](#), there is no shared internal branch and $\tilde{F}_4(A, B; C, D) = 0$.

To test this prediction, we simulated 100 replicates of four populations related by this topology. In the GWAS panel populations, we simulated purely environmental phenotypes with a difference in mean phenotype (as outlined above), conducted a GWAS, ascertained SNPs, and then used these SNPs to construct polygenic scores and computed \hat{q} in the test panel. The results

are consistent with our theoretical expectations: the test statistic is biased for the topology with $\tilde{F}_4(A, B; C, D) > 0$ (Fig. 1d), but unbiased when $\tilde{F}_4(A, B; C, D) = 0$ (Fig. 1e).

Given the population model, $\tilde{\mathbf{F}}_{\text{XC}} = \mathbf{0}$ for the unconfounded topology, making \tilde{F}_{Gr} a vector of zeros. Therefore, rerunning the GWAS including \tilde{F}_{Gr} did not change the outcome of the already unbiased test (Fig. 1g). For the confounded topology, the structure in $\tilde{\mathbf{F}}_{\text{XC}}$ reflects the deepest split in the phylogeny and is aligned with T . \tilde{F}_{Gr} is therefore an indicator of which GWAS panel individuals are on which side of the deepest split, and including it as a covariate in the GWAS eliminated the bias for the confounded topology (Fig. 1f).

Quantifying error in estimators of \tilde{F}_{Gr}

As outlined above, in practice \tilde{F}_{Gr} cannot be observed directly, and must be estimated with error from the data. To illustrate the impact of this estimation error on the performance of the direct estimator, the sample PCs, and LMMs in a simple, well-understood case, we performed simulations using three different versions of our toy model, varying the length of the internal branch shared between the test and GWAS panels. Specifically, since \tilde{F}_{Gr} is known in this toy model, we could compute the error in either the direct or the PCA-based estimator as one minus the squared correlation between \tilde{F}_{Gr} and the corresponding estimator. We took all of these vectors to be standardized, so this is

$$\text{Error} = 1 - (\hat{\mathbf{x}}^T \tilde{\mathbf{F}}_{Gr})^2, \quad (29)$$

where $\hat{\mathbf{x}}$ represents the appropriate estimator.

For each simulation, we estimated \tilde{F}_{Gr} as in Equation 24, using L genome-wide SNPs with a frequency of greater than 1% in both the test and GWAS panels. For PCA, we computed sample PCs via singular value decomposition of the genotype matrix using the same set of SNPs that were used to compute \tilde{F}_{Gr} . We then took \hat{U}_1 (i.e. the first sample PC) as the PCA-based estimator of \tilde{F}_{Gr} (McVean 2009). In all of these simulations, we held the GWAS and test panel sample sizes constant at $N, M = 1,000$ and varied the number of SNPs (L) as a way to vary the accuracy of the estimators. We simulated 100 replicates for each topology and plotted the resulting averages across these replicates in Fig. 2.

First, we simulated a scenario of complete overlap, where there was a single population split, and individuals in both the GWAS and test panels were independently drawn as 50:50 mixtures from the two populations on either side of the split (Fig. 2a). When the GWAS sample size (M) was on the same order as the number of SNPs (L), the direct estimator had a smaller error than the first PC (Fig. 2b) and consequently reduced the bias in \hat{q} by a larger amount (Fig. 2c). Intuitively, the direct estimator singles out the relevant axis of population structure because we have already identified it ourselves in the test panel. In contrast, PCA has to find this axis “on its own” in the high dimension GWAS panel genotype data, and thus pays an additional cost. However, when $M \ll L$, PCA no longer has to pay this additional cost, and its performance improves to match that of the direct estimator.

We next simulated under the same toy model of partial overlap in population structure between test and GWAS panels that we considered above in Fig. 1a. This resulted in an increase in the error of the direct estimator relative to the complete overlap case because the test panel genotype contrasts are less informative about the relevant axis of structure in the GWAS panel. In contrast, the error in \hat{U}_1 was unchanged, as the amount of structure

in the GWAS panel is the same as in Fig. 2a. Notably, in this case the direct estimator still outperformed PCA when $M/L \approx 1$, but PCA performed better as M/L decreases.

Finally, in Fig. 2g, we reduced the internal branch length even further, which caused \tilde{F}_{Gr} to perform poorly even when $L \gg M$. Intuitively, because the correlation between allele frequency contrasts is so small, the direct estimator requires a very large number of SNPs to produce an accurate estimate. In this case, PCA outperforms the direct estimator once the detection threshold (see Equation 21) is crossed. Similar to the two previous models, the reduction of bias in \hat{q} closely tracks the error in the estimators of population structure (Fig. 2i). However, we note that across all three population models, the reduction is slightly larger than expected for \hat{U}_1 (Supplementary Fig. S1).

LMMs are too constrained to control for strong confounding

For each of the three scenarios outlined above, we also estimated effect sizes using a standard LMM (Fig. 2c, f, i, teal points), implemented in GCTA (Yang et al. 2011, 2014). Interestingly, we found that when PC1 is very poorly estimated, the LMM outperformed PC1, and approximately matched the performance of our direct estimator for the complete overlap case (Fig. 2c). Increases in L initially led to improvements in the performance of the LMM, but these improvements began to level off at roughly the same point that PC1 crossed the detection threshold and started to become well estimated. Past this point, the residual bias when using the LMM remained flat at a constant proportion of the uncorrected bias.

We gained two valuable insights from these results. First, although LMMs have an interpretation in terms of the PCs of the GRM (i.e. Equation 23), the fact that the LMM reduces bias even when PC1 is below the detection threshold indicates that LMMs do not pay the same cost in high-dimensional data that the fixed-effect PCA method does. One interpretation of this result is that because the LMM includes all of the PCs, it can use lower sample PCs to capture variance that lies along theoretical PC 1, but which is missed by sample PC 1 due to its inaccurate estimation. Perhaps a more straightforward interpretation, however, is that the accuracy of the LMM in this regime depends on the accuracy of the individual entries of the GRM, which are all estimated pairwise, and thus not impacted by the costs of high dimensionality.

The second insight is that when there is strong confounding along one particular axis, the Normal prior on the effect of each PC assumed by the LMM is too restrictive. Intuitively, because PC 1 explains only a small amount of the total variance, we can think of fitting this LMM as roughly equivalent to first estimating the variance scaling parameter, σ^2 , using all other PCs (i.e. $\hat{\sigma}^2 = \frac{1}{M-2} \sum_{i=2}^{M-1} \frac{(y^T \hat{U}_i)^2}{\lambda_i}$) and then estimating the effect size, β_ℓ , under a model where α_1 is constrained by the prior $\alpha_1 \sim N(0, \lambda_1 \hat{\sigma}^2)$. Because there is no confounding along the other PCs, this limits the amount of confounding along PC 1 that can be effectively controlled, resulting in significant residual bias in the test statistic. Thus, while LMMs are clearly effective at modeling background genetic variation under an assumption of evolutionary neutrality, they may often be too constrained to effectively capture strong environmental confounders that do not fit this pattern. This results supports the commonly used combined approach in which fixed effects covariates (e.g. PCs) are used to model the most significant axes of structure, with an LMM included to model background genetic variation on less significant axes.

Alternative approaches based on test panel PCs

Finally, we considered alternative ways of using the test panel data to control for stratification bias. One natural idea is to

