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 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 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 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. Finally, we use the understanding gained from this analysis to develop a method that uses patterns of genetic similarity between the two panels to guard against these biases, and show that this method can provide better protection against confounding than the standard PCA-based approach.

Author Summary

Complex traits are influenced by both genetics and the environment. Human geneticists increasingly use polygenic scores, calculated as the weighted sum of trait-associated alleles, to predict genetic effects on a phenotype. Differences in polygenic scores across groups would therefore seem to indicate differences in the genetic basis of the trait, which are of interest to researchers across disciplines. However, because polygenic scores are usually computed using effect sizes estimated using population samples, they are susceptible to confounding due to both the genetic background and the environment. Here, we use theory from population and

statistical genetics, together with simulations, to study how environmental and background genetic effects can confound tests for association between polygenic scores and axes of ancestry variation. We then develop a simple method to protect these tests from confounding and show that our approach succeeds in situations that are difficult to protect with standard methods. Our work helps clarify how bias in the distribution of polygenic scores is produced and provides a tool to researchers wishing to protect their analyses from confounding.

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 then 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 assumptions that genetic and environmental effects combine additively, 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 these assumptions are met, polygenic scores would seem to provide a means of separating out 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 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, and statistical bias correction approaches, such as 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 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 [12] or phenotype biased migration [9]. Medically, it is interesting to know if polygenic score differences or gradients represent real underlying gradients

in the average genetic effect [13], whether those gradients are caused by non-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 [14, 15, 16, 17]. Several research groups have developed and applied methods to detect these signals [18, 12, 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, 18, 12, 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, 23]. 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].

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 [35] 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 [36, 37, 38, 39, 40]. Less accurate polygenic scores are then less able to capture evolution of the mean polygenic score, all else equal [39]. 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 [41]. 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 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 and axes of ancestry variation. We model the covariance of genotypes in a GWAS and test panel in terms of an underlying population genetic model, and write down 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 demonstrate that including this particular axis of ancestry variation 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 show how to interpret