Testing for differences in polygenic scores in the presence of confounding

Jennifer Blanc[1*] and Jeremy J. Berg[1*]

[1]Department of Human Genetics, University of Chicago, Chicago, IL, USA
[*]To whom correspondence should be addressed: jgblanc@uchicago.edu, jjberg@uchicago.edu

Abstract

Polygenic scores have become an important tool in human genetics, enabling the prediction of individual 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 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 simple model of genetic relatedness to describe how confounding in the estimation panel biases the distribution of polygenic scores in a way 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. We then use the understanding gained from this analysis to develop a simple method that leverages the patterns of genetic similarity between the two panels to guard against these biases, and show that this method can provide better protection against confounding that the standard PCA-based approach.

1 Introduction

The calculation of polygenic scores [1] 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 panel. These effect estimates can be combined into polygenic scores in a distinct prediction panel by taking a sum of the genotypes of individuals in that panel, weighted by the estimated effects. Under the relatively strict assumption that variation in the phenotype is not correlated with variation in ancestry the GWAS panel and that the prediction panel individuals experience a similar distribution of environments to the GWAS panel individuals, these scores can be viewed as an estimate of each individual's expected phenotype, given their genotypes at the included sites. If this assumption is met, polygenic scores would seem to provide a means of partially separating genetic and environmental 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 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 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 panel. Ancestry stratification is a long recognized problem in the GWAS study design [2], 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, to statistical bias correction approaches, like the inclusion of genetic principal components as covariates [3], linear mixed models [4, 5], and LD score regression [6]. While these approaches have largely been successful in minimizing the number of false positive single variant associations [7], effect size estimates can still exhibit slight stratification biases without significantly altering the false discovery rates for individual associations, and these biases can be compounded when aggregating across loci, ultimately leading to confounded predictions in which the ancestry associated effects are mistaken for direct genetic effects.

Separation of direct genetic effects from correlations between ancestry and environment or 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 [8, 9, 10, 11]. It is natural to ask if 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 or phenotype biased migration. Medically, it is interesting to know if polygenic score differences or gradients represent real underlying gradients in the average genetic effect [12], whether those gradients are caused by neutral 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.

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 who differ in ancestry allows us to aggregate the small changes in allele frequency induced by selection on a polygenic trait into a detectable signal [13, 14, 15, 16]. Several research groups have developed and applied methods to detect these signals [17, 18, 19, 20, 21, 22, 23, 24]. 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 [25, 26, 17, 18, 27, 28, 29, 20, 30, 31, 19], only for these signals to weaken substantially or disappear entirely when re-evaluated using effects estimated in the larger and more genetically homogeneous UK Biobank [32, 33, 22]. Further analysis 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 [32, 33]. Though more recent work using approaches that should be robust to this confounding have provided support for some of the original signals (see below) [34, 35], ancestry stratification remains a concern.

Recently, in the context of selection tests, Chen et al [34] proposed a strategy to mitigate the impact of stratification by carefully choosing the GWAS panel so that even if residual stratification biases in effect size estimates exist, they will be unlikely to confound the test (see also [36, 37] for examples 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 [27], 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 a cost of reduced power: polygenic scores are generally less accurate when the effect sizes used to compute them are ported to genetically divergent samples [38, 39, 40, 41, 42]. Less accurate polygenic scores are then less able to detect a given polygenic selection event, all else equal. These decays in polygenic score accuracy also pose a significant challenge to their use in medicine, as scores that are predictive for some and not for others may exacerbate health inequities [43]. Thus, realizing the potential of polygenic scores in both basic science and medical applications will require the use of large and 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 of precise methods to protect against this bias.

In this paper, we develop a general model for bias in polygenic score predictions and propose a simple method to correct for bias in tests for an association between polygenic scores axes of ancestry variation. We model a joint sample of a GWAS and test panel in which we estimate effects in the former, then compute polygenic scores and test for association with an arbitrary axis of ancestry variation measured in the latter. We first characterize the bias in the distribution of the test panel polygenic scores as a function of the cross panel genetic similarity matrix, relying on simple and standard models for the distribution of phenotypes and genotypes across the two panels. We then show how the bias in the association between polygenic scores and a specific axis of ancestry variation in the test panel depends on a linear combination of the entries of the cross-panel genetic similarity matrix, where the test axis, measured in the test panel, specifies the combination. We then show that including this particular linear combination of genetic similarities as a covariate in the GWAS is sufficient to protect against bias in the association between polygenic scores and the chosen axis of ancestry variation in the test panel. We provide theoretical arguments linking this approach to the standard PCA approach of including top genetic principal components of the GWAS panel, and we show using simulations that our approach works when the PCA approach works, but also in cases where it fails. We then discuss several ways that this approach can provide clarity in polygenic score analyses.

2 Model

2.1 Genotypes and phenotypes

We assume that individuals in the GWAS panel are phenotyped and that the trait includes a contribution from a large number of 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

$$\begin{aligned} y &= \sum_{\ell} \beta_\ell G_\ell + e \\ &= u + e \end{aligned} \tag{1}$$

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

In our model, the environmental effect on each individual is an independent Normally distributed random variable with variance $1 - h^2$. However, the expected environmental effect for each individual can differ in some arbitrary but unknown way across individuals, reflecting underlying differences in their socio-environmental context. We write the distribution of environmental effects as $e \sim MVN\left(c, \left(1 - h^2\right)\mathbf{I}\right)$, where c is the vector of expected environmental effects.

Randomness in the genetic effect arises from segregation of genotypes within the panel, with a covariance structure given by relatedness matrix $\mathbf{F}_{GG}$. Again, however, the expected genetic effect can vary across individuals, reflecting the action of non-neutral processes such as natural selection, assortative mating, phenotype biased migration, or phenotype dependent ascertainment. We therefore write the distribution of genetic effects as $u \sim MVN\left(v, h^2\mathbf{F}_{GG}\right)$, where v is the vector of expected genetic effects across GWAS panel individuals given their ancestry and the set of non-neutral forces that have acted throughout the evolutionary history of the population. The sum $v + c$ therefore gives the vector of individuals' expected phenotypes, given their ancestry and socio-environmental contexts. We assume that these are not observed.

The genotypes at individual sites covary across panels due to shared ancestry. We assume that each variant was present at some frequency λ_ℓ in a set of shared ancestors common to both panels, and model the pattern of variation around the expected genotype ($2\lambda_\ell$) due to variation in relatedness. We write X_ℓ for the vector of mean centered genotypes measured in N test panel individuals, and G_ℓ for the vector of mean centered genotypes in the M GWAS panel individuals. The expected pattern of covariance at site ℓ, both within and across panels, is

$$Var\left(\begin{bmatrix} X_\ell \\ G_\ell \end{bmatrix}\right) = 4\lambda_\ell\left(1 - \lambda_\ell\right)\mathbf{F} \tag{2}$$

where

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{XX} & \mathbf{F}_{XG} \\ \mathbf{F}_{GX} & \mathbf{F}_{GG} \end{bmatrix} \tag{3}$$

is the joint relatedness matrix for the two panels. The matrices $\mathbf{F}_{GG}$ and $\mathbf{F}_{XX}$ describe within panel relatedness, while $\mathbf{F}_{GX} = \mathbf{F}_{XG}^\top$ contains the cross panel coefficients. We assume that individual

sites are unlinked, so that the only covariance among sites is that arising from shared ancestry among individuals, i.e. population structure.

2.2 Polygenic score association tests

We suppose that marginal associations, denoted $\hat{\beta}_\ell$, are computed in the GWAS panel via ordinary least squares, and that a total of S sites are ascertained for inclusion in the polygenic score, computed in the test panel as

$$Z = \sum_{\ell}^{S} \hat{\beta}_\ell X_\ell. \tag{4}$$

We then consider testing the polygenic scores for an association with an arbitrary test vector T, which might represent an eco-geographic variable of interest (e.g latitude [18] or an encoding of whether one lives in a particular geographic region or not [9, 44]), the fraction of an individual's genome assigned to a particular "ancestry group"[17, 20], or one of the top genetic principal components of the test panel genotype matrix [21]. We refer to this as a "polygenic score association test".

As a test statistic we consider the observed covariance between the test vector and the polygenic scores, i.e.

$$\hat{q} = \frac{1}{N-1} Z^\top T. \tag{5}$$

Under the null model, $\hat{q}$ has mean 0, reflecting the fact that genetic drift is directionless, and variance $V_A^{PGS}\left(\frac{1}{N} + f_{TX}\right)$, where $V_A^{PGS} = 4\sum_{\ell=1}^{S} \hat{\beta}_\ell^2 \lambda_\ell \left(1-\lambda_\ell\right)$, and $f_{TX} = T^\top \left(F_{XX} - \mathbf{I}\right) T$ captures variance arising from relatedness among test panel individuals. Under alternative models including the effect of natural selection or other non-neutral processes, $\hat{q}$ also has mean zero, reflecting our uncertainty about the direction of the effect, but an increased variance.

We can also reframe our test as a statement about the association between the effect estimates and a set of genotype contrasts, $r_\ell = X_\ell^\top T$, which measure the association between the test vector and the genotypes at each site. Then, writing r and $\hat{\beta}$ for the vectors of genotype contrasts and effect estimates, we can equivalently write our test statistic as

$$\hat{q} = \frac{1}{N-1} \hat{\beta}^\top r, \tag{6}$$

which will help make theoretical connections between different interpretations of our results and our method of correcting for stratification bias.

3 Results

Now, given these simple modeling assumptions, we describe how the relationship between the GWAS and test panels impact the distribution of polygenic scores and the association between the polygenic scores and the test vector. We first consider the case where no attempt is made to correct for population structure, before discussing the role of the standard PCA approach in the context of our method.

3.1 Estimating marginal effects

Conditional on the genetic and environmental effects, and genotypes at the focal site, the marginal effect size estimate is given by

$$\begin{aligned}\hat{\beta}_\ell \mid G_\ell, u, e &= \frac{y^\top G_\ell}{G_\ell^\top G_\ell} \\ &= \frac{u^\top G_\ell}{G_\ell^\top G_\ell} + \frac{e^\top G_\ell}{G_\ell^\top G_\ell} \\ &= \sigma_\ell + \epsilon_\ell.\end{aligned} \tag{7}$$

where σ_ℓ reflects the association between the total genetic effect on the trait and the genotype at the focal site (including the effect of the focal allele if it is causal) and ϵ_ℓ reflects the association with the environmental effect.

Additionally, we can decompose the genetic effect into the causal contribution from the focal site and the contribution from the background, i.e. $u = \beta_\ell G_\ell + u_{-\ell}$, so that σ_ℓ can be further decomposed as $\sigma_\ell = \beta_\ell + \sigma_{-\ell}$ where $\sigma_{-\ell}$ is the association between the focal site and the background contribution from all other sites. Because drift is directionless, an individual variant is as likely to be positively associated with either the genetic background or the environmental effect as it is to be negatively associated. As a result, the expected residual contribution to the marginal effect is zero for both the genetic and environmental contributions, i.e. $\mathbb{E}\left[\sigma_{-\ell} \mid u_{-\ell}\right] = 0$ and $\mathbb{E}\left[\epsilon_\ell \mid e\right] = 0$, so that the expected marginal effect is equal to the allele's causal effect: $\mathbb{E}\left[\hat{\beta}_\ell \mid u_{-\ell}, e\right] = \beta_\ell$.

3.2 Stratification bias induces ancestry associated bias in polygenic scores

The deviation of an allele's estimated causal effect size from its expectation depends on how the genotypes in the GWAS panel (G_ℓ) deviate from their neutral expectation. Because these deviations are correlated with the deviations of the test panel genotypes (X_ℓ) due to shared ancestry, the estimated effect sizes can become correlated with the pattern of genotypic variation in the test panel for reasons that have nothing to do with the actual genetic effect of the variant. This leads to a bias in the expectation of the polygenic scores, and in turn a bias in polygenic score association tests. The expectation of the vector of polygenic scores in the test panel (Z) is

$$\mathbb{E}\left[Z\right]^\top = \frac{S}{(M-1)}\left(v^\top + c^\top\right)\tilde{\mathbf{F}}_{GX} \tag{8}$$

(supplement section 6.1) where

$$\tilde{\mathbf{F}}_{GX} = \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell/(M-1)}\right], \tag{9}$$

is the standardized cross panel genetic similarity matrix, where both the GWAS and test panel genotypes have been standardized by the variance in the GWAS panel, rather than by the variance in the combined sample.