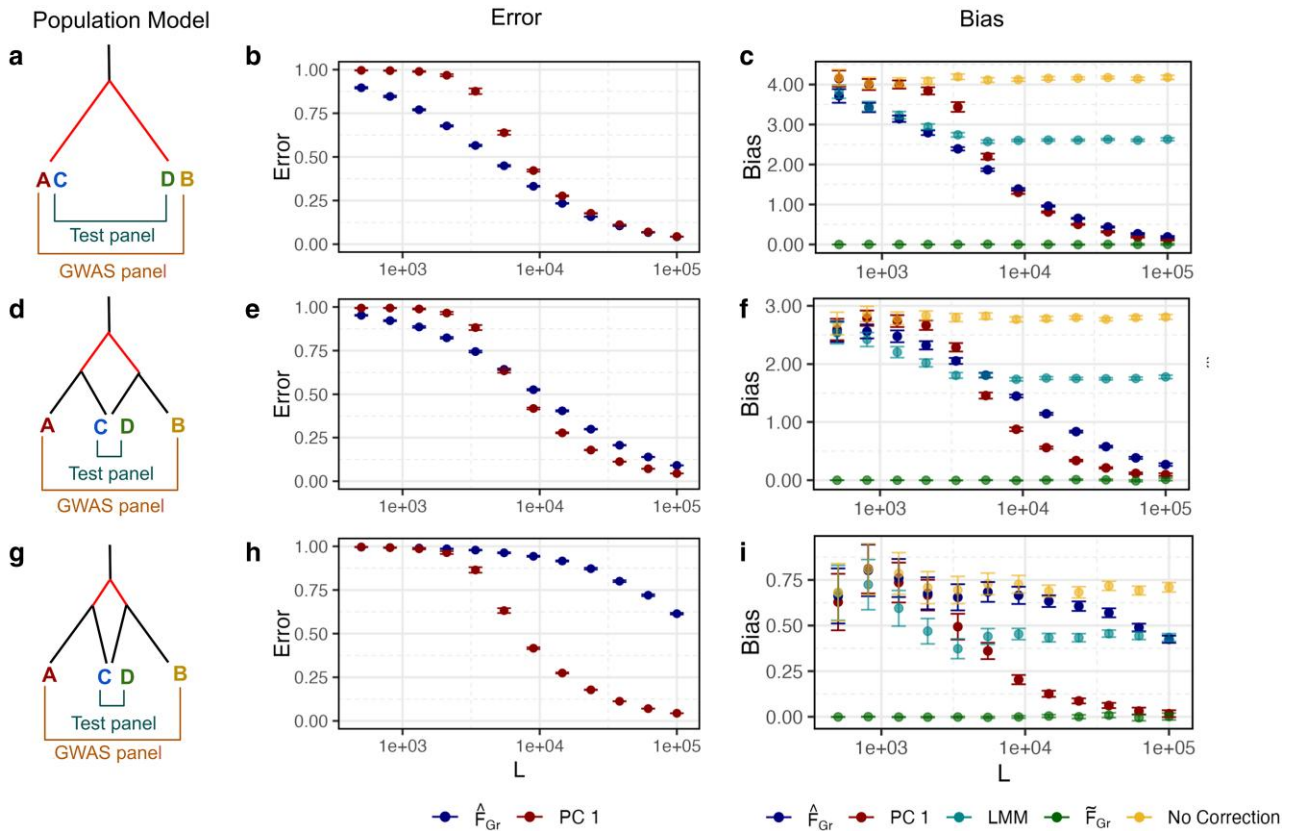


Fig. 2. Error in estimators of \hat{F}_{Gr} depends on the number of SNPs used to compute them. a) We simulated a population model with a single split and sampled an equal proportion of individuals from each population to make a GWAS and test panel. d, g) Here, we simulated population models with two splits and sampled individuals in the overlapping structure configuration. b, e, h) As \hat{F}_{Gr} is known for these population models, we computed the error in \hat{U}_1 and \hat{F}_{Gr} as estimators of \hat{F}_{Gr} using Equation 29. For both estimators, error decreased as the number of SNPs increased. We hold the number of GWAS panel individuals constant at $M = 1,000$ so as L increases the ratio of M/L decreases. The error in \hat{U}_1 does not depend on the population model as the depth of the deepest split is constant across models. Error in \hat{F}_{Gr} increases as overlap between panels decreases. c, f, i) Bias in \hat{q} computed from using the estimators as covariates in the GWAS follows from the error in the estimators themselves. We also include bias in \hat{q} when effect sizes were estimated using a linear mixed model.

compute PCs in the test panel and project them into the GWAS panel. However, the best that this approach could possibly do would be to equal the performance of the direct estimator, \hat{F}_{Gr} . To see this, consider that if PC1 of the test panel were estimated with perfect accuracy, then it would be identical to the test vector, T , and so projecting it into the GWAS panel would be identical to computing \hat{F}_{Gr} . Noise in the test panel PCs would then only serve to reduce the accuracy further, increasing the amount of residual bias above that of the direct estimator approach. A more promising approach is a joint PCA, in which we compute PCs for a combined sample that includes both the GWAS and test panel individuals, and then include the GWAS panel individuals' position on these joint PCs when estimating effect sizes (see Fig. 3). This approach increases the amount of information that PCA has about axes of population structure that are shared between the two populations (e.g. in the toy models considered here, it increases the amount of information about the internal branch that is shared between the two panels), which should lead to improved performance. However, it also increases the complexity of the structure within the panel from which the PCs are derived, which could create complications.

For the models depicted in Fig. 2a and d, we found that this additional structure is not an issue. The first joint sample PC, when detectable, captures the deepest split in the population model. And because the combined sample size is twice as big as the

GWAS panel sample size, PC1 is estimated with less noise, resulting in improved bias reduction relative to the GWAS only PCA (Fig. 3a and b).

However, for the model depicted in Fig. 2g, the first joint sample PC did not always capture the deepest split. In this model, there is very little drift between the first and second population splits, and estimation noise can lead to inconsistency in which sample PC corresponds to each theoretical PC. As a result, including only joint PC1 resulted in unexpected behavior: the bias switched signs as L increased (Fig. 3e), and the absolute value of the bias was not a monotonic function of L (Fig. 3f). While the (absolute) bias did ultimately converge toward zero as L increased, it is notable that joint PC1 performed worse than PC1 of the GWAS panel alone for some large values of L (Fig. 3f), despite outperforming it at lower values. However, this shortcoming of the joint PCA approach can be overcome simply by including more PCs (in this case, 3 joint PCs is sufficient). The sign switching of the bias and nonmonotonic behavior of the absolute bias both still occur, but the absolute bias of the joint PC approach outperformed the GWAS panel only approach.

Grid simulations

Our exploration of the four-population toy models above is valuable for comparing the performance of different methods in settings where patterns of population structure are simple enough

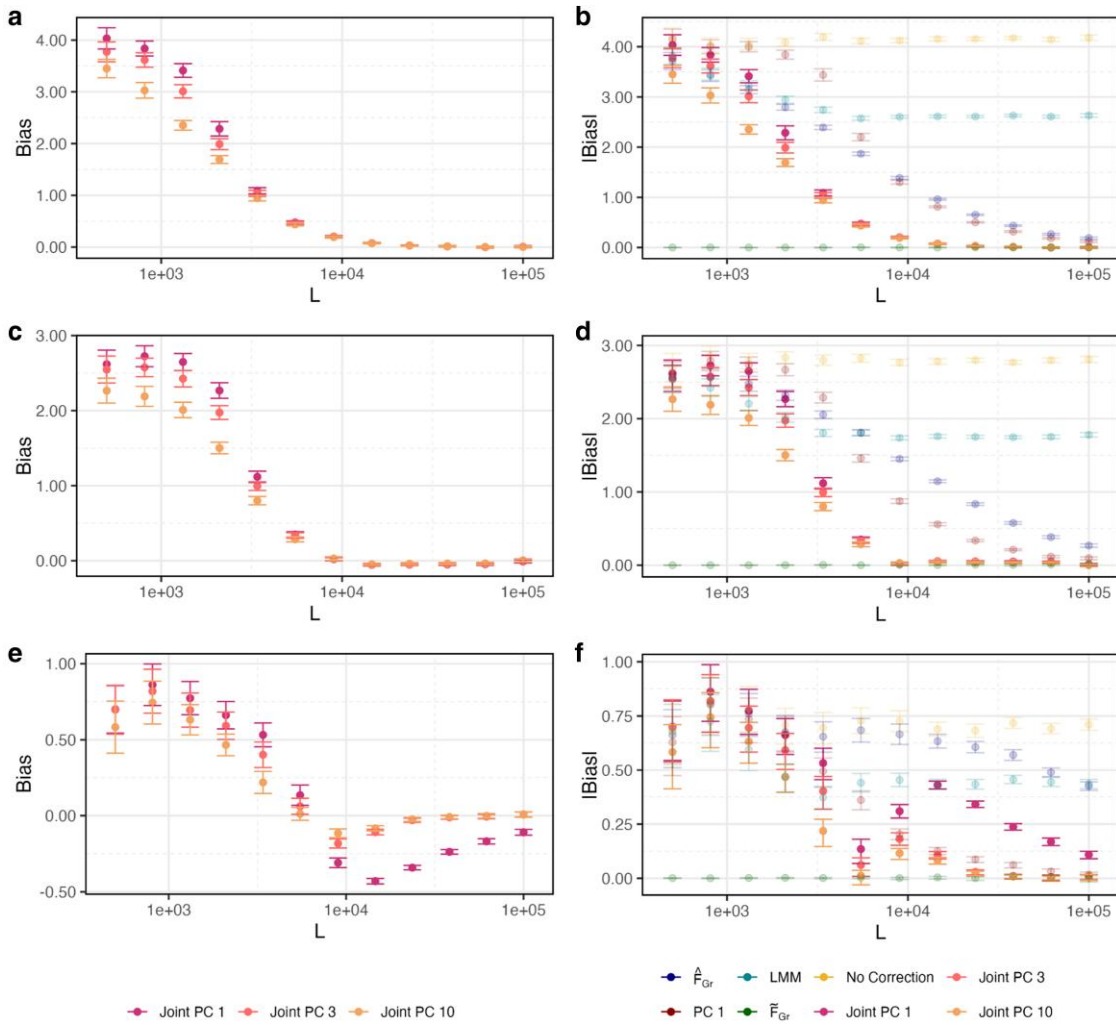


Fig. 3. Residual bias in \hat{q} when including top PCs from the combined GWAS and test panel as covariates in the GWAS. a) Signed bias for the population model in Fig. 2a when effect sizes are estimated using either the top 1, 3, or 10 sample PCs from the combined GWAS and test panels as covariates in the GWAS. b) Absolute value of the bias and comparison of the joint PCA approach (solid dots) to the approaches in Fig. 2 (transparent dots). c, d) Same plots for the population model in Fig. 2d and (e, f) for the model in Fig. 2g.

that they can be understood intuitively. However, to illustrate the theoretical observation that controlling for \hat{F}_{Gr} is sufficient to protect against confounding in polygenic score association tests, even when there is confounding along other axes, we need a model with more complex structure.

To this end, we conducted another set of coalescent simulations under a symmetric two-way migration model on a six-by-six lattice grid, building off of a framework developed by Zaidi and Mathieson (2020). We sampled an equal number of individuals per deme to comprise both the GWAS and test panels, with total sample sizes $N, M = 1,440$. We then simulated three different distributions of purely environmental phenotypes across the GWAS panel individuals. For each scenario, we estimated effect sizes, ascertained associated sites, and tested for an association between polygenic score and latitude, longitude, or membership in the single confounded deme, depending on the example. We compared the performance of the direct estimator approach (i.e. \hat{F}_{Gr}) with standard tools: the top 10 GWAS sample PCs, and a standard linear mixed model. In each case, we used the same set of $L = 20,000$ SNPs that were found at a frequency greater than 1% in both panels for both estimators to compute all estimates of population structure (see the Discussion section for remarks on the joint PCA approach in the grid model).

For the first example, the confounder, c , was a linear function of an individual’s position on the latitudinal axis (Fig. 4a). When we estimated effect sizes with no correction for population structure, the spatial distribution of the resulting polygenic scores reflected the distribution of the environmental confounder. Consequently, an association test using latitude as the test vector was biased. However, including \hat{F}_{Gr} or the top 10 sample PCs as covariates in the GWAS model was sufficient to produce effect sizes that were unbiased with respect to the latitudinal genotype contrasts in the test panel, ensuring resulting association test was unbiased. The linear mixed model did reduce stratification compared to the uncorrected polygenic scores but did not fully protect the association test. This outcome is consistent with our argument above that for this example the LMM is too constrained in its assumption that the confounding along each PCs is proportional to the corresponding eigenvalue.

In the second example, we simulated confounding along the diagonal, resulting in uncorrected polygenic scores that were correlated with both latitude and longitude in the test panel and an association test that was biased along both axes (Fig. 4b). When we computed \hat{F}_{Gr} using latitude as the test vector, the resulting effect sizes were uncorrelated with latitudinal genotype contrasts, but remained susceptible to bias along other axes (e.g. longitude).

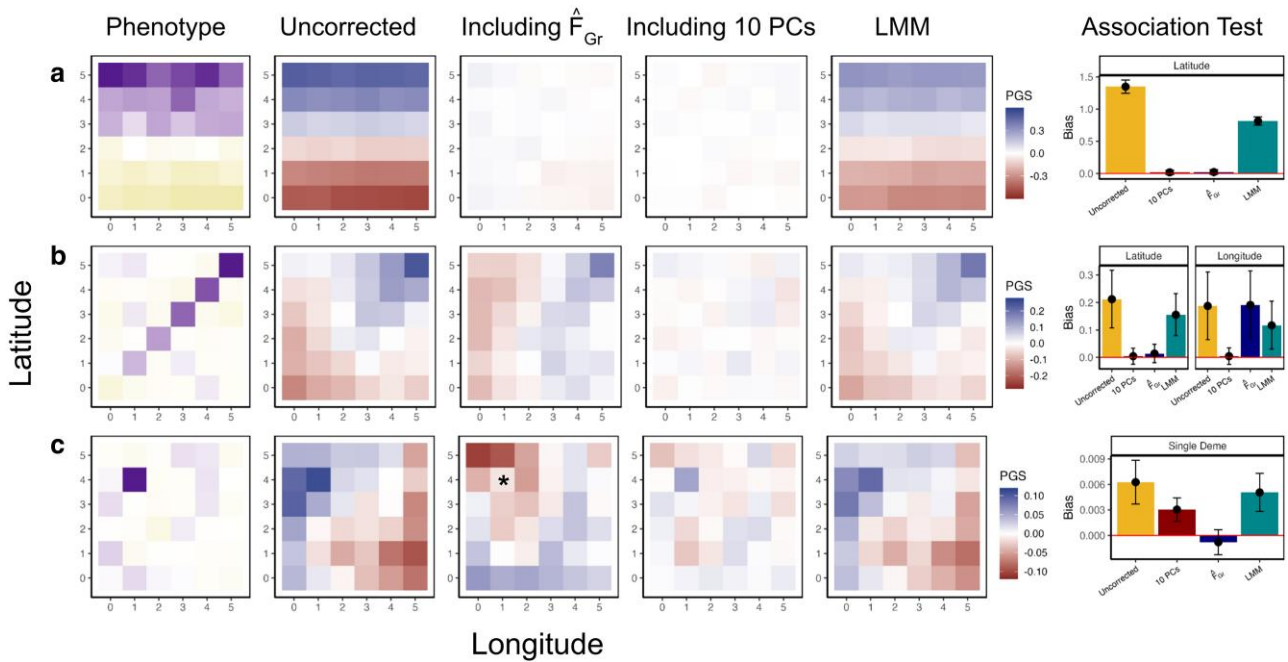


Fig. 4. Stratification bias in more complex demographic scenarios. GWAS and test panel individuals were simulated using a stepping-stone model with continuous migration. In the GWAS panel, the phenotype was nonheritable and stratified along either latitude a), the diagonal b), or in a single deme c). When effect sizes were estimated in a GWAS without correction for stratification, polygenic scores constructed in the test panel recapitulated the spatial distribution of the confounder (second column). Including \hat{F}_{Gr} (test vector is latitude for a and b, belonging to * deme for c) in the GWAS model eliminated bias in polygenic scores along the test axis (third column) which is also reflected in the association test bias (sixth column). We also compare the direct approach to including the top 10 PCs (fourth column) and LMMs (fifth column).

This example illustrates the targeted nature of this approach, as using effect sizes from a GWAS including \hat{F}_{Gr} ensures that the association test for the corresponding test vector are unbiased (assuming \hat{F}_{Gr} is well estimated) but does not remove all bias along other axes. Including 10 sample PCs protected both the latitudinal and longitudinal association tests, while the LMM did not.

In the third example, we simulated an increased environmental effect in a single deme, a scenario that induced a more complex spatial pattern in the uncorrected polygenic scores (Fig. 4c), and that previous work has shown to be difficult to correct for with standard tools (Mathieson and McVean 2012; Zaidi and Mathieson 2020). We then took the test vector to be an indicator for whether the test panel individuals were sampled from the deme with the environmental effect or not, and computed \hat{F}_{Gr} using these contrasts. In this scenario, including \hat{F}_{Gr} as a covariate in the GWAS resulted in an unbiased test statistic. In contrast, the top ten sample PCs and the LMM did not.

Quantifying error in population structure estimators

Next, we wanted to better understand the role that error in our population structure estimators played in these simulations. In contrast to the four-population toy model, it is not straightforward to compute \hat{F}_{Gr} given our underlying demographic model, particularly for the case of testing a single deme against all others. As a result, we could not directly measure the error in \hat{F}_{Gr} or sample PCs as estimators of \hat{F}_{Gr} . Instead, we used the fact that under this demographic model individuals within a deme are exchangeable, and therefore have the same values of both \hat{F}_{Gr} and theoretical PCs. This allowed us to estimate the error in \hat{F}_{Gr} by computing the fraction of the total variance in \hat{F}_{Gr} that can be attributed to variance of individual values within demes and variance of deme means across replicates (see Direct estimator section). For the PCs, the relationship between the order of the underlying

theoretical PCs and the order of the sample PCs may differ across replicates due to the noisiness of the sample PCs, so it is not obvious how to compute the variance of the deme means across replicates. We therefore used only the within deme variances, so our estimates of the error for the PCs are technically estimates of a lower bound on the error (see Principal components section). However, we note that for our estimation of the error in \hat{F}_{Gr} , we found that the variance within demes was by far the larger contributor, so we expect this to be a relatively tight bound. We then varied the number of SNPs used to compute our estimators of population structure from $L = 20,000$ down to $L = 2,000$, and observed how differences in the estimated error of our population structure estimators translate to differences in the amount of bias in the polygenic score association test statistic.

In Fig. 4a and b, \hat{F}_{Gr} corresponds to latitude, so we expect it to be captured by the top two population PCs (Novembre and Stephens 2008). For $L = 20,000$ (the number of SNPs used in Fig. 4), we estimated the lower bound on the error in sample PCs 1 and 2 to be 0.011. Across the range of L values we tested, the estimated bound was no greater than 0.053 (Fig. 5a), and including 10 PCs consistently removes bias in \hat{q} (Fig. 5b). Similarly, we estimated the error in \hat{F}_{Gr} for latitude to be 0.012 when $L = 20,000$ with a maximum of 0.059 when $L = 2,000$. Although these estimates are nearly identical to the values we observe for the first two PCs, the bias in \hat{q} is slightly higher (Fig. 5b). We observed a similar result in the 4 population toy model (Supplementary Fig. S1), so this may be the same phenomenon, or it may be that PCs 3–10 are capturing some of the residual latitudinal signal that is not captured by the first two.