3.3 Bias in polygenic scores leads to biased polygenic score associations

The bias in the polygenic score association ($\hat{q}$) follows straightforwardly from the bias in the polygenic scores. The expectation of $\hat{q}$ is

$$\begin{aligned}\mathbb{E}\left[\hat{q}\right] &= \frac{1}{N-1}\mathbb{E}\left[Z\right]^{\top} T \\ &= \frac{S}{(N-1)(M-1)}\left(v^{\top}+c^{\top}\right)\tilde{\mathbf{F}}_{GX}T. \end{aligned} \tag{10}$$

We therefore expect the polygenic score association test to be biased when the test vector (T) aligns with the vector of expected phenotypes ($v+c$) in a space defined by the cross panel genetic similarity matrix ($\tilde{\mathbf{F}}_{XG}$).

This result therefore provides explicit theoretical justification for the procedure of choosing GWAS and test panels that do not overlap in population structure as a means to avoiding bias in polygenic score analyses [1, 34, 45], as $\mathbb{E}\left[\hat{q}\right]=0$ if $\tilde{\mathbf{F}}_{XG}=\mathbf{0}$. Any polygenic score association test performed using a pair of GWAS and test panels that meet this non-overlapping population structure criterion will therefore be unbiased, no matter what specific test is performed, or how the expected phenotypes are distributed. Stratification in the GWAS may still bias individual effects, but these residual biases are indistinguishable from noise from the perspective of the polygenic score association test. This can lead to a reduction in the accuracy of the polygenic scores, but does not incur bias.

Alternatively, recalling the site-by-site interpretation of the polygenic score association (Eq (6)), we can also express the bias as

$$\begin{aligned}\mathbb{E}\left[\hat{q}\right] &= \frac{S}{(N-1)(M-1)}\mathbb{E}\left[\hat{\beta}^{\top}r\right] \\ &= \frac{S}{(N-1)(M-1)}\left(v^{\top}+c^{\top}\right)\tilde{F}_{Gr} \end{aligned} \tag{11}$$

where

$$\begin{aligned}\tilde{F}_{Gr} &= \mathbb{E}\left[\frac{G_{\ell}r_{\ell}^{\top}}{G_{\ell}^{\top}G_{\ell}/(M-1)}\right] \\ &= \tilde{\mathbf{F}}_{GX}T, \end{aligned} \tag{12}$$

is a vector whose length is equal to the number of individuals in the GWAS panel (M). For each individual m in the GWAS panel, $\tilde{F}_{Gr,m}$ contains a linear combination of the their genetic similarities to the test panel individuals, where the combination is specified by the test vector. Alternatively, but equivalently, we can interpret $\tilde{F}_{Gr,m}$ as the expected correlation across sites between the standardized genotypes of GWAS panel individual m and the genotype contrasts of the association test (r), standardized by the square root of the variance in the GWAS panel. Expressing the bias in this way leads to two key insights. First, the test statistic is biased when the vector of expected phenotypes ($v+c$) aligns with the vector of correlations between GWAS panel genotypes and test panel genotype contrasts ($\tilde{F}_{Gr}$). Second, for a specific hypothesis corresponding to a specific set of genotype contrasts (r), we need only to ensure that the linear combinations of genetic similarities all sum to zero (i.e. that $\tilde{F}_{Gr}$ is a vector of zeros), which is a narrower condition

than the complete non-overlapping population structure criterion (i.e. $\tilde{\mathbf{F}}_{XG} = \mathbf{0}$). While it would be difficult to meet this criterion via sample construction without also meeting the broader non-overlapping population structure criterion, it can be achieved statistically, which is the topic we turn to next.

3.4 Controlling for stratification bias in polygenic association tests

3.4.1 A simple fixed-effects procedure to remove stratification bias

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 [46, 47, 42]. Another obvious solution is to include the vectors of expected genetic and environmental effects, v 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, v and c are typically not measureable, so this is generally not an option.

Our analysis suggests a different approach. Namely, if $\tilde{F}_{Gr}$ is known, then including it as a co-variate in the GWAS will remove bias in effect size estimates that is correlated with the genotype contrasts relevant to the test (r). In this case, the OLS estimated marginal effect at site ℓ, conditional on G_ℓ, u and e, is

$$\begin{aligned}\hat{\beta}'_\ell \mid G_\ell, u, e &= \frac{u^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} + \frac{e^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} \\ &= \sigma'_\ell + \epsilon'_\ell.\end{aligned} \tag{13}$$

where $\mathbf{P} = \left(\mathbf{I} - \frac{1}{\|\tilde{F}_{Gr}\|}\tilde{F}_{Gr}\tilde{F}_{Gr}^\top\right)$ removes variation along the axis specified by $\tilde{F}_{Gr}$ from the estimated effects, and renders them uncorrelated with the 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

$$\mathbb{E}\left[Z\right]^\top \approx \frac{S}{(M-1)}\left(v^\top + c^\top\right)\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \tag{14}$$

where

$$\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \approx \mathbf{P}\tilde{\mathbf{F}}_{GX} \tag{15}$$

captures cross panel relatedness along all axes of variation other than that specified by $\tilde{F}_{Gr}$. However, regressing out variation aligned with $\tilde{F}_{Gr}$ ensures that $\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} T = 0$, and it follows that

$$\begin{aligned}\mathbb{E}\left[\hat{q}\right] &\approx \frac{S}{(N-1)(M-1)}\left(v^\top + c^\top\right)\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} T \\ &\approx 0.\end{aligned} \tag{16}$$

3.4.2 Implementing the bias correction

In practice, $\tilde{F}_{Gr}$ is not known, but can be estimated from genome-wide data as a linear combination of the GWAS panel genotypes (G_ℓ), weighted by the genotype contrasts of the test (r_ℓ), each standardized by the variance in the GWAS panel

$$\hat{F}_{Gr} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{G_\ell r_\ell}{G_\ell^\top G_\ell / (M-1)}. \tag{17}$$

We then estimate marginal effects in the GWAS under the model

$$y = G_\ell \beta_\ell + \gamma \hat{F}_{Gr} + \varepsilon, \tag{18}$$

and then ascertain SNPs for inclusion in the polygenic scores via standard methods. In our simulations (see below and Methods), we also re-estimate their effects in a joint model as

$$y = \mathbf{G}_S \beta_S + \gamma \hat{F}_{Gr} + \varepsilon \tag{19}$$

where the S subscript indicates that only the ascertained sites are included. We then use this set of ascertained sites with jointly re-estimated effect sizes to perform the association tests in our simulated examples below.

3.4.3 Predicting GWAS panel genotypes

In the previous sections, we characterized our approach as an effort to estimate a specific linear combination of the entries of the cross panel genetic similarity matrix between GWAS and test panels. In this section, we give a different interpretation of $\hat{F}_{Gr}$ in terms of a linear prediction of the standardized GWAS panel genotypes (i.e. $\tilde{G}_\ell = \frac{G_\ell}{\sqrt{G_\ell^\top G_\ell/(M-1)}}$) given the set of genotype contrasts, also standardized by the GWAS panel genotype variance (i.e. $\tilde{r}_\ell = \frac{r_\ell}{\sqrt{G_\ell^\top G_\ell/(M-1)}}$). We consider the matrix-variate regression model

$$\tilde{\mathbf{G}} = \rho \tilde{r}^\top + \varepsilon. \tag{20}$$

In this model, $\tilde{\mathbf{G}}$ is an $M \times L$ matrix containing the standardized genotypes ($\tilde{G}_\ell$) as columns, θ is a vector of length M containing one regression coefficient per individual in the GWAS panel, $\tilde{r}$ is the length L vector of standardized genotype contrasts, and ε is the $M \times L$ matrix of residuals. The OLS estimate of θ is

$$\hat{\rho} = \frac{1}{\tilde{r}^\top \tilde{r}} \sum_{\ell} \frac{G_\ell r_\ell}{G_\ell^\top G_\ell / (M-1)}. \tag{21}$$

Then, for each site ℓ, a linear prediction of $\tilde{G}_\ell$ given $\tilde{r}_\ell$ is equal to the product of $\hat{\theta}$ and $\tilde{r}_\ell$:

$$\mathbb{E}\left[\tilde{G}_\ell \mid \tilde{r}_\ell\right] = \hat{\rho}\tilde{r}_\ell. \tag{22}$$

The expression for our estimate of $\hat{F}_{Gr}$ in equation (17) differs from the expression for $\hat{\rho}$ by only a scalar (L^{-1} in equation (17) vs $\left(\tilde{r}^\top \tilde{r}\right)^{-1}$ here), so we can intuitively interpret our bias correction method as including $\mathbb{E}\left[\tilde{G}_\ell \mid \tilde{r}_\ell\right]$ as a covariate when estimating the effect at site ℓ.

3.4.4 Relationship to PCA

A standard approach to controlling for population stratification in polygenic scores is to include the top K principle components of the GWAS panel genotype matrix as covariates in the GWAS, for some suitably large value of K. How does our approach relate to the this standard approach? In general, $\hat{F}_{Gr}$ can be written as a linear combination of the PCs of the GWAS panel genotype matrix, where the weights of this linear combination are a function of the covariance between genotype contrasts r and the SNP loadings for each PC.

Specifically, we can write $\hat{F}_{Gr}$ as

$$\hat{F}_{Gr} = \sum_i \alpha_i U_i \tag{23}$$

where U_i is the ith PC of the standardized GWAS panel genotype matrix, and the weights are given by

$$\alpha_i = \lambda_i Cov(V_i, \tilde{r}) \tag{24}$$

where λ_i is the ith singular value of the standardized GWAS panel genotype matrix, V_i is the vector of SNP loadings for the ith PC, and $\tilde{r}$ contains the standardized genotype contrasts (i.e. $\tilde{r}_\ell = \frac{r_\ell}{\sqrt{G_\ell^\top G_\ell / \sqrt{(M-1)}}}$). Estimating the marginal associations with $\hat{F}_{Gr}$ as a covariate can therefore be understood as fitting a model in which *all* PCs of the GWAS panel genotype matrix are included as covariates, but the relative magnitude of the contributions from different PCs are fixed, and we estimate only a single slope that scales the contributions from all of the PCs jointly, i.e.

$$y = G_\ell \beta_\ell + \gamma \left(\sum_i \alpha_i U_i^{\tilde{G}} \right) + e \tag{25}$$

where the α_i are fixed according to equation (24).

As a corollary, if we perform a polygenic score association test using GWAS effect size estimates in which the top K PCs of the GWAS panel genotypes are included as covariates, the necessary and sufficient condition for the included PCs to protect against bias from unmeasured confounders in a particular polygenic score association test is that $\hat{F}_{Gr}$ falls within the span of those top K PCs, i.e. if $\alpha_i \approx 0$ for $i > K$. An alternative interpretation of the PC correction approach is that it operates on a hypothesis that the major axes of confounding in a given GWAS panel (i.e. $v + c$ in our notation) can be captured by the included PCs. A given association test will also be unbiased if this condition is met, but it is more difficult to assess whether this is the case.

3.4.5 Downward bias with true signal

Our procedure differs from standard approaches for computing polygenic scores in that it uses the test panel genotype data twice: once when controlling for stratification in the GWAS panel, and a second time when testing for polygenic selection. Should we not then worry that our procedure will regress out real signal when it exists? The answer is no, so long as the number of SNPs used to compute the correction is large relative to the number included in the polygenic score.