Next, we explored the role of error in our population structure estimators for the more difficult single deme test and confounder case (Fig. 4c). We again computed the error in \hat{F}_{Gr} as we varied L , with estimates that ranged from 0.04 to 0.18 as L decreased (Fig. 5a). For larger values of L , the error was small enough that

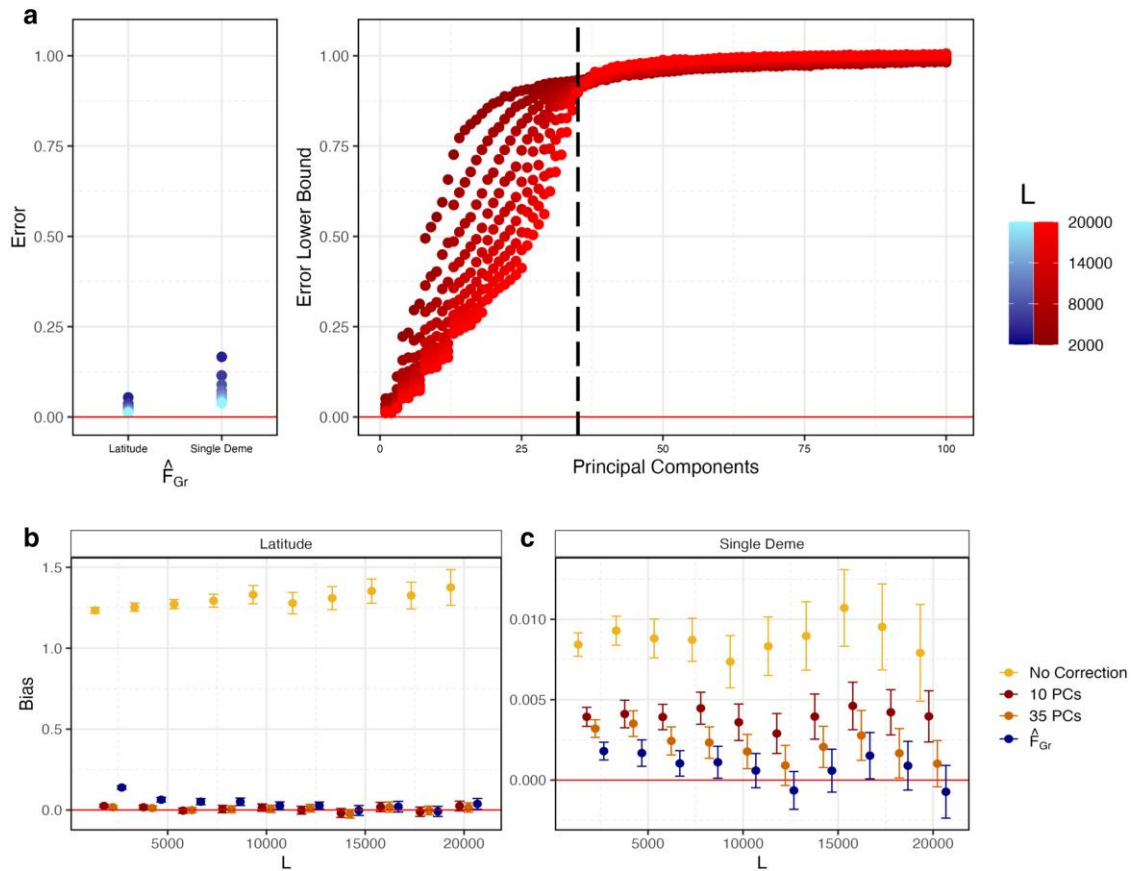


Fig. 5. Quantifying error in estimates of \hat{F}_{Gr} and sample PCs for the six-by-six stepping-stone demographic model used in Fig. 4, individuals within a deme are exchangeable and have the same \hat{F}_{Gr} and theoretical PC value. Therefore, we used variation within demes to estimate the error in \hat{F}_{Gr} and a lower bound for the error in sample PCs (see [Direct estimator](#) and [Principal components](#) sections for details) for different values of L , holding $M = 1,400$ constant. The dashed vertical line indicates PC 35, the last theoretical PC we expect to capture real structure. b) When latitude was the test vector, both sample PCs and \hat{F}_{Gr} were well estimated and bias in \hat{q} was reduced. c) When a single deme indicator variable was the test vector, higher PCs are needed to capture \hat{F}_{Gr} . These sample PCs are not well estimated, and residual bias remains when 35 PCs are used for most values of L .

confidence intervals on the bias overlapped zero, but this was not true when we reduced L so that the error was larger (Fig. 5c). As shown above (Fig. 4c), with $L = 20,000$, we found that 10 PCs were not sufficient to remove the bias. This could either be because \hat{F}_{Gr} is not captured by the top 10 theoretical PCs, or because \hat{F}_{Gr} can be captured by 10 theoretical PCs, but the sample PCs are too noisy as estimates. Given that there are 36 demes in our simulations and that individuals within demes are exchangeable, only the top 35 theoretical PCs capture real population structure, while the rest correspond to sampling variance. As a result, if the sample PCs are sufficiently well estimated, then only 35 should be required to remove the bias. In practice, we find that when using 35 PCs for larger values of L , the bias is closer to zero than it was with 10 PCs, but the confidence intervals still do not always overlap zero, and the bias is generally greater than it is when we use our direct estimator, \hat{F}_{Gr} (Fig. 5c). As expected, the performance with 35 sample PCs decreases further with an increase in the error, but is always intermediate between 10 PCs and \hat{F}_{Gr} . All of this is consistent with the observation that the error in the higher sample PCs (i.e. 11–35) is extremely high across the range of L values we explored (Fig. 5a).

PCs succeed by capturing structure relevant to the test, not by capturing the confounder

Finally, to the extent that the PCs did succeed in removing bias in our simulations, we wanted to understand whether it was

because they successfully captured the confounder or because they captured the relevant axis of structure for the test (see [Relationship between \$\hat{F}_{Gr}\$ and PCA](#) section). To this end, for each of the three grid scenarios in the $L = 20,000$ case, we computed the cumulative proportion of variance in the confounder, c , that could be explained by the first J sample PCs, for J up to 100 (Fig. 6). We found that while the confounding axis was well captured by sample PCs 1 and 2 for latitude (Fig. 6a), it was not well captured by the top 10, 35, or even 100 PCs for the diagonal (Fig. 6b) or single deme confounders (Fig. 6e). In contrast, when we took our estimator, \hat{F}_{Gr} , as a proxy for \hat{F}_{Gr} , we found that the PCs explained a considerably higher fraction of the variance. For the first two cases, the test axis was latitude, so this is unsurprising. However, this was true even for the single deme case, and results from the fact that relatedness among adjacent demes leads to a smoothing effect (Supplementary Fig. S2), which makes \hat{F}_{Gr} easier for the PCs to capture.

Discussion

Interpreting patterns in the distribution of polygenic scores is difficult, especially when confounding cannot be ruled out. Because most well-powered GWAS are conducted on population samples where the relationship between genetic background, ancestry, and the environment is not well controlled, stratification bias remains a significant concern (Berg et al. 2019; Sohail et al. 2019;

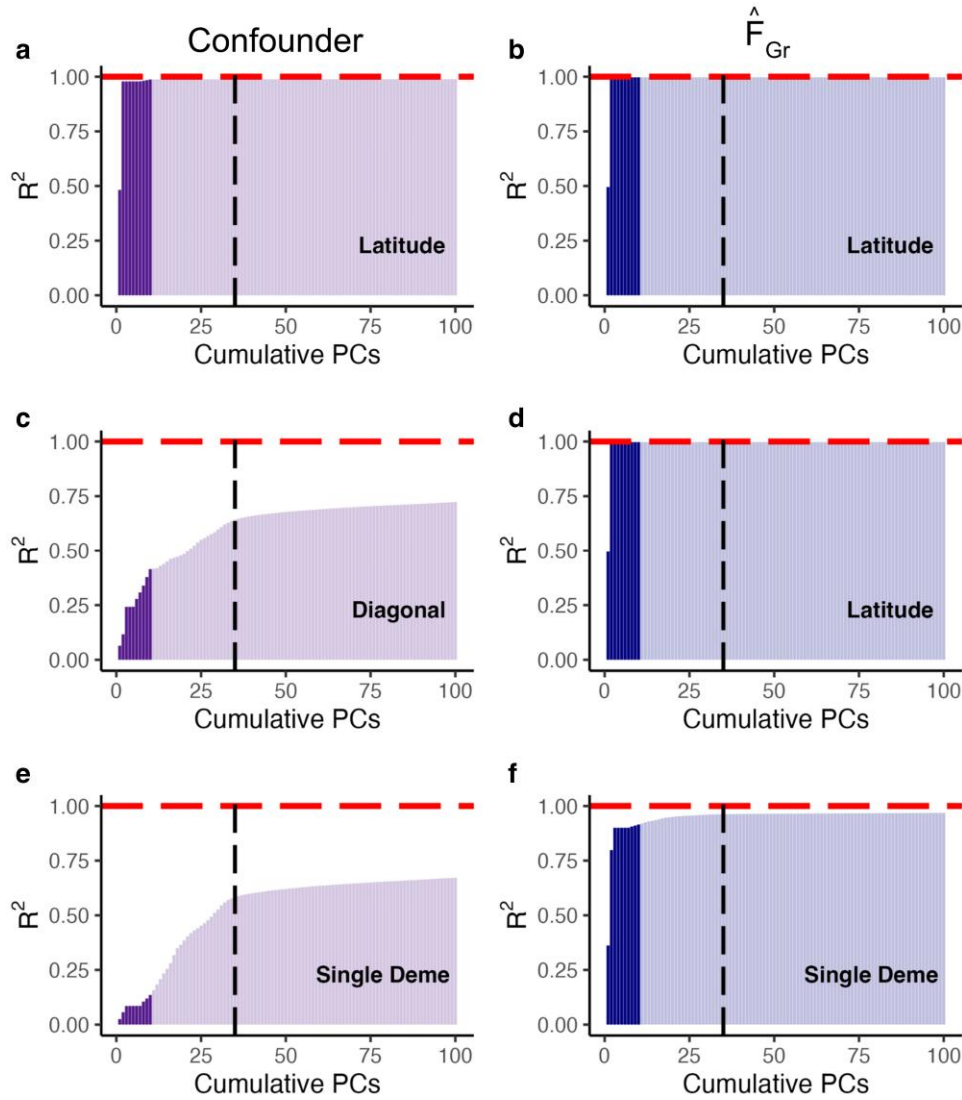


Fig. 6. Different patterns of confounding and \hat{F}_{Gr} are captured by different GWAS panel sample PCs. For the three possible combinations of confounding and polygenic score association tests in Fig. 4, we plot the variance in either the confounder (a, c, e) or \hat{F}_{Gr} (b, d, f) explained by cumulative GWAS panel sample PCs, with the top 10 PCs highlighted in a darker color. As \hat{F}_{Gr} is unknown for this model, we estimated the error in \hat{F}_{Gr} as 0.011 and 0.04 for latitude and the single deme, respectively, and therefore assume it is a decent proxy for \hat{F}_{Gr} . In (a, b), both the confounder and \hat{F}_{Gr} (and therefore \hat{F}_{Gr}) represent variation along latitude and are well captured by the first two PCs. For (c, d) the confounder varies along the diagonal and these individual deme level differences are not well captured by top sample PCs. In contrast, the test vector is still latitude and \hat{F}_{Gr} is again well captured by PCs 1 and 2. Finally, in (e, f), both the confounder and the test vector represent membership in a single deme and therefore not as well captured by top sample PCs.

Cox et al. 2023; Ding et al. 2023; Tan et al. 2024). Here, under standard modeling assumptions, we characterize patterns of stratification bias in the distribution of polygenic scores as a function of the expected genetic similarity between GWAS and test panels. We find that for any given polygenic score association test axis, the amount of bias in the association test statistic depends on the strength of stratification along a single axis of population structure in the GWAS panel, where each individual’s position on this axis is given by the product of the expected cross-panel GRM and the test axis (i.e. $\hat{F}_{Gr} = \hat{F}_{GX}T$).

The ability to ensure a given polygenic score association test is unbiased in practice, therefore depends on the accuracy with which we can model \hat{F}_{Gr} via covariates included in the GWAS. For the standard PCA-based approach, the inconsistency of the sample PCs as estimators of population structure is therefore a plausible explanation for the signatures of residual stratification bias that have been reported across many GWAS datasets (Berg et al. 2019; Sohail et al. 2019; Ding et al. 2023), though such signals

might also arise simply from not including enough PCs, even if they are well estimated. The inconsistency of the sample PCs as estimators is a well-known result in random matrix theory (Baik et al. 2005; Johnstone and Paul 2018), and we are not the first to notice the connection to stratification bias in GWAS and polygenic scores (Bloemendal and Chen 2019), but the phenomenon is not widely acknowledged in the GWAS literature.

Based on our theoretical analysis, we also propose a direct estimator of the target axis of population structure using the test panel genotype data. We show that, under the optimal condition of complete overlap in structure between panels and a large sample size in the test panel (Fig. 2a and c), this estimator outperforms, or at least equals, the standard PCA-based estimator. However, a limitation of this direct approach is that the performance relative to PCA degrades as the amount of variance explained by \hat{F}_{Gr} in the GWAS panel decreases (Fig. 2b and c). As a result, it is best suited for cases where the GWAS cohort and test panels are drawn from the same sample, ensuring a

high degree of overlap in structure between panels and making \hat{F}_{Gr} easier to estimate accurately. We also expect this method to perform best when the test panel is large relative to the amount of variance explained by the test vector, so that the relevant genotype contrasts, r , are well estimated.

We also considered a joint PCA approach, in which we compute PCs on a the combined GWAS and test panels and then use the positions of GWAS panel individuals on these axes as covariates in the estimation of effect sizes. In the toy model examples to which we applied this approach, it outperformed all other methods, provided we properly accounted for the increase in the complexity of structure in the combined panel by increasing the number of PCs. While we did not apply this joint PCA method in the more complex grid simulations, it is clear that doing so would offer an improvement over the standard GWAS panel PCA approach. This is because in our grid simulations, the test panel is composed of exactly the same number of individuals from each deme as in the GWAS panel. The *theoretical* PCs of the two panels are therefore identical, and computing PCs on the joint panel in these simulations would be identical to computing PCs on a GWAS panel with twice the sample size. It is less obvious how the joint PCA approach would perform in even more complex scenarios, where both the GWAS and test panels have complex structure but are not sampled in the same way. The increase in total sample size would be expected to increase the accuracy of the PCs, but the increase in the complexity of the structure may create complications. Based on our simulations of the four-population toy model with the deepest population splits (i.e. Fig. 2g), it is plausible that the increased complexity can always be overcome simply by including enough PCs, even if they are individually not well-estimated. Assuming this is the case, and because we expect that practitioners would always try to include all PCs that explain any significant amount of structure, our results suggest that the joint PCA approach is the optimal choice among those that we considered. A more exhaustive exploration of this approach is beyond our present scope but would be valuable given our results.

Several recent papers have proposed additional alternative methods for improved control of population structure in GWAS and polygenic scores. These proposals include: (1) using PCs of rare variants (as opposed to common variants) (Zaidi and Mathieson 2020), (2) using PCs of external reference datasets in addition to the PCs of the GWAS panel (Sarmanova et al. 2020), and (3) using local ancestry assignments (in lieu of global linear estimators) (Hu et al. 2025). Our results highlight the importance of developing tools to better estimate the error in population structure estimates (Haag et al. 2025), and it would be interesting to assess the merits of these alternative methods through this lens. Ideally, future methods development might allow each set of GWAS summary statistics to be accompanied by statistics summarizing the accuracy of the population structure estimates used to control for stratification. These estimates could then be used in downstream analyses to provide quantitative statements about the extent to which a particular polygenic score association test is or is not protected from stratification bias. We also note that tests for association between polygenic scores and axes of ancestry variation are closely related to bivariate LD score regression, which combines effect estimates for one trait and frequency/genotype contrasts from an independent dataset (Bulik-Sullivan B et al. 2015; Field et al. 2016; Berg et al. 2019). Previous work in the context of polygenic selection tests raised concerns about spurious inflation of the LD score slope due to background selection (Berg et al. 2019). It would be interesting to revisit this issue more fully in light of our present results.

Our model contains several elements that differ from reality. It is worth highlighting what these are, and what their effects are. For example, our model ignores linkage between sites and assumes that we use marginal effects, rather than jointly estimated effects, to construct our polygenic scores. First, linkage between sites does not change the fundamental point that controlling for \hat{F}_{Gr} is sufficient to render the effect size estimates uncorrelated with the test panel genotype contrasts under the null. This is holds whether effects are estimated marginally or jointly. However, in practice, we still prefer to estimate effects jointly, for at least two reasons. The first is simply because doing so increases the accuracy of the polygenic score, thereby boosting our power. The second reason is that, in the presence of residual stratification (e.g. if our estimator, \hat{F}_{Gr} , has high error), polygenic scores constructed with jointly estimated effects should be less biased than those constructed using marginal effects. This is because when effect sizes are estimated marginally, each site experiences the entirety of the stratification effect, and therefore gets a “full dose” of it. The stratification effect is then being added into the polygenic score multiple times across SNPs. This is why we find the bias in the polygenic score association test statistic to be proportional to the number of loci included in the polygenic score. In contrast, if effects were estimated jointly, the stratification effect will be spread out more evenly across sites, and so we would expect the effect on the polygenic score to be less extreme, but not eliminated.

We also ignored the ways that stratification bias can impact the ascertainment process. In short, if biased effect sizes are used to ascertain sites, this can generate biases in the frequency distribution of ascertained sites (see e.g. Figure 6 in Zaidi and Mathieson 2020). However, successfully controlling for \hat{F}_{Gr} is also sufficient to eliminate this source of bias in a given polygenic score association test. We provide a more detailed explanation in [Supplementary Section S8](#). Another issue is that, throughout our simulations, we often estimate effect sizes while attempting to control for stratification only along the target axis of the test. We do this to highlight our main point that controlling for the target axis is sufficient to render the association test unbiased, but readily acknowledge that it does not address all of the negative consequences of stratification bias. For example, bias along other axes will function as additional noise in the process of ascertaining SNPs, and in the polygenic scores themselves, which is expected to reduce power. Therefore, it is still desirable to include top PCs or use an LMM alongside \hat{F}_{Gr} , even when \hat{F}_{Gr} is well estimated.

We also wish to emphasize that our results are relevant to a broader set of analyses than those explicitly covered by our model. For example, with a slight shift in perspective, our model is applicable to studies that use GWAS summary statistics together with coalescent methods to test for signals of directional polygenic selection (Field et al. 2016; Edge and Coop 2019; Song et al. 2021; Stern et al. 2021). The key recognizing that such methods use patterns of haplotype variation to estimate genotype contrasts between the sampled present day individuals and a set of unobserved ancestors, and then ask whether these estimated genotype contrasts correlate with effect size estimates for a trait of interest. Thus, in such an analysis, there also exists an \hat{F}_{Gr} that describes the extent to which individuals in the GWAS panel are more closely related to the present-day sample or the hypothetical ancestors. For both the coalescent approaches, as well as methods relying on direct comparisons of polygenic scores, both the evolutionary hypothesis being tested and the degree of susceptibility to bias follow directly from the set of genotype contrasts used in the test. Some prior work has suggested that certain

coalescent methods for testing for polygenic selection are more robust to stratification bias than others (Song et al. 2021; Stern et al. 2021), but our results show that this cannot be true: two methods that test the same evolutionary hypothesis using the same set of estimated effect sizes necessarily have the same susceptibility to stratification bias. If there are differences in robustness to stratification bias among methods, then this must come either from changing the evolutionary hypothesis being tested or from overall differences in the statistical power of the methods.