To see why, consider that we can rewrite our covariate in equation (18) in terms of a sum of the contribution from our focal site and all other sites as

$$y = G_\ell \beta_\ell + \gamma \left(\frac{L-1}{L} \hat{F}_{Gr,-\ell} + \frac{1}{L} \frac{G_\ell r_\ell}{G_\ell^\top G_\ell} \right) + e \tag{26}$$

where $\hat{F}_{Gr,-\ell}$ is the estimate of $\tilde{F}_{Gr}$ that one would obtain using all sites other than the focal one. Thus, because our covariate includes a contribution from the focal site itself, controlling for it induces a slight bias in the estimated effect size, the sign of which depends on the sign of r_ℓ. If r_ℓ is positive, $\hat{\beta}_\ell$ has a slight negative bias, whereas if r_ℓ is negative, the bias will be positive (a similar effect has been noticed in the context of correcting for PCs of gene expression data [48], and is the motivation for the "leave-one-chromosome-out" approach to fitting linear mixed models [49]). Because the variance in r_ℓ is similar across sites, the product $r_\ell \hat{\beta}_\ell$ will be biased toward 0 by a factor of approximately $\left(1 - \frac{1}{L}\right)$ for each site, owing to the fact that the focal site contributes approximately $\frac{1}{L}$ of our estimate $\hat{F}_{Gr}$. Our test statistic is a sum over contributions of $r_\ell \hat{\beta}_\ell$ from S independent sites, so including $\hat{F}_{Gr}$ as a covariate when estimating effect sizes induces a downward bias of approximately $\left(1 - \frac{S}{L}\right)$, i.e.

$$\mathbb{E}\left[\hat{q} \mid q\right] \approx q \left(1 - \frac{S}{L}\right). \tag{27}$$

where q is the value of our test statistic we would expect to obtain if the causal effects were known. Our procedure is therefore expected to remove a small amount of signal alongside the stratification bias, but this effect will be minimal as long as the number of sites used to compute $\hat{F}_{Gr}$ far exceeds the number of sites included in the polygenic score. The key principle is that when there is overlapping structure between the GWAS and test panels, our ability to separate real signal from stratification bias depends on an assumption that the signal is sparse, but the confounding is dense. Notably, because the standard PCA approach protects against stratification bias in the test by capturing the same axis of cross panel variation (see above), it will induce the same downward bias.

Real populations exhibit linkage disequilibrium, so the appropriate value of L corresponds to an "effective" number of independent sites, rather than the actual total number of sites used to compute the correction. For human population samples imputed to common reference panels, the effective number of SNPs is typically on the order of at least half a million [50], suggesting that even for a polygenic score that included, for example, 10,000 SNPs, the downward bias should be no more than 2%. Further concern about downward biases in applications could be amerliorated via a "leave one chromosome out" scheme [49], in which an estimate of $\hat{F}_{Gr}$ for each of the 22 autosomes would be computed using only genotypes from the other 21.

3.5 Applications: Theory and Simulations

In this section, using both theory and simulations, we consider a number of concrete examples with varying degree of alignment between the axis of stratification and axis of population structure relevant to the polygenic association test, demonstrating how these biases play out in practice, and how our proposed correction reduces false positives to nominal levels.

3.5.1 Toy Model

We first consider a toy model with four populations (labelled A, B, C and D), which are related to one another by an evenly balanced population phylogeny (Figure 1) with no structure within each of the populations. The GWAS panel is composed of an equal mixture of individuals from populations A and B, and we test for a difference in mean polygenic score between populations C and D under two different topologies, one where A and C are sister to one another (Figure 1A), and another where A and B are sister (Figure 1C). For simplicity, we simulate a purely environmental phenotype (i.e. $h^2 = 0$) with a difference in mean between populations A and B equal to Δ_{AB}, so that $e_i = \frac{\Delta_{AB}}{2} + \varepsilon_i$ if individual i belongs to population A, and $e_i = -\frac{\Delta_{AB}}{2} + \varepsilon_i$ if they belong to population B, where $\varepsilon_i \sim N(0,1)$ (Figure 1B). Following from equation (7), the marginal effect size estimate for site ℓ is then

$$\begin{aligned} \hat{\beta}_\ell \mid G_\ell, e &= \frac{G_\ell^\top e}{G_\ell^\top G_\ell} \\ &= \frac{\Delta_{AB}\left(p_{A,\ell} - p_{B,\ell}\right)}{\frac{2}{M} G_\ell^\top G_\ell} + \frac{G_\ell^\top \varepsilon}{G_\ell^\top G_\ell} \end{aligned} \tag{28}$$

where $p_{A,\ell}$ and $p_{B,\ell}$ are the population allele frequencies at site ℓ (see also equation 2.3 in the supplement of [51]).

Then, using these these effect sizes to test for a difference between populations C and D. The expected value of $\hat{q}$ is

$$\begin{aligned} \mathbb{E}\left[\hat{q}\right] &= \Delta_{AB} \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{(p_{A,\ell} - p_{B,\ell})(p_{C,\ell} - p_{D,\ell})}{\frac{1}{M} G_\ell^\top G_\ell}\right] \\ &= \Delta_{AB} S \tilde{F}_4(A, B; C, D) \end{aligned} \tag{29}$$

where $\tilde{F}_4(A, B; C, D)$ is a version of Patterson's F_4 statistic, [52, 53] standardized by the genotypic variance in the GWAS panel, and measures the amount of genetic drift common to population A and B that is also shared by populations C and D. Writing the expectation of $\hat{q}$ in terms of this modified F_4 statistic helps illustrate the role of cross panel population structure in driving stratification bias in polygenic scores for a simple example. The effect estimate at site ℓ is a linear function of $p_{A,\ell} - p_{B,\ell}$, so the resulting selection test will be biased if $p_{A,\ell} - p_{B,\ell}$ is correlated with $p_{C,\ell} - p_{D,\ell}$. This is true for the demographic model in Figure 1A, where shared drift on the internal branch generates such a correlation, yielding a positive value for $\tilde{F}_4(A, B; C, D)$ and therefore a biased test statistic, but not for the model in Figure 1B, where there is no shared internal branch, $\tilde{F}_4(A, B; C, D) = 0$, and the test statistic is unbiased.

To test this prediction, we simulated 100 replicates of four populations related by this topology. In each simulation, populations on the same side of the tree are separated by 100 generations of evolution, while populations on opposite sides are separated by 200 generations. The diploid population size is 10,000 for each population and in the ancestral populations throughout the simulation, so the F_{ST} between pairs of populations on the same side of the tree is approximately 0.005, and 0.01 for populations on opposite sites. In contrast to our theoretical model, in our simulations we include linkage over short scales by simulating 200 independent 1Mb linkage blocks per simulation replicate. We then simulated purely environmental phenotypes in the GWAS panel populations (A and B), with a difference in mean phenotype, as outlined above, conducted a GWAS

to ascertain SNPs, and then used these ascertained SNPs to construct polygenic scores and test for a difference in mean between populations C and D. To assess whether the difference was significantly different from zero, we compare to an empirical null distribution constructed by permuting the signs of the of the effect estimates across sites (see Methods). The results are consistent with our theoretical expectations: the test is biased for the topology with $\tilde{F}_4(A, B; C, D) > 0$ (Figure 1D), but unbiased when $\tilde{F}_4(A, B; C, D) = 0$ (Figure 1E).

We then use all genome-wide SNPs with a frequency of greater than 5% in both the test and GWAS panel to estimate $\hat{F}_{Gr}$, as in equation (17). The resulting vector is, as expected, tightly correlated with population label for topology with $\tilde{F}_4(A, B; C, D) > 0$, but reflects random noise for the case when $\tilde{F}_4(A, B; C, D) = 0$ (Figure S2). When we re-run the polygenic score association test using effects estimated with $\hat{F}_{Gr}$ included as a covariate, the bias in $\hat{q}$ is removed (Figure S1), and false positive rates are restored to nominal levels for the confounded topology (Figure 1F). For the unconfounded topology, the test is already unbiased, so the correction has no effect (Figure 1G).

Next, we sought to confirm that including $\hat{F}_{Gr}$ does not regress out true signals of polygenic score divergence, consistent with our theoretical argument above. To do this, we modified our simulations by adding causal loci to make the trait heritable, with $h^2 = 0.3$, and sampled the sign of the effect for these causal loci so as to generate a correlation between the effect and the frequency differences between populations C and D. We compare the true positive rate across simulations under three different schemes for obtaining effect sizes: 1) estimating effects with no bias correction, 2) including $\hat{F}_{Gr}$ as a covariate, or 3) including known population ID in the GWAS. Correcting for population ID is included as a gold standard for this simple model, but would be unrealistic for a real sample as real populations do not conform to idealized population models. For each scheme, we consider both the case where the identity of the causal variants are known but the effects are estimated vs ascertaining the most significant variant per LD block using p-values. We compare all of these schemes to association tests performed using the true causal variants with their true causal effect sizes.

When the identities of the causal sites are known and there is no environmental stratification, all three schemes for estimating effetcts perform similarly, and equally as well as when the causal effects themselves are known exactly (Figure 2, top left). As expected, when the environmental stratification is in the same direction as the true mean difference, the uncorrected effect sizes reject the null at a higher rate than when the causal effects are known, while our method and population ID perform as well as having the causal effects (Figure 2, top middle). The same is true when the environmental effect is in the opposite direction as the true signal, except the uncorrected effects have reduced power, as expected (Figure 2, top right). When the identities of the causal variants are unknown and associated sites must be ascertained, we observe largely similar patterns, with our method and the population ID method having slightly reduced power relative to when the identities of the causal variants are known (Figure 2, bottom row). Our simulation results support our theoretical argument that including $\hat{F}_{Gr}$ does not remove real signals when $S << L$.

3.5.2 Grid Simulations

To further explore our correction procedure in more complex scenarios, we conducted another set of coalescent simulations under a symmetric two-way migration model on a six-by-six lattice grid. We sampled an equal number of individuals per deme to comprise both the GWAS and test panels and

then simulate several different distribution of purely environmental phenotypes across the GWAS panel individuals. We consider three different scenarios for the distribution of phenotypes: one where the mean varies with latitude, one where it increases only along the diagonal of the grid, and one in which only a single deme has an elevated mean, a scenario which is difficult to correct for with standard tools [54, 55]. For each scenario, we estimate effect sizes, ascertain associated sites, and test for an association between polygenic score and latitude, longitude, or membership in the single confounded deme, depending on the example.

For the first example, we simulated a latitudinal confounder such that c_i is a linear function of an individual's position on the latitudinal axis (Figure 3A). When we do an uncorrected GWAS and use those effect sizes to build polygenic scores in the test panel, their spatial distribution reflects the distribution of the environmental confounder. Consequently, an association test using latitude as the test vector is biased. However, when we compute $\hat{F}_{Gr}$ and include it as a covariate in the GWAS model, effect sizes are unbiased with respect to the latitudinal genotype contrasts in the test panel and the resulting association test is unbiased.

In the second example, we simulate confounding along the diagonal, resulting in uncorrected polygenic scores that are systemically correlated with both latitude and longitude in the test panel and the association test is biased along both axes (Figure 3B). However, when we compute $\hat{F}_{Gr}$ using latitude as the test vector, the resulting effect sizes are only uncorrelated with latitudinal genotype contrasts and polygenic scores remain susceptible to bias along any other axis (e.g. longitude). This example highlights the targetted nature of this method, as using effect sizes from a GWAS including $\hat{F}_{Gr}$ does not remove all bias, but does makes the association test using those effect sizes for the pre-specified test vector unbiased.

In the first two examples, the environmental confounder reflected broad geographic structure and could therefore be easily captured by the first two principal components of the GWAS panel. Thus, while our approach was effective, the standard PCA method was as well. In our next example we simulated single deme confounder which induces a more complex spatial pattern in the uncorrected polygenic scores (Figure 3C). Taking latitude to be the test vector, $\hat{F}_{Gr}$ again protects the selection test from bias only along the latitudinal axis. In this case, while the confounder is not fully captured by the first 10 principle components (see Figure S5), including 10 PCs in the GWAS successfully controls for stratification because $\hat{F}_{Gr}$ falls within their span.