Finally, we note that even if \tilde{F}_{Gr} is known exactly, the interpretation of the results of polygenic score association tests is limited by the many assumptions that must be made in any polygenic score analysis (Novembre and Barton 2018). For example, these analyses use effect sizes estimated in a one set of genetic and environmental background, and there is no guarantee that the effects will be the same in other backgrounds. Effect size heterogeneity can cause many difficulties with the interpretation of positive associations between polygenic scores and axes of population structure (as several papers have noted, see Novembre and Barton 2018; Rosenberg et al. 2019; Harpak and Przeworski 2021). Another difficulty with interpretation arises from allelic turnover (Carlson et al. 2022) and differences in tagging across populations. A given polygenic score will have less power to detect differences between populations that are genetically more distant from the GWAS panel, and this can lead to a biased picture of how selection has actually affected the trait across populations (Yair and Coop 2022). However, none of these phenomena are expected to generate false signals of directional selection where none exists. This is because the fact that the effect size might vary across populations has no impact on the correlation between the effect size measured in only one of the populations and patterns of allele frequency differentiation among populations. One subtle caveat to this claim is that certain forms of directional interaction effects (e.g. directional dominance) could, in principle, create correlations between the direction of recent allele frequency change on the lineage leading to the GWAS panel individuals and the average effect as estimated under an additivity assumption. This would violate the null model. However, there is little evidence for substantial interaction variance among common variants in human complex traits, so this is unlikely to be an issue in practice.

Moving beyond the specific issue of associations between polygenic scores and population structure axes, we note that GWAS can also be impacted by other forms of genetic confounding beyond the simple associations between ancestry and genetic background that we consider here, include dynastic effects, assortative mating, and stabilizing selection (Veller and Coop 2023). Therefore, while our results provide a pathway to a more rigorous approach for protecting against stratification bias in polygenic score association tests, addressing a known problem in their implementation, continued care in the interpretation of polygenic score analyses is always warranted.

Materials and methods

Simulating genotypes

We used *msprime* (Kelleher et al. 2016) to simulate genotypes under different models with 100 replicates per model. The first model, shown in Fig. 1, has two population splits, 200 and 100 generations in the past, for a total of 4 present day populations. We fix the population size for all present and past populations to 10,000 diploid individuals. We then sample 5,000 individuals per population and create two configurations of GWAS and test panels

($N, M = 10,000$) based on the diagrams in Fig. 1a and c. For every model replicate, we simulate a large number of independent sites and downsample to $L = 10,000$ SNPs with minor allele frequency (MAF) > 0.01 in both GWAS and test panels. We use these genotype simulations for Fig. 1 and Supplementary Fig. S3. When the populations in the GWAS and test panel are nonsister (i.e. Fig. 1a), the average within panel F_{ST} (Bhatia et al. 2013) was 0.01, whereas in the configuration in Fig. 1c the average F_{ST} was 0.005.

For Fig. 2, we use the same model setup but adjust the split times to 12/0, 12/4, and 12/10 generations in the past for population models A, B, and C, respectively. The average F_{ST} for the overlapping structure scenario is approximately 0.0006. To reduce computational burden, we scale down the sample size to 1,000 individuals per panel (500 per population). We simulate a large number of independent SNPs and downsample to L sites (MAF > 0.01 in both panels) which we vary from 500 to 100,000.

For Fig. 4, we use a model, modified from Zaidi and Mathieson (2020), that is a 6×6 stepping-stone model where structure extends infinitely far back with a symmetric migration rate of $m = 0.01$. We fix the effective population size to 1,000 diploid individuals and sample 80 individuals per deme, which we split equally into GWAS and test panels ($N, M = 1,440$). As above, we simulate a large number of independent SNPs and down-sample to $L = 20,000$ SNPs with MAF > 0.01 in both panels.

Simulating phenotypes

To study the effect of environmental stratification on association tests, we first simulated nongenetic phenotypes for an individual i in the GWAS panel as $y_i \sim N(0, 1)$. In our discrete 4 population models, we then generate a phenotypic difference between populations by adding Δ_{AB} to y_i for individuals in population B. For Fig. 1, we vary Δ_{AB} from 0 to 0.1 standard deviations. In order to compare across models and values of $\frac{L}{M}$ in Fig. 2, we compute Δ_{AB} as $\frac{5,000}{0.05 \times L}$.

In our grid simulations, we generated three different phenotypic gradients where the largest phenotypic shift was always equal to Δ . To generate a latitudinal gradient (Fig. 4a), we added $\frac{\Delta}{4}$ to y_i for individuals in row 1, $2\frac{\Delta}{4}$ for individuals in row 2, etc. For Fig. 4b, we generated a gradient along the diagonal by adding $\frac{\Delta}{4}$ to the phenotype for individuals in deme (1,1), $2\frac{\Delta}{4}$ for individuals in deme (2,2), etc. For Fig. 4c, we shifted the phenotype of individuals in deme (1,4) by Δ . For all grid simulations in Fig. 4, we set $\Delta = 0.2$. In order to compare across values of L in Fig. 5, we compute Δ as $\frac{60}{0.015}$.

To study the effect of controlling for stratification in cases where there is a true signal of association between polygenic scores and the test vector (Supplementary Fig. S3), we used our 4 population demographic model and followed the protocol outlined in Zaidi and Mathieson (2020) to simulate a neutral trait with $h^2 = 0.3$. We first randomly select 300 variants to be causal and sample their effect sizes from $\beta_\ell \sim N(0, \sigma_\ell^2 [p_\ell(1-p_\ell)]^\alpha)$, where σ_ℓ^2 is a frequency independent scale of the variance in effect sizes, p_ℓ is allele frequency in the GWAS panel, and α is a scaling factor controlling the relationship between allele frequency and effect size. We set $\alpha = -0.4$ and $\sigma_g^2 = \sigma_\ell^2 \sum_{\ell=1}^{200} [2p_\ell(1-p_\ell)]^{\alpha+1} = 0.3$.

To simulate a signal of true difference in polygenic score in the test panel, we calculate the frequency difference $p_{D,\ell} - p_{C,\ell}$ at all 300 causal sites in the test panel and flip the sign of the effect sizes in the GWAS panel such that $p_D - p_C > 0$ and $\beta_\ell > 0$ with probability θ . θ therefore controls the strength of the association with $\theta = 0.5$ representing no expected association and $\theta = 1$ representing the most extreme case where trait increasing alleles are always at a higher frequency in population D. We use θ ranging from 0.5 to

0.62. We then draw the environmental component of the phenotype $e_{i,k} \sim N(0, 1 - h^2)$ and generate an environmental confounder by adding $\Delta_{AB} \in \{-0.1, 0, 0.1\}$ to $e_{i,k}$ for individuals in population B.

Computing covariates

For each polygenic score association test, we computed \hat{F}_{Gr} . We first construct T as either population ID, latitude, or the single deme of interest, depending on the simulation. Given this test vector, we compute $r = \mathbf{X}^T T$ using the `plink2` (Chang et al. 2015) function `--glm`. Finally we compute \hat{F}_{Gr} (see Equation 24) using `--score` in `plink2`, taking care to standardize by the variance in the GWAS panel genotypes. Additionally, we used `plink2` (Chang et al. 2015) `--pca` or `--pca approx` to compute sample PCs from the GWAS panel genotype matrix.

Genome-wide association study

For each set of phenotypes, we first carried out three separate marginal association GWASs using the regression equations below,

- 1) $y = \beta_\ell G_\ell + \epsilon$
- 2) $y = \beta_\ell G_\ell + \omega \hat{F}_{Gr} + \epsilon$
- 3) $y = \beta_\ell G_\ell + \omega_1 \hat{U}_1 + \dots + \omega_j \hat{U}_j + \epsilon$.

Additionally, we conducted a fourth GWAS, $y = \beta_\ell G_\ell + \omega \hat{F}_{Gr} + \epsilon$, for the discrete 4 population model where \hat{F}_{Gr} is known. All fixed-effect GWASs were done using the `plink2` (Chang et al. 2015) function `--glm`. Additionally, for Figs. 2 and 4, we estimate effect sizes using a linear mixed with the `--mlma` function in the GCTA software (Yang et al. 2011, 2014).

We then ascertain S SNPs based on minimum p -value for inclusion in the polygenic score. For Fig. 1 and 4, we set $S = 300$. In order to compare across values of $\frac{1}{M}$ in Figs. 2 and 5, we set $S = 0.05 \times L$ and $S = 0.015 \times L$, respectively. For Supplementary Fig. S3, we use estimated effect sizes at the 300 causal sites (bottom row) and the top 300 sites ascertained on p -value (top row).

Polygenic score association test

We construct polygenic scores for the individuals in the test panel as $\hat{Z}_i = \sum_{\ell=1}^S \hat{\beta}_\ell X_{i,\ell}$, where $\hat{\beta}_\ell$ is the estimated effect size from the joint model and $X_{i,\ell}$ is the mean-centered genotype value for the ℓ th variant.

For each replicate, we then compute the test statistic $\hat{q} = \frac{1}{N} \hat{Z}^T T$ by multiplying the vector of polygenic scores for individuals in the test panel by the test vector. Finally, we compute the bias in \hat{q} across each set of 100 replicates as $\mathbb{E}[\hat{q} - q]$.

Estimating the error in population structure estimators for the grid model

Direct estimator

Consider that the value of $\hat{F}_{Gr,ij}$, the entry of \hat{F}_{Gr} for the i th individual in the j th deme, can be decomposed as

$$\hat{F}_{Gr,ij} = \left(\hat{F}_{Gr,ij} - \overline{\hat{F}_{Gr,j}} \right) + \left(\overline{\hat{F}_{Gr,j}} - \hat{F}_{Gr,j} \right) + \hat{F}_{Gr,j}, \quad (30)$$

where $\overline{\hat{F}_{Gr,j}} = \frac{1}{m_j} \sum_i^{m_j} \hat{F}_{Gr,ij}$ is the empirical average of $\hat{F}_{Gr,ij}$ within deme j (m_j is the number of individuals in deme j), and $\hat{F}_{Gr,j}$ is the entry of the true population structure axis \hat{F}_{Gr} , for all individuals in deme j . Individuals within demes are exchangeable in our model, so the deviations $(\hat{F}_{Gr,ij} - \overline{\hat{F}_{Gr,j}})$ and $(\overline{\hat{F}_{Gr,j}} - \hat{F}_{Gr,j})$ both represent

sources of error in our estimator. The fraction of variance in \hat{F}_{Gr} that is attributable to error is therefore

$$\text{error} = \frac{\mathbb{E}_j \left[\text{Var}_i \left(\hat{F}_{Gr,ij} - \overline{\hat{F}_{Gr,j}} \right) \right] + \text{Var}_j \left(\overline{\hat{F}_{Gr,j}} - \hat{F}_{Gr,j} \right)}{\text{Var}(\hat{F}_{Gr})}. \quad (31)$$

We can estimate $\mathbb{E}_j[\text{Var}_i(\hat{F}_{Gr,ij} - \overline{\hat{F}_{Gr,j}})]$ as

$$\frac{1}{H} \sum_h \frac{1}{J} \sum_j \frac{1}{m_j - 1} \sum_i^{m_j} \left(\hat{F}_{Gr,ijh} - \overline{\hat{F}_{Gr,jh}} \right)^2, \quad (32)$$

where h indexes replicate simulations and H is the total number of replicates ($H = 100$ in our case), J gives the total number of demes (36 in our case), m_j is the number of individuals in deme j , and

$$\overline{\hat{F}_{Gr,jh}} = \frac{1}{m_j} \sum_i^{m_j} \hat{F}_{Gr,ijh} \quad (33)$$

is the empirical mean entry for deme j in replicate h .

To estimate the contribution of variance in the per-deme means, we compute the variance across replicates for a given deme and then take the average of these values across demes:

$$\frac{1}{J} \sum_j \frac{1}{H - 1} \sum_h \left(\overline{\hat{F}_{Gr,jh}} - \frac{1}{H} \sum_\ell^H \overline{\hat{F}_{Gr,j\ell}} \right)^2. \quad (34)$$

(here, the sums over ℓ and h are both sums over replicates—one for the mean, and one for the variance—but we use different letters to avoid confusion).

The denominator, in turn, can be estimated straightforwardly as

$$\frac{1}{M - 1} \sum_i^M \left(\hat{F}_{Gr,i} - \frac{1}{M} \sum_\ell^M \hat{F}_{Gr,\ell} \right)^2, \quad (35)$$

where we now use ℓ to index individuals within the mean calculation. Our estimate of the error is then given by summing (32) and (34) and dividing by (35).

Principal components

To estimate the error in the sample PCs, we follow similar steps, except that it is not obvious how to compute the variance of the per deme means, as the relationship between the order of the underlying population PCs and the sample PCs may differ across replicates due to the noisiness of the sample PCs. We therefore include only the variance among individuals within demes in our estimate of the error, which makes it an estimate of a lower bound on the error, rather than a direct estimate. The PCs are automatically standardized to have a variance of 1, so that for the k th PC, a lower bound on the error is given by

$$\text{error}_k > \mathbb{E}_j \left[\text{Var}_i \left(\hat{U}_{ijk} - \overline{\hat{U}_{jk}} \right) \right], \quad (36)$$

which we estimate as

$$\frac{1}{H} \sum_h \frac{1}{J} \sum_j \frac{1}{m_j - 1} \sum_i^{m_j} \left(\hat{U}_{ijkh} - \frac{1}{m_j} \sum_\ell^{m_j} \hat{U}_{ijkh} \right)^2. \quad (37)$$

Data availability

All of the code developed to produce the figures and simulations in this article is available in the github repository: <https://github.com/jgblanc/PGS-differences-confounding>. We used the existing software plink2 <https://www.cog-genomics.org/plink/2.0/>, GCTA <https://yanglab.westlake.edu.cn/software/gcta/#Overview>, msprime <https://tskit.dev/msprime/docs/stable/intro.html>, bcftools <https://samtools.github.io/bcftools/bcftools.html>, R <https://www.r-project.org/>, and python <https://www.python.org/>.

Supplemental material available at GENETICS online.

Acknowledgments

We would like to thank members of the Berg, Novembre, and Steinrücken labs, as well as members of the University of Chicago genetics community for helpful discussions and feedback during the development of this project. We thank Andy Dahl and Matthew Stephens in particular for many helpful conversations. Additionally, we thank Matthew Stephens, John Novembre, and Xuanyao Liu for support at all stages of this work, as well as Maggie Steiner, Vivaswat Shastry, and Maryn Carlson for help troubleshooting and additional insights. Finally, we thank Graham Coop, Jeff Spence, Arjun Biddanda, and Yuval Simons for comments on the manuscript.

Funding

This work was supported by the National Human Genome Research Institute (F31HG011821 to J.G.B.) and the National Institute of General Medical Sciences (R35GM151257 to J.J.B.).

Conflicts of interest

The author(s) declare no conflicts of interest.

Literature cited

Abdellaoui A, Dolan CV, Verweij KJH, Nivard MG. 2022. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nat Genet.* 54:1345–1354.

Abdellaoui A, Hugh-Jones D, Yengo L, Kemper KE, Nivard MG, Veul L, Holtz Y, Zietsch BP, Frayling TM, Wray NR, et al. 2019. Genetic correlates of social stratification in Great Britain. *Nat Hum Behav.* 3(12):1332–1342. doi:10.1038/s41588-022-01158-0.

Baik J, Ben Arous G, Piché S. 2005. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices.

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet.* 10(8):e1004412. doi:10.1371/journal.pgen.1004412.

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, et al. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife.* 8:e39725. doi:10.7554/eLife.39725.

Berg JJ, Zhang X, Coop G. 2017. Polygenic adaptation has impacted multiple anthropometric traits. Preprint, *Evolutionary Biology*.

Bhatia G, Patterson N, Sankararam S, Price AL. 2013. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* 23(9):1514–1521. doi:10.1101/gr.154831.113.

Bloemendal A, Chen C. 2019. PCA and stratification in GWAS/a primer on random matrix theory.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, Duncan L, Perry JRB, Patterson N, Robinson EB, et al. 2015. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 47(11):1236–1241. doi:10.1038/ng.3406.

Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 47(3):291–295. doi:10.1038/ng.3211.

Bulmer MG. 1971. The effect of selection on genetic variability. *Am Nat.* 105(943):201–211. doi:10.1086/282718.

Carlson MO, Rice DP, Berg JJ, Steinrücken M. 2022. Polygenic score accuracy in ancient samples: quantifying the effects of allelic turnover. *PLoS Genet.* 18(5):e1010170. doi:10.1371/journal.pgen.1010170.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience.* 4(1):s13742–015. doi:10.1186/s13742-015-0047-8.

Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, Cucca F, Okada Y, Chiang CW. 2020. Evidence of polygenic adaptation in sardinia at height-associated loci ascertained from the Biobank Japan. *Am J Hum Genet.* 107(1):60–71. doi:10.1016/j.ajhg.2020.05.014.

Cox SL, Nicklisch N, Francken M, Wahl J, Meller H, Haak W, Alt KW, Rosenstock E, Mathieson I. 2023. Socio-cultural practices affect sexual dimorphism in stature in Early Neolithic Europe. Pages: 2023.02.21.529406 Section: New Results.

Ding Y, Hou K, Xu Z, Pimplaskar A, Petteer E, Boulier K, Privé F, Vilhjálmsdóttir BJ, Olde Loohuis LM, Pasaniuc B. 2023. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature.* 618:774–781. doi:10.1038/s41586-023-06079-4.

Edge MD, Coop G. 2019. Reconstructing the history of polygenic scores using coalescent trees. *Genetics.* 211(1):235–262. doi:10.1534/genetics.118.301687.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, et al. 2016. Detection of human adaptation during the past 2000 years. *Science.* 354(6313):760–764. doi:10.1126/science.aag0776.

Guo J, Wu Y, Zhu Z, Zheng Z, Trzaskowski M, Zeng J, Robinson MR, Visscher PM, Yang J. 2018. Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat Commun.* 9(1):1–9. doi:10.1038/s41467-017-02088-w.

Haag J, Jordan AI, Stamatakis A. 2025. Pandora: a tool to estimate dimensionality reduction stability of genotype data. *Bioinf Adv.* 5(1). doi:10.1093/bioadv/vbaf040.

Harpak A, Przeworski M. 2021. The evolution of group differences in changing environments. *PLoS Biol.* 19(1):e3001072. doi:10.1371/journal.pbio.3001072.

Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, Carslake D, Hemani G, Paternoster L, Smith GD, et al. 2019. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun.* 10(1):333. doi:10.1038/s41467-018-08219-1.

Hoffman GE. 2013. Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PLoS One.* 8(10):e75707. doi:10.1371/journal.pone.0075707.

Hu S, Ferreira LA, Shi S, Hellenthal G, Marchini J, Lawson DJ, Myers SR. 2025. Fine-scale population structure and widespread conservation of genetic effect sizes between human groups across traits. *Nat Genet.* 57:379–389. doi:10.1038/s41588-024-02035-8.

- Johnstone IM, Paul D. 2018. PCA in high dimensions: An orientation. *Proc IEEE*. 106(8):1277–1292. doi:[10.1109/JPROC.2018.2846730](https://doi.org/10.1109/JPROC.2018.2846730).
- Josephs EB, Berg JJ, Ross-Ibarra J, Coop G. 2019. Detecting adaptive differentiation in structured populations with genomic data and common gardens. *Genetics*. 211(3):989–1004. doi:[10.1534/genetics.118.301786](https://doi.org/10.1534/genetics.118.301786).
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 42(4):348–354. doi:[10.1038/ng.548](https://doi.org/10.1038/ng.548).
- Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*. 12(5):e1004842. doi:[10.1371/journal.pcbi.1004842](https://doi.org/10.1371/journal.pcbi.1004842).
- Kerminen S, Martin AR, Koskela J, Ruotsalainen SE, Havulinna AS, Surakka I, Palotie A, Perola M, Salomaa V, Daly MJ, et al. 2019. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *Am J Hum Genet*. 104(6):1169–1181. doi:[10.1016/j.ajhg.2019.05.001](https://doi.org/10.1016/j.ajhg.2019.05.001).
- Kremer A, Le Corre V. 2012. Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity (Edinb)*. 108(4):375–385. doi:[10.1038/hdy.2011.81](https://doi.org/10.1038/hdy.2011.81).
- Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. *Sci (New York, N.Y.)*. 265(5181):2037–2048. doi:[10.1126/science.8091226](https://doi.org/10.1126/science.8091226).
- Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 467(7317):832–838. doi:[10.1038/nature09410](https://doi.org/10.1038/nature09410).
- Latta R. 1998. Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *Am Nat*. 151(3):283–292. doi:[10.1086/286119](https://doi.org/10.1086/286119).
- Latta RG. 2004. Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytol*. 161(1):51–58. doi:[10.1046/j.1469-8137.2003.00920.x](https://doi.org/10.1046/j.1469-8137.2003.00920.x).
- Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, Crawford A, Hemani G, Davey Smith G, Timpson NJ. 2020. Is population structure in the genetic Biobank era irrelevant, a challenge, or an opportunity? *Hum Genet*. 139(1):23–41. doi:[10.1007/s00439-019-02014-8](https://doi.org/10.1007/s00439-019-02014-8).
- Le MK, Smith OS, Akbari A, Harpak A, Reich D, Narasimhan VM. 2022. 1,000 ancient genomes uncover 10,000 years of natural selection in Europe. *Pages: 2022.08.24.505188 Section: New Results*.
- Le Corre V, Kremer A. 2003. Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*. 164(3):1205–1219. doi:[10.1093/genetics/164.3.1205](https://doi.org/10.1093/genetics/164.3.1205).
- Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. 2012. Improved linear mixed models for genome-wide association studies. *Nat Methods*. 9(6):525–526. doi:[10.1038/nmeth.2037](https://doi.org/10.1038/nmeth.2037).
- Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, et al. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*. 47(3):284–290. doi:[10.1038/ng.3190](https://doi.org/10.1038/ng.3190).
- Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet*. 100(4):635–649. doi:[10.1016/j.ajhg.2017.03.004](https://doi.org/10.1016/j.ajhg.2017.03.004).
- Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 51(4):584–591. doi:[10.1038/s41588-019-0379-x](https://doi.org/10.1038/s41588-019-0379-x).
- Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*. 528(7583):499–503. doi:[10.1038/nature16152](https://doi.org/10.1038/nature16152).
- Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 44(3):243–246. doi:[10.1038/ng.1074](https://doi.org/10.1038/ng.1074).
- McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet*. 5(10):e1000686. doi:[10.1371/journal.pgen.1000686](https://doi.org/10.1371/journal.pgen.1000686).
- Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. 2020. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*. 9:e48376. doi:[10.7554/eLife.48376](https://doi.org/10.7554/eLife.48376).
- Novembre J, Barton NH. 2018. Tread lightly interpreting polygenic tests of selection. *Genetics*. 208(4):1351–1355. doi:[10.1534/genetics.118.300786](https://doi.org/10.1534/genetics.118.300786).
- Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*. 40(5):646–649. doi:[10.1038/ng.139](https://doi.org/10.1038/ng.139).
- Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics*. 192(3):1065–1093. doi:[10.1534/genetics.112.145037](https://doi.org/10.1534/genetics.112.145037).
- Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2(12):e190. doi:[10.1371/journal.pgen.0020190](https://doi.org/10.1371/journal.pgen.0020190).
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38(8):904–909. doi:[10.1038/ng1847](https://doi.org/10.1038/ng1847).
- Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 20(4):R208–R215. doi:[10.1016/j.cub.2009.11.055](https://doi.org/10.1016/j.cub.2009.11.055).
- Pritchard JK, Rienzo AD. 2010. Adaptation—not by sweeps alone. *Nat Rev Genet*. 11(10):665–667. doi:[10.1038/nrg2880](https://doi.org/10.1038/nrg2880).
- Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P, Stone JL, Sullivan PF, Ruderfer DM, et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 460(7256):748–752. doi:[10.1038/nature08185](https://doi.org/10.1038/nature08185).
- Racimo F, Berg JJ, Pickrell JK. 2018. Detecting polygenic adaptation in admixture graphs. *Genetics*. 208(4):1565–1584. doi:[10.1534/genetics.117.300489](https://doi.org/10.1534/genetics.117.300489).
- Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature*. 461(7263):489–494. doi:[10.1038/nature08365](https://doi.org/10.1038/nature08365).
- Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, Powell JE, Vinkhuyzen A, Berndt SI, Gustafsson S, et al. 2015. Population genetic differentiation of height and body mass index across Europe. *Nat Genet*. 47(11):1357–1362. doi:[10.1038/ng.3401](https://doi.org/10.1038/ng.3401).
- Rosenberg NA, Edge MD, Pritchard JK, Feldman MW. 2019. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evol Med Public Health*. 2019(1):26–34. doi:[10.1093/emph/eoy036](https://doi.org/10.1093/emph/eoy036).
- Sarmanova A, Morris T, Lawson DJ. 2020. Population stratification in GWAS meta-analysis should be standardized to the best available reference datasets. *Pages: 2020.09.03.281568 Section: New Results*.

- Schraiber JG, Edge MD, Pennell M. 2024. Unifying approaches from statistical genetics and phylogenetics for mapping phenotypes in structured populations. *PLoS Biol.* 22(10):e3002847. doi:10.1371/journal.pbio.3002847.
- Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CW, Hirschhorn J, Daly MJ, Patterson N, et al. 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife.* 8:e39702. doi:10.7554/eLife.39702.
- Song W, Shi Y, Wang W, Pan W, Qian W, Yu S, Zhao M, Lin GN. 2021. A selection pressure landscape for 870 human polygenic traits. *Nat Hum Behav.* 5(12):1731–1743. doi:10.1038/s41562-021-01231-4.
- Stern AJ, Speidel L, Zaitlen NA, Nielsen R. 2021. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am J Hum Genet.* 108(2):219–239. doi:10.1016/j.ajhg.2020.12.005.
- Tan T, Jayashankar H, Guan J, Nehzati SM, Mir M, Bennett M, Agerbo E, Ahlskog R, Pinto de Andrade Anapaz V, Asvold BO, et al. 2024. Family-GWAS reveals effects of environment and mating on genetic associations. *bioRxiv* 24314703. <https://doi.org/10.1101/2024.10.01.24314703>, 2024–10, preprint: not peer reviewed.
- Trochet H, Pelletier J, Tadros R, Hussin J. 2021. Comparison of polygenic risk scores for heart disease highlights obstacles to overcome for clinical use. Technical report, *bioRxiv*. Section: New Results Type: article. *bioRxiv* 243287. <https://doi.org/10.1101/2020.08.09.243287>, preprint: not peer reviewed.
- Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet.* 44(9):1015–1019. doi:10.1038/ng.2368.
- Uricchio LH, Kitano HC, Gusev A, Zaitlen NA. 2019. An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evol Lett.* 3(1):69–79. doi:10.1002/evl3.97.
- Veller C, Coop G. 2023. Interpreting population and family-based genome-wide association studies in the presence of confounding. Pages: 2023.02.26.530052 Section: New Results.
- Vilhjálmsdóttir BJ, Nordborg M. 2013. The nature of confounding in genome-wide association studies. *Nat Rev Genet.* 14(1):1–2. doi:10.1038/nrg3382.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet.* 101(1):5–22. doi:10.1016/j.ajhg.2017.06.005.
- Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. 2020. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun.* 11(1):3865. doi:10.1038/s41467-020-17719-y.
- Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet.* 46(11):1173–1186. doi:10.1038/ng.3097.
- Yair S, Coop G. 2022. Population differentiation of polygenic score predictions under stabilizing selection. *Philos Trans R Soc B Biol Sci.* 377(1852):20200416. doi:10.1098/rstb.2020.0416.
- Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 88(1):76–82. doi:10.1016/j.ajhg.2010.11.011.
- Yang J, Zaitlen N, Goddard M, Visscher P, Price A. 2014. Mixed model association methods: advantages and pitfalls. *Nat Genet.* 46(2):100–106. doi:10.1038/ng.2876.
- Zaidi AA, Mathieson I. 2020. Demographic history mediates the effect of stratification on polygenic scores. *eLife.* 9. doi:10.7554/eLife.61548.
- Zhang Y, Pan W. 2015. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genet Epidemiol.* 39(3):149–155. doi:10.1002/gepi.21879.
- Zoledziewska M, Sidore C, Chiang CWK, Sanna S, Mulas A, Steri M, Busonero F, Marcus JH, Marongiu M, Maschio A, et al. 2015. Height-reducing variants and selection for short stature in Sardinia. *Nat Genet.* 47(11):1352–1356. doi:10.1038/ng.3403.

Editor: J. Yang