In our final example, we again simulated a single deme confounder , but this time took the test vector to be an indicator for whether the test panel individuals were sampled from that deme or not. In this case, the test vector cannot be captured by the top 10 PCs (or, indeed, a much larger numbers of PCs , see Figure S3), so this method fails, while our method succeeds at protecting the test from bias. Conceptually, $\hat{F}_{Gr}$ is more successful than top principal components in this case because it is able to capture more subtle structure in the GWAS panel. We also tested our approach for each of these for scenarios over a range of magnitudes for the confounding effect. In our simulations, $\hat{F}_{Gr}$ protects all association tests from confounding effects of up to two phenotypic standard deviations, whereas the top 10 principle components begin to fail in some scenarios for more modest strengths of the confounding effect (Figure S4).

4 Discussion

Understanding associations between polygenic scores and patterns of ancestry variation is important to many areas of human genetics. Because most well-powered GWAS are conducted on population samples in which the relationship between environment and ancestry is not well controlled, stratification bias remains a significant concern [32, 33, 42, 56]. Here, we characterize patterns of stratification bias in the distribution of polygenic scores as a function of the genetic similarity between GWAS and test panels, and show how test panel genotypes can be used to protect against stratification bias when testing hypotheses about the relationship between polygenic scores and ancestry. The advantage of this approach is that it can succeed in situations where PCA fails because the bias lies along an axis of variation in the GWAS panel which is not captured by the top PCs.

Another advantage of our approach has to do with the tradeoff between statistical power and robustness in polygenic selection analyses, as researchers motivated by the difficulties with stratification bias and height [32, 33] have tended to take conservative approaches meant to trade away statistical power in exchange for robustness to biases due to ancestry stratification. This includes the non-overlapping population structure study design [34], but also choices such as focusing only on genome wide significant hits, or throwing away all effect size information except for the sign. An advantage of the stronger guarantee against bias due to ancestry stratification provided by our method is thus that it may enable the use of less conservative methods for computing polygenic scores, which do not trade away so much statistical power.

The disadvantage of the approach is that it requires that the test panel genotype are available and the test axis specified before carrying out the GWAS. Thus, the benefits of our approach cannot immediately be extended to situations where the individual level data for both panels is not available. Tests for association between polygenic scores and axes of ancestry variation are closely related to bivariate LD score regression as applied to a combination of effect estimates for one trait and frequency/genotype contrasts from an independent datasets [57, 19, 32], an approach which does not require the individual level genotypes, but does require access to ancestry-matched LD scores. Previous work in the context of polygenic selection tests raised concerns about inflation of the slope in this approach due to background selection [32]. A more complete analysis of bivariate LD score regression applied in this context would be welcome.

We have focused on analyses of the relationship between polygenic scores and patterns of ancestry variation, but our results are relevant for polygenic prediction more generally. For example, our approach could be used to protect polygenic score predictions from bias associated with ancestry associated socio-cultural variables measured in a particular prediction panel, even if those variables are not measured the GWAS panel. A related idea is to control for axes of ancestry variation identified in a large external reference dataset that captures as broad of a swath of human genetic variation as possible [58].

Our results are also relevant for analyses that use GWAS summary statistics and coalescent methods to test for signals of directional polygenic selection on variants associated with a particular trait [19, 23, 59, 60]. One way of thinking about these coalescent approaches is to recognized that they 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, if the relevant genotype

contrasts are estimated using the coalescent methods before conducting the GWAS, then these estimated genotype contrasts could be used to immunized the coalescent analysis against stratification bias, similar to how we immunize specific polygenic score comparisons in this paper.

A related point is that for both the coalescent approaches as well as methods relying on direct comparison 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 authors have suggested that certain coalescent methods of testing for polygenic selection are more robust to stratification bias than others [59, 60], but our results show that this cannot be true: two different methods that test the same evolutionary hypothesis using the same set of estimated effect sizes necessarily have the same susceptibility to stratification bias. Therefore, any observed differences in robustness to stratification bias among methods must come either from changing the evolutionary hypothesis being tested or from differences in ability to accurately estimate the relevant genotype contrasts (i.e. overall differences in statistical power).

We also note several limitations of our approach. First, our model assumes that all loci share the same underlying genetic relatedness matrix. For simple demographic histories like that of our toy model, this is unlikely to be an issue, but could be an issue for more complex demographic histories experienced by real human populations, as mutations of different ages experience the demographic history differently. One solution to this problem would be to bin variants by allele frequency or estimated age [61] and compute different corrections for different frequency/age bins similar to the suggestion by [55] to use principal components computed on rare variants to correct for stratification along axes of recent genetic divergence.

Another limitation is that in real human datasets with complex patterns of genetic variation and confounding, our approach alone will not be sufficient to mitigate all of the negative consequences of confounding. For example, while a GWAS conducted using only our covariate to control for stratification biases would result in effect size estimates that are uncorrelated with the target axis of population structure under the null, stratification along other axes would still function as additional noise in the effect size estimates. This would result in the ascertainment of tag SNPs that are less tightly coupled to their causal variants, and less accurate polygenic scores, reducing power to detect real signals. Therefore, control of population structure via inclusion of top PCs and/or linear mixed models, in addition to the inclusion of our projected test vector, is still desirable.

Finally, we note that the interpretation of results derived from applying our method will be limited by standard assumptions made in many polygenic score association tests and caution interpretation in light of these assumptions [62]. Notably, we rely on common tag SNPs, focus on the additive genetic component of a trait, and ignore the potential impact of interactions, either among genetic variants, or between genetic variants and the environment. Further, GWAS can also be impacted by other forms of genetic confounding beyond the simple associations between ancestry and genetic background that we consider here, including that arising from indirect effects, assortative mating, and stabilizing selection [63]. Finally, the interpretation of associations between polygeic scores and axes of ancestry variation remains challenging in general, as differences in polygenic scores do not necessarily translate into differences in phenotype [12], and associations between polygenic scores and patterns of ancestry variation may have counterintuitive evolutionary explanations [64, 41, 40]. Therefore, while our results provide a means to protect against stratification bias in polygenic score association tests, addressing a known problem in their implementation, continued care in the interpretation of these analyses is warranted.

5 Materials and Methods

5.1 Simulating genotypes

We used *msprime* [65] to simulate genotypes under 2 different models with 100 replicates per model. The first model, shown in Figure 1, has two population splits, 200 and 100 generations in 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. Each replicate consists of 200 independent 1Mb chromosomes with per base mutation and recombination rates equal to 1×10^{-8}. We then sample 20,000 individuals per population and create two configurations of GWAS and test panels based on the diagrams in Figure 1A and Figure 1C. We down sample the test panels to 2,000 individuals with an equal number from each of the two populations. Finally, we filter out variants with MAF < 0.05 in either panel from both panels. When the populations in the GWAS and test panel are non-sister (i.e Figure 1A) the average F_{ST} [66] was 0.01, whereas in the configuration in Figure 1C the average F_{ST} was 0.05.

The second model, modified from [67], is a 6×6 stepping stone model where structure extends infinitely far back with a symmetric migration rate of $m = 0.01$. As above, we fix the effective population size to 10,000 diploid individuals and simulate 200 chromosomes per individual. To reduce computational burden, we scale down to 10kb chromosomes with a per base mutation and recombination rates equal to 1×10^{-7}. We then sample 20 individuals per deme to create a test panel ($N = 720$) and 60 individuals per deme to create the GWAS panel ($M = 2160$). Again we filter out SNPs with MAF < 0.05 in either panel.

5.1.1 Simulating phenotypes

To study the effect of environmental stratification on association tests, we first simulated non-genetic phenotypes for an individual i in the GWAS panel as $y_i \sim N(0, 1)$. In our discrete 4 population model (Figure 1) we then generate a phenotypic difference between populations by adding Δ_{AB} to y_i for individuals in population B. In our grid simulations we generate a phenotypic gradient by adding $\frac{j\Delta}{J}$, to individuals in J specific grid squares depending on the desired stratification pattern (see Figure 3 and Figure S4) where Δ varies from 0 to 2 standard deviations.

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 (Figure 2), we used our 4 population demographic model and followed the protocol outlined in [67] to simulate a neutral trait with $h^2 = 0.3$. We first randomly select 1 causal variant per chromosome and sample their effect sizes $\beta_\ell \sim N(0, \sigma_i^2[p_\ell(1 - p_\ell)]^\alpha)$. σ_i^2 is the frequency independent component of genetic variance, 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 $\sigma_g^2 = \sigma_i^2 \sum_{\ell=1}^{200} [2p_\ell(1 - p_\ell)]^{\alpha+1} = 0.3$. 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 200 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 then draw the

environmental component of the phenotype $e_{i,k} \sim N(\mu_k, 1-h^2)$ where again $\mu_k = 0$ for individuals in population A and $\mu_k = \Delta_{AB}$ for individuals in population B.

5.1.2 Computing $\hat{F}_{Gr}$

In order to increase the precision in our estimates of the genotype contrasts, and therefore of $\hat{F}_{Gr}$, we first standardize the test panel genotypes and test vector by an estimate of the test panel genetic similarity matrix. We therefore compute $r^* = \mathbf{X}^\top \hat{\mathbf{F}}_{XX}^{-1} T$ where we first calculate $\hat{\mathbf{F}}_{XX}$ using the plink2 [68] function `--make-rel cov square` and then invert $\hat{\mathbf{F}}_{XX}$ using `eigen()` in R. Next, we multiply by T and use the plink2 function `--glm` to compute r^*. Finally we compute $\hat{F}_{Gr}$ as stated in (17) using `--sscore`.

5.1.3 GWAS and SNP ascertainment

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

1. $y = \beta_\ell G_\ell + \epsilon$
2. $y = \beta_\ell G_\ell + \gamma \hat{F}_{Gr} + \epsilon$
3. $y = \beta_\ell G_\ell + \gamma \mathrm{ID} + \epsilon$
4. $y = \beta_\ell G_\ell + \gamma_1 \mathrm{PC}_1 + ... + \gamma_k \mathrm{PC}_k + \epsilon$

where ID contains the test axis of interest as measured directly in the GWAS panel (either population ID, latitude or the single deme of interest). This set-up allows us to compare our covariate $\hat{F}_{Gr}$ to a GWAS done with no correction and one done with perfect correction for the test of interest. Additionally, we compared our results to a model that includes some number of top genetic principal components in the GWAS panel. All GWASs were done using the plink2 [68] function `--glm`.

We then ascertain 200 SNPs to do our association test by picking the SNP with the lowest p-value per chromosome. Additionally, for heritable trait simulations, we know the identity of the underlying causal sites and can choose to use those sites (Figure 2). For each set of sites, either those chosen by p-value or causal sites, we use R to re-estimate effect sizes jointly using the `lm()` under the four models above where we instead include all 200 SNPs along with the covariates.

5.1.4 Polygenic Score Association Test

For each replicate we first compute the test statistic $\hat{q} = \frac{\hat{\beta}^\top r^*}{\sqrt{N \hat{V}_A^{PGS}}}$ by multiplying the set of ascertained effect size estimates by r^* for corresponding site, and calculating $V_A^{PGS} = 4 \sum_{\ell=1}^{S} \hat{\beta}_\ell p_\ell (1-p_\ell)$ where p_ℓ is the allele frequency in the test panel.

To evaluate the significance of our association statistic, we compute p-values using an empirical null approach. Specifically, we obtain psuedosamples of $\hat{q}$ by randomizing the sign of the effect sizes and recomputing $\hat{q}$. We then obtain a p-value from this sign-flipping null as,

$$P = 2 * \frac{\sum_i^N I(|\hat{q}_i| > |\hat{q}_{obs}|)}{K}, \tag{30}$$

where K is the number of psuedosamples, here set to 1,000, and $I()$ is an indicator function that equals 1 if the psuedosample $|\hat{q}_i|$ is greater than the observed $||hat q_{obs}|$. Finally within a set of 100 evolutionary replicates for each model and confounder combination we compute the positive rate as the proportion of p-values $<= 0.05$.

Code availability

All of the code developed to produce the figures and simulations in this paper is available in the github repository: `https://github.com/jgblanc/polygenic_adaptation_stratification`.

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 for comments on the manuscript. We also acknowledge funding support from the National Human Genome Research Institute (F31HG011821 to JGB).

References

1. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748–752. https://doi.org/10.1038/nature08185.
2. Lander ES, Schork NJ. Genetic dissection of complex traits. Science (New York, NY). 1994;265(5181):2037–2048. https://doi.org/10.1126/science.8091226.
3. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006;38(8):904–909. https://doi.org/10.1038/ng1847.
4. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010;42(4):348–354. https://doi.org/10.1038/ng.548.

5. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics. 2015;47(3):284–290. https://doi.org/10.1038/ng.3190.

6. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. 2015;47(3):291–295. https://doi.org/10.1038/ng.3211.

7. Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, Crawford A, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? Human Genetics. 2020;139(1):23–41. https://doi.org/10.1007/s00439-019-02014-8.

8. Kerminen S, Martin AR, Koskela J, Ruotsalainen SE, Havulinna AS, Surakka I, et al. Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. The American Journal of Human Genetics. 2019;104(6):1169–1181. https://doi.org/10.1016/j.ajhg.2019.05.001.

9. Abdellaoui A, Hugh-Jones D, Yengo L, Kemper KE, Nivard MG, Veul L, et al. Genetic correlates of social stratification in Great Britain. Nature Human Behaviour. 2019;3(12):1332–1342. https://doi.org/10.1038/s41562-019-0757-5.

10. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nature Communications. 2019;10(1):333. https://doi.org/10.1038/s41467-018-08219-1.

11. Trochet H, Pelletier J, Tadros R, Hussin J. Comparison of polygenic risk scores for heart disease highlights obstacles to overcome for clinical use. bioRxiv; 2021. `https://www.biorxiv.org/content/10.1101/2020.08.09.243287v2`.

12. Rosenberg NA, Edge MD, Pritchard JK, Feldman MW. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. Evolution, Medicine, and Public Health. 2019;2019(1):26–34. https://doi.org/10.1093/emph/eoy036.

13. Latta R. Differentiation of Allelic Frequencies at Quantitative Trait Loci Affecting Locally Adaptive Traits. The American Naturalist. 1998;151(3):283–292. https://doi.org/10.1086/286119.

14. Latta RG. Gene flow, adaptive population divergence and comparative population structure across loci. New Phytologist. 2004;161(1):51–58. https://doi.org/10.1046/j.1469-8137.2003.00920.x.

15. Pritchard JK, Di Rienzo A. Adaptation – not by sweeps alone. Nature Reviews Genetics. 2010;11(10):665–667. https://doi.org/10.1038/nrg2880.

16. Pritchard JK, Pickrell JK, Coop G. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology. 2010;20(4):R208–R215. https://doi.org/10.1016/j.cub.2009.11.055.

17. Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nature Genetics. 2012;44(9):1015–1019. https://doi.org/10.1038/ng.2368.

18. Berg JJ, Coop G. A Population Genetic Signal of Polygenic Adaptation. PLoS Genetics. 2014;10(8):e1004412. https://doi.org/10.1371/journal.pgen.1004412.

19. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760–764. https://doi.org/10.1126/science.aag0776.

20. Racimo F, Berg JJ, Pickrell JK. Detecting Polygenic Adaptation in Admixture Graphs. Genetics. 2018;208(4):1565–1584. https://doi.org/10.1534/genetics.117.300489.

21. Josephs EB, Berg JJ, Ross-Ibarra J, Coop G. Detecting Adaptive Differentiation in Structured Populations with Genomic Data and Common Gardens. Genetics. 2019;211(3):989–1004. https://doi.org/10.1534/genetics.118.301786.

22. Uricchio LH, Kitano HC, Gusev A, Zaitlen NA. An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evolution Letters. 2019;3(1):69–79. https://doi.org/10.1002/evl3.97.

23. Edge MD, Coop G. Reconstructing the History of Polygenic Scores Using Coalescent Trees. Genetics. 2019;211(1):235–262. https://doi.org/10.1534/genetics.118.301687.

24. Stern AJ, Speidel L, Zaitlen NA, Nielsen R. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. The American Journal of Human Genetics. 2021;108(2):219–239.

25. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832–838. https://doi.org/10.1038/nature09410.

26. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics. 2014;46(11):1173–1186. https://doi.org/10.1038/ng.3097.

27. Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, et al. Population genetic differentiation of height and body mass index across Europe. Nature Genetics. 2015;47(11):1357–1362. https://doi.org/10.1038/ng.3401.

28. Zoledziewska M, Sidore C, Chiang CWK, Sanna S, Mulas A, Steri M, et al. Height-reducing variants and selection for short stature in Sardinia. Nature Genetics. 2015;47(11):1352–1356. https://doi.org/10.1038/ng.3403.

29. Berg JJ, Zhang X, Coop G. Polygenic Adaptation has Impacted Multiple Anthropometric Traits. Evolutionary Biology; 2017. `http://biorxiv.org/lookup/doi/10.1101/167551`.

30. Guo J, Wu Y, Zhu Z, Zheng Z, Trzaskowski M, Zeng J, et al. Global genetic differentiation of complex traits shaped by natural selection in humans. Nature communications. 2018;9(1):1–9.

31. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499–503.

32. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. eLife. 2019;8:e39725. https://doi.org/10.7554/eLife.39725.

33. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife. 2019;8:e39702. https://doi.org/10.7554/eLife.39702.

34. Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, et al. Evidence of Polygenic Adaptation in Sardinia at Height-Associated Loci Ascertained from the Biobank Japan. The American Journal of Human Genetics. 2020;107(1):60–71. https://doi.org/10.1016/j.ajhg.2020.05.014.

35. Howe LJ, Nivard MG, Morris TT, Hansen AF, Rasheed H, Cho Y, et al. Within-sibship GWAS improve estimates of direct genetic effects; 2021. `https://www.biorxiv.org/content/10.1101/2021.03.05.433935v1`.

36. Tucci S, Vohr SH, McCoy RC, Vernot B, Robinson MR, Barbieri C, et al. Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia. Science. 2018;361(6401):511–516. https://doi.org/10.1126/science.aar8486.

37. Le MK, Smith OS, Akbari A, Harpak A, Reich D, Narasimhan VM. 1,000 ancient genomes uncover 10,000 years of natural selection in Europe. bioRxiv. 2022; p. 2022–08.

38. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human demographic history impacts genetic risk prediction across diverse populations. The American Journal of Human Genetics. 2017;100(4):635–649.

39. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nature communications. 2020;11(1):1–9.

40. Carlson MO, Rice DP, Berg JJ, Steinrücken M. Polygenic score accuracy in ancient samples: Quantifying the effects of allelic turnover. PLOS Genetics. 2022;18(5):e1010170. https://doi.org/10.1371/journal.pgen.1010170.

41. Yair S, Coop G. Population differentiation of polygenic score predictions under stabilizing selection. Philosophical Transactions of the Royal Society B: Biological Sciences. 2022;377(1852):20200416. https://doi.org/10.1098/rstb.2020.0416.

42. Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations. Genetics; 2022. `http://biorxiv.org/lookup/doi/10.1101/2022.09.28.509988`.

43. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nature genetics. 2019;51(4):584–591.

44. Abdellaoui A, Dolan CV, Verweij KJH, Nivard MG. Gene–environment correlations across geographic regions affect genome-wide association studies. Nature Genetics. 2022; p. 1–10. https://doi.org/10.1038/s41588-022-01158-0.

45. Le MK, Smith OS, Akbari A, Harpak A, Reich D, Narasimhan VM. 1,000 ancient genomes uncover 10,000 years of natural selection in Europe; 2022. `https://www.biorxiv.org/content/10.1101/2022.08.24.505188v1`.

46. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. The American Journal of Human Genetics. 2017;100(4):635–649. https://doi.org/10.1016/j.ajhg.2017.03.004.

47. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. eLife. 2020;9:e48376. https://doi.org/10.7554/eLife.48376.

48. Dahl A, Guillemot V, Mefford J, Aschard H, Zaitlen N. Adjusting for Principal Components of Molecular Phenotypes Induces Replicating False Positives. Genetics. 2019;211(4):1179–1189. https://doi.org/10.1534/genetics.118.301768.

49. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature genetics. 2015;47(3):284–290.

50. Li MX, Yeung JMY, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Human Genetics. 2012;131(5):747–756. https://doi.org/10.1007/s00439-011-1118-2.

51. Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, et al. Population genetic differentiation of height and body mass index across Europe. Nature genetics. 2015;47(11):1357–1362.

52. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489–494.

53. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient Admixture in Human History. Genetics. 2012;192(3):1065–1093. https://doi.org/10.1534/genetics.112.145037.

54. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nature Genetics. 2012;44(3):243–246. https://doi.org/10.1038/ng.1074.

55. Zaidi AA, Mathieson I. Demographic history impacts stratification in polygenic scores. Genetics; 2020. `http://biorxiv.org/lookup/doi/10.1101/2020.07.20.212530`.

56. Cox SL, Nicklisch N, Francken M, Wahl J, Meller H, Haak W, et al.. Socio-cultural practices affect sexual dimorphism in stature in Early Neolithic Europe; 2023. https://www.biorxiv.org/content/10.1101/2023.02.21.529406v1.

57. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature Genetics. 2015;47(11):1236–1241. https://doi.org/10.1038/ng.3406.

58. Sarmanova A, Morris T, Lawson DJ. Population stratification in GWAS meta-analysis should be standardized to the best available reference datasets; 2020. https://www.biorxiv.org/content/10.1101/2020.09.03.281568v1.

59. Stern AJ, Speidel L, Zaitlen NA, Nielsen R. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. The American Journal of Human Genetics. 2021;108(2):219–239. https://doi.org/10.1016/j.ajhg.2020.12.005.

60. Song W, Shi Y, Wang W, Pan W, Qian W, Yu S, et al. A selection pressure landscape for 870 human polygenic traits. Nature Human Behaviour. 2021;5(12):1731–1743. https://doi.org/10.1038/s41562-021-01231-4.

61. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale sequencing data. PLOS Biology. 2020;18(1):e3000586. https://doi.org/10.1371/journal.pbio.3000586.

62. Novembre J, Barton NH. Tread Lightly Interpreting Polygenic Tests of Selection. Genetics. 2018;208(4):1351–1355. https://doi.org/10.1534/genetics.118.300786.

63. Veller C, Coop G. Interpreting population and family-based genome-wide association studies in the presence of confounding; 2023. https://www.biorxiv.org/content/10.1101/2023.02.26.530052v1.

64. Harpak A, Przeworski M. The evolution of group differences in changing environments. PLOS Biology. 2021;19(1):e3001072. https://doi.org/10.1371/journal.pbio.3001072.

65. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS computational biology. 2016;12(5):e1004842.

66. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome research. 2013;23(9):1514–1521.

67. Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on polygenic scores. Elife. 2020;9:e61548.

68. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742–015.

6 Supplementary Material

6.1 Expectation of polygenic scores using marginal effects

Conditional on the GWAS panel genotypes and the genetic and environmental components of the phenotype, the marginal association at site ℓ is given by

$$\begin{aligned}\hat{\beta}_\ell \mid G_\ell, u, e &= \frac{y^\top G_\ell}{G_\ell^\top G_\ell} \\ &= \frac{u^\top G_\ell}{G_\ell^\top G_\ell} + \frac{e^\top G_\ell}{G_\ell^\top G_\ell} \\ &= \sigma_\ell + \epsilon_\ell.\end{aligned} \tag{31}$$

(equation (7) in the main text). Thus, conditioning on all variables in the GWAS panel and the genotypes in the test panel, the polygenic scores are given by

$$Z^\top \mid \mathbf{G}, u, e = \sum_{\ell=1}^{S} \hat{\beta}_\ell X_\ell^\top \tag{32}$$

$$= \sum_{\ell=1}^{S} \sigma_\ell X_\ell^\top + \sum_{\ell=1}^{S} \epsilon_\ell X_\ell^\top \tag{33}$$

$$= \sum_{\ell=1}^{S} \frac{u^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top + \sum_{\ell=1}^{S} \frac{e^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top. \tag{34}$$

Now, taking the expectation over the joint distribution of genotypes in the GWAS and test panels, the expected polygenic scores conditional only on u and e are

$$\mathbb{E}_{\mathbf{G},\mathbf{X}}\left[Z^\top \mid u, e\right] = \sum_{\ell=1}^{S} u^\top \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell}\right] + \sum_{\ell=1}^{S} e^\top \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell}\right] \tag{35}$$

$$= S\left(u^\top \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell}\right] + e^\top \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell}\right]\right) \tag{36}$$

$$= \frac{S}{M-1}\left(u^\top \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell/(M-1)}\right] + e^\top \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell/(M-1)}\right]\right) \tag{37}$$

$$= \frac{S}{M-1}\left(u^\top + e^\top\right)\tilde{\mathbf{F}}_{GX} \tag{38}$$

where

$$\tilde{\mathbf{F}}_{GX} = \mathbb{E}_{G_\ell, X_\ell}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell/(M-1)}\right]. \tag{39}$$

Finally, taking the expectations over u and e, we have

$$\mathbb{E}\left[Z^\top\right] = \frac{S}{M-1}\left(v^\top + c^\top\right)\tilde{\mathbf{F}}_{GX} \tag{40}$$

6.2 Expectation of polygenic scores while controlling for $\tilde{\mathbf{F}}_{Gr}$

Now, controlling for $\tilde{\mathbf{F}}_{\mathbf{Gr}}$, the marginal effects are

$$\begin{aligned}\hat{\beta}'_\ell \mid G_\ell, u, e &= \frac{u^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} + \frac{e^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} \\ &= \sigma'_\ell + \epsilon'_\ell,\end{aligned} \tag{41}$$

where $\mathbf{P} = \left(\mathbf{I} - \frac{1}{\|\tilde{F}_{Gr}\|}\tilde{F}_{Gr}\tilde{F}_{Gr}^\top\right)$

Thus, conditional on the variables in the GWAS panel, the polygenic scores are

$$Z^\top \mid \mathbf{G}, u, e = \sum_{\ell=1}^{S} \hat{\beta}'_\ell X_\ell^\top \tag{42}$$

$$= \sum_{\ell=1}^{S} \sigma'_\ell X_\ell^\top + \sum_{\ell=1}^{S} \epsilon'_\ell X_\ell^\top \tag{43}$$

$$= \sum_{\ell=1}^{S} \frac{u^\top \mathbf{P} G_\ell X_\ell^\top}{G_\ell^\top \mathbf{P} G_\ell} + \sum_{\ell=1}^{S} \frac{e^\top \mathbf{P} G_\ell X_\ell^\top}{G_\ell^\top \mathbf{P} G_\ell}. \tag{44}$$

Taking the expectation over the the joint distribution of genotypes in the two panels, the expected polygenic scores conditional on u and e are

$$\mathbb{E}_{\mathbf{G},\mathbf{X}}\left[Z^\top \mid u, e\right] = \sum_{\ell=1}^{S} u^\top \mathbf{P} \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top \mathbf{P} G_\ell}\right] + \sum_{\ell=1}^{S} e^\top \mathbf{P} \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top \mathbf{P} G_\ell}\right] \tag{45}$$

$$= S\left(u^\top + e^\top\right) \mathbf{P} \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top \mathbf{P} G_\ell}\right] \tag{46}$$

$$\approx \frac{S}{M-1}\left(u^\top + e^\top\right) \mathbf{P} \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{G_\ell^\top G_\ell / M-1}\right] \tag{47}$$

$$\approx \frac{S}{M-1}\left(u^\top + e^\top\right) \mathbf{P} \tilde{\mathbf{F}}_{GX} \tag{48}$$

$$\approx \frac{S}{M-1}\left(u^\top + e^\top\right) \left(\tilde{\mathbf{F}}_{GX} - \frac{1}{\|\tilde{F}_{Gr}\|} \tilde{F}_{Gr} \tilde{F}_{Gr}^\top \tilde{\mathbf{F}}_{GX}\right) \tag{49}$$

$$\approx \frac{S}{M-1}\left(u^\top + e^\top\right) \tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \tag{50}$$

where the approximation sign comes from assuming that we can ignore the $\mathbf{P}$ in the denominator, which is reasonable so long as $\tilde{F}_{Gr}$ explains only a small fraction of the variance in the GWAS panel genotypes. Taking the expectation over u and e, we have

$$\mathbb{E}\left[Z^\top\right] \approx \frac{S}{M-1}\left(v^\top + c^\top\right) \tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \tag{51}$$

6.3 Expectation of $\hat{q}$ while controlling for $\tilde{\mathbf{F}}_{Gr}$

The expectation of $\hat{q}$ then becomes

$$\mathbb{E}\left[\hat{q}\right] \approx \frac{S}{(N-1)(M-1)} \left(v^\top + c^\top\right) \tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} T \tag{52}$$

$$\approx \frac{S}{(N-1)(M-1)} \left(v^\top + c^\top\right) \left(\tilde{\mathbf{F}}_{GX} - \frac{1}{\|\tilde{F}_{Gr}\|} \tilde{F}_{Gr} \tilde{F}_{Gr}^\top \tilde{\mathbf{F}}_{GX}\right) T \tag{53}$$

$$\approx \frac{S}{(N-1)(M-1)} \left(v^\top + c^\top\right) \left(\tilde{\mathbf{F}}_{GX} T - \frac{1}{\|\tilde{F}_{Gr}\|} \tilde{F}_{Gr} \tilde{F}_{Gr}^\top \tilde{\mathbf{F}}_{GX} T\right) \tag{54}$$

$$\approx \frac{S}{(N-1)(M-1)} \left(v^\top + c^\top\right) \left(\tilde{F}_{Gr} - \frac{1}{\|\tilde{F}_{Gr}\|} \tilde{F}_{Gr} \tilde{F}_{Gr}^\top \tilde{F}_{Gr}\right) \tag{55}$$

$$\approx \frac{S}{(N-1)(M-1)} \left(v^\top + c^\top\right) \left(\tilde{F}_{Gr} - \tilde{F}_{Gr}\right) \tag{56}$$

$$\approx 0 \tag{57}$$

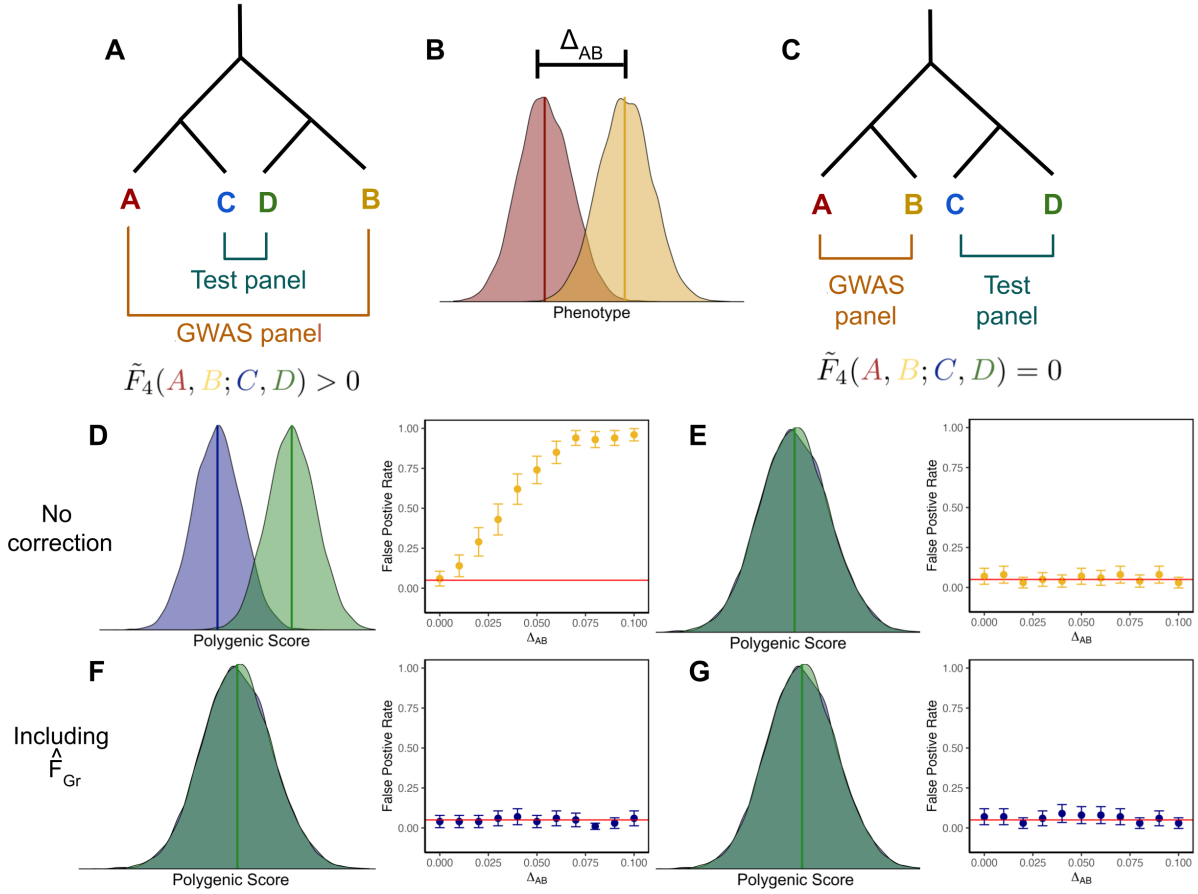

Figure 1: **Schematic of two different panel configurations. The effect of stratification depends on the overlapping structure between the GWAS and test panels.** (A) and (C) Two different topologies used to create the GWAS and test panels. (B) Stratification was modeled in the GWAS panel by drawing 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 false positive rate for $\hat{q}$ increases with the magnitude of stratification in the GWAS. (E) However, when there is no overlapping structure between panels, there is expected difference in mean polygenic scores between C and D and the false positive rate remains at 5% regardless of the magnitude of stratification. (F) and (G) Including $\hat{F}_{Gr}$ as a covariate in the GWAS controls for stratification, keeping the false positive at nominal levels regardless of Δ_{AB} or the overlapping structure between GWAS and test panels. False positive rates were calculated as the number of simulations with $p < 0.05$ out of 100 simulations for each Δ_{AB}.

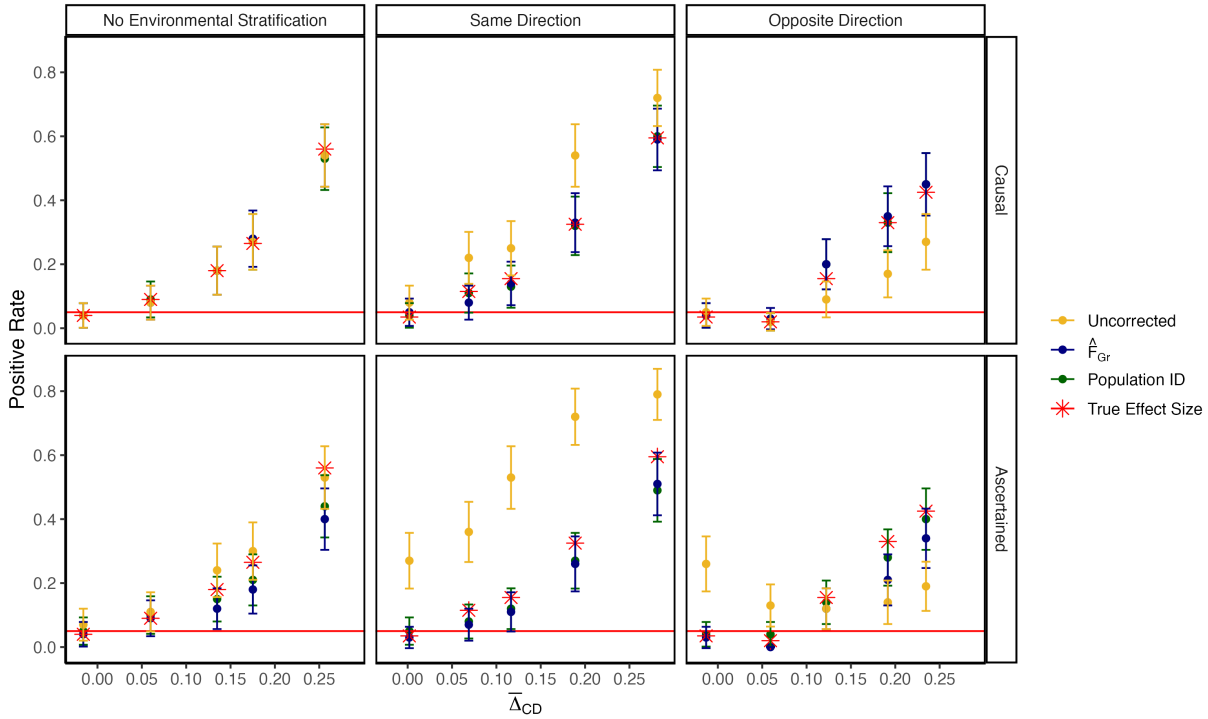

Figure 2: **Including $\hat{F}_{Gr}$ as a covariate in the GWAS model does not regress out true signal.** Here we generate a true difference in average polygenic scores by flipping the sign of a proportion of causal effect sizes to align with allele frequency contrasts, $p_{D,\ell} - p_{C,\ell}$, in the test panel. $\bar{\Delta}_{CD}$ is the true difference in polygenic scores averaged across 100 replicates. The top row is the positive rate in our association test using estimated effect sizes at causal sites. Regardless of the direction of the stratification effect, our correction procedure perfectly captures the real signal (computed using the true effect sizes). The bottom row is the positive rate when polygenic scores were computed using sites ascertained on p-value. While there is some loss of power due to imperfect tagging, our approach maintains power to detect true signal.

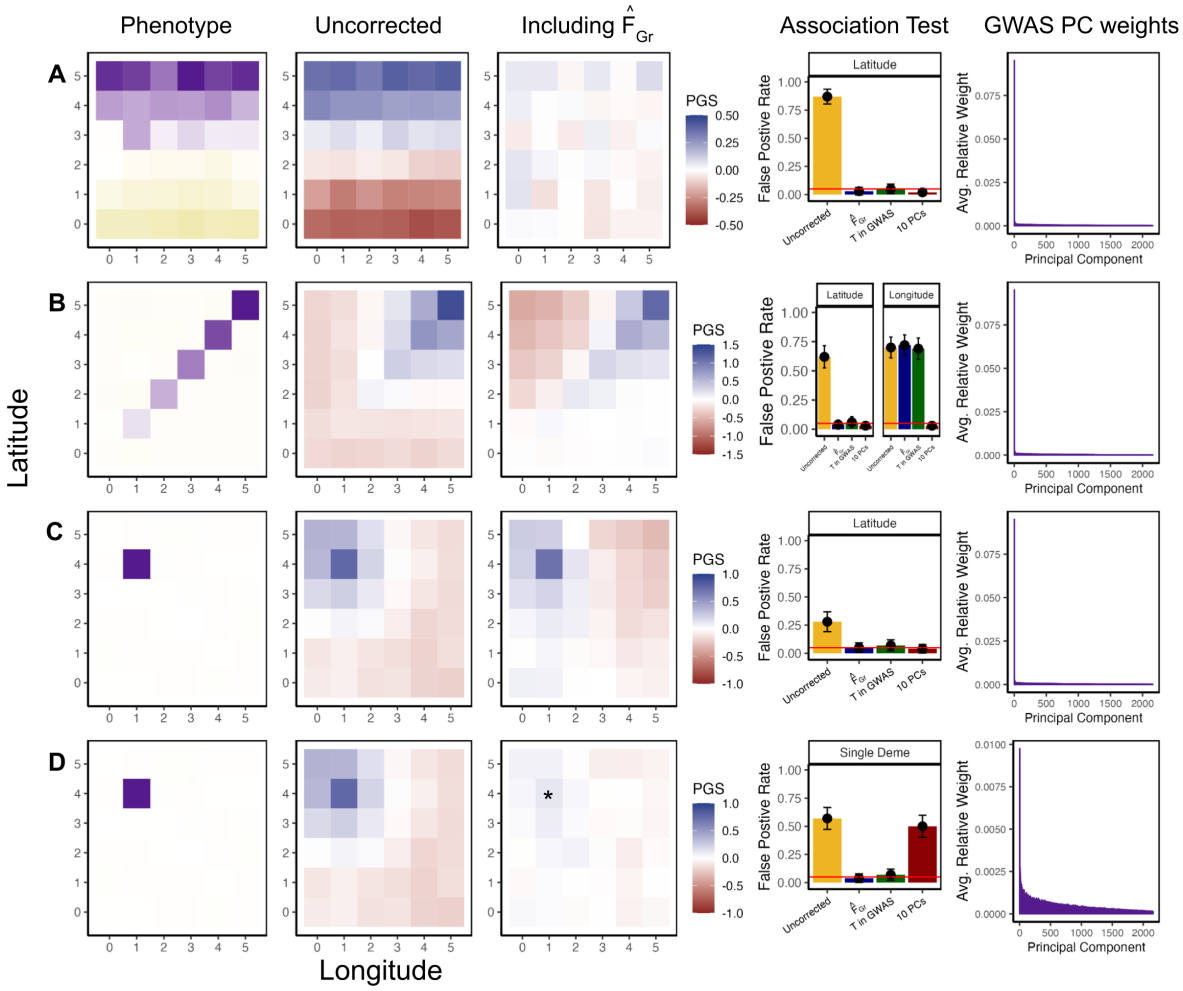

Figure 3: **Including $\hat{F}_{Gr}$ as a covariate in the GWAS model controls for stratification bias specific to the association test of interest.** GWAS and test panel individuals were simulated using stepping-stone model with continuous migration. In the GWAS panel the phenotype is non-heritable and stratified along either latitude (A), the diagonal (B), or in a single deme (C, D). When effect sizes are estimated in a GWAS with no correction for stratification, polygenic scores constructed in the test panel recapitulate the spatial distribution of the confounder (second column). Including $\hat{F}_{Gr}$ (latitude for A-C, belonging to * deme for D) in the GWAS model eliminates bias in polygenic scores along the test axis (third column) and the false positive rate for $\hat{q}$ across 100 replicates remains at nominal levels (fourth column). We also compare our approach to including the top 10 PCs and including the test axis, as define in the GWAS panel (latitude for A-C, single deme for D), directly. Finally we plot the relative weight of all $M-1$ GWAS panel PCs on $\hat{F}_{Gr}$, averaged across 100 replicates (fifth column). When the test axis is latitude (A-C), $\hat{F}_{Gr}$ loads mainly onto one of the first two PCs, while a single deme test (D) has weight on the full range of PCs.

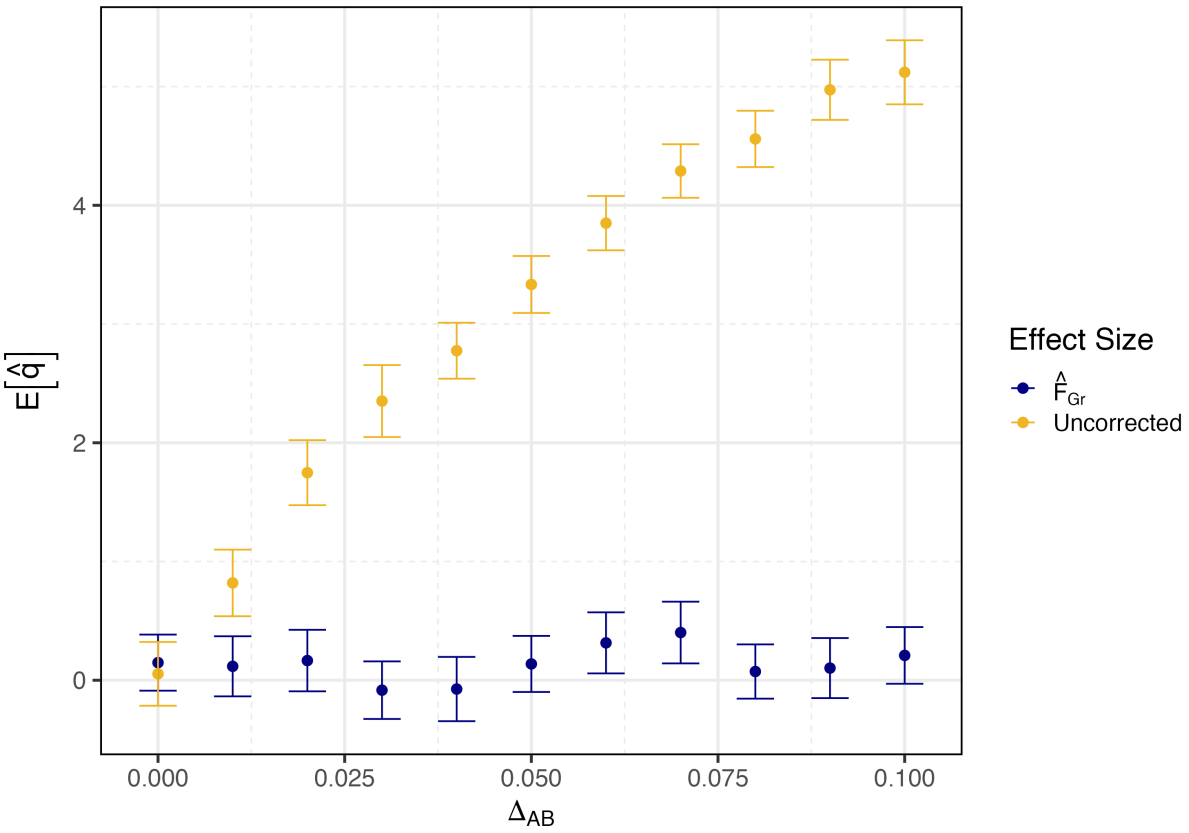

Figure S1: **Including $\hat{F}_{Gr}$ as a covariate removes bias in $\hat{q}$.** In our simulations of the 4 population toy model with overlapping topology (Figure 1A), increasing the magnitude of the expected mean difference in phenotypes for populations A and B, increases the expectation of $\hat{q}$ when effect sizes are uncorrected. When we include $\hat{F}_{Gr}$ as a covariate in the GWAS, $\mathbb{E}[\hat{q}]$ is unbiased.

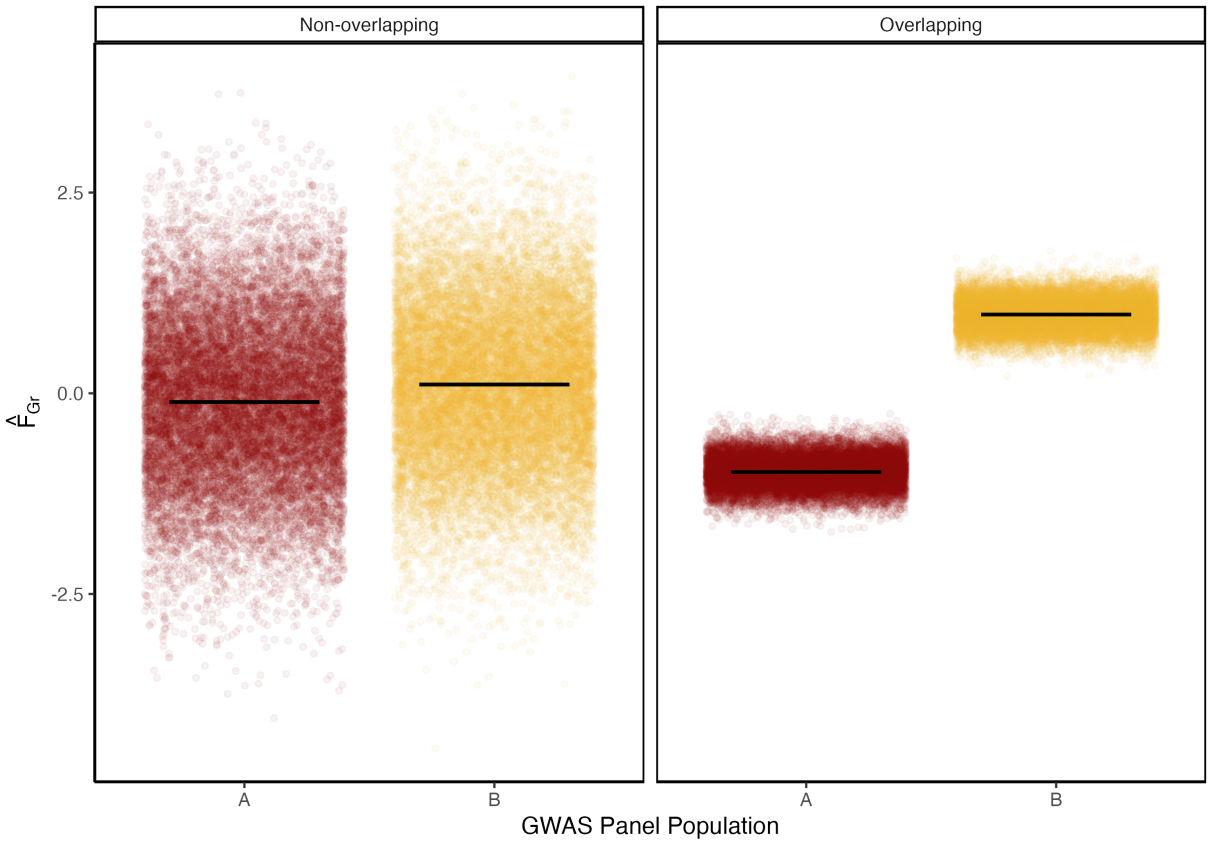

Figure S2: $\hat{F}_{Gr}$ **separates populations in the GWAS panel when test and GWAS panels have overlapping structure.** In the 4 population toy model, when the topology is non-overlapping (Figure 1C), $\hat{F}_{Gr}$, shown here for a single replicate, is purely noise and does not reflect population membership, due to the fact the genotype contrasts in the test panel are already uncorrelated with structure in the GWAS panel. When the topology is overlapping (Figure 1A), $\hat{F}_{Gr}$ separates individuals in the GWAS panel based on population membership.

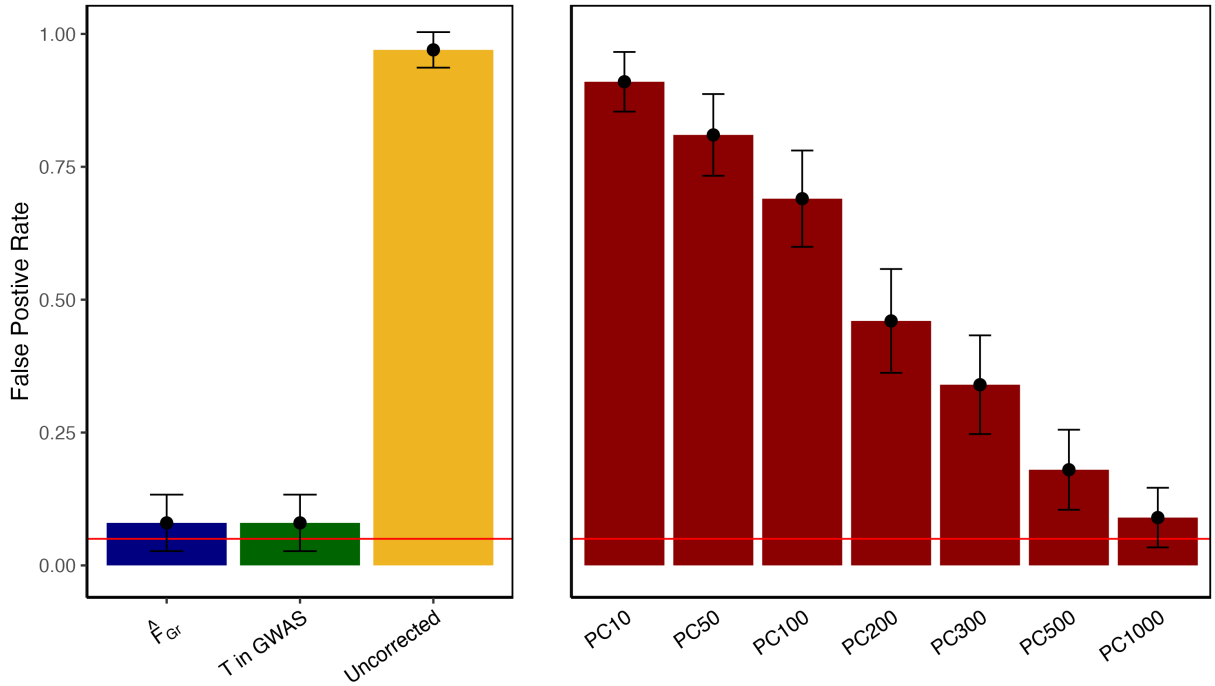

Figure S3: $\hat{F}_{Gr}$ **controls for stratification for a single deme association test.** Here we simulate a 2 standard deviation confounder in a single deme in the GWAS panel (Figure 3D) and conduct an association test for the difference in polygenic score between the same single deme and the rest of the grid in the test panel. Including $\hat{F}_{Gr}$ or the actual test vector, as defined in the GWAS individuals, as a covariate in the GWAS model, controls for stratification and the resulting test is unbiased. We then vary the number of principal components of the GWAS panel included as covariates, demonstrating that it takes 1,000 PCs to fully control for stratification.

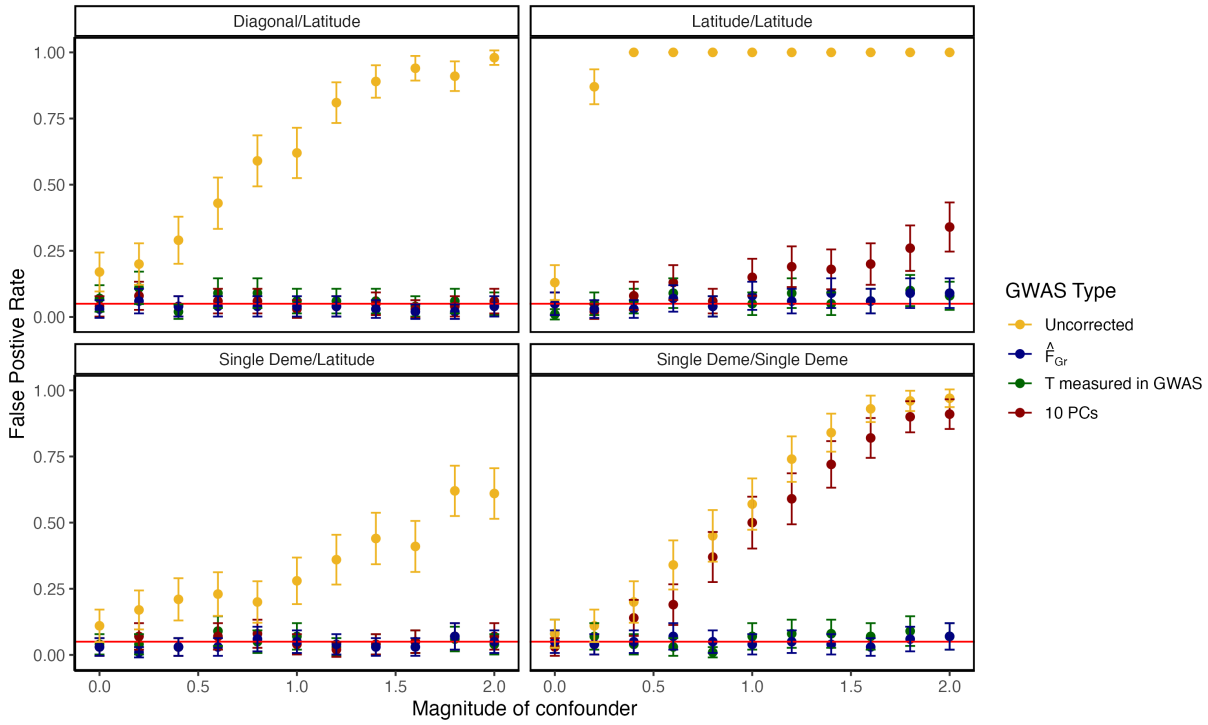

Figure S4: $\hat{F}_{Gr}$ **controls for stratification for a range confounding effects.** For the for simulation scenarios in Figure 3 we plot the false positive rate for the given association test for a range of confounding effects where the magnitude of the confounder is the difference between the most extreme phenotypic shifts in units of standard deviations. For all situations tested, our approach protects the test from stratification bias, mirroring the gold standard of measuring the test vector itself in GWAS panel individuals. In two confounder/test scenarios (right column) the standard approach of including the top 10 GWAS panel principal components begins to fail with large enough confounding.

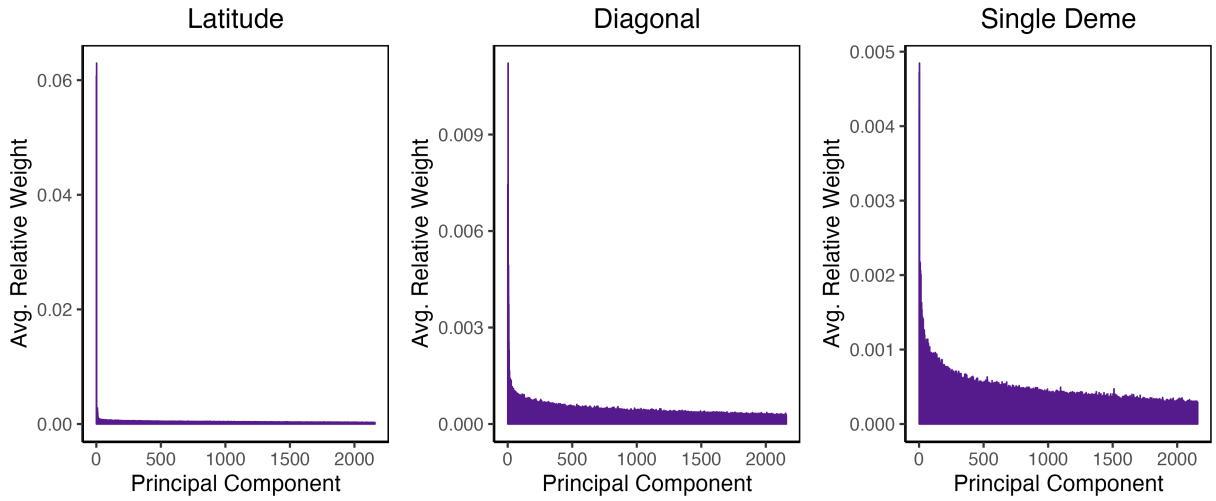

Figure S5: **Different confounding patterns load onto different GWAS principal components.** For the three possible confounding patterns tested in Figure 3 we plot the average relative loading of the confounder onto each of the $M-1$ GWAS panel principal components averaged across all 100 replicates. As the stratification pattern gets more complex, the confounder is more likely to load onto lower PCs, reflecting more subtle structure in the GWAS panel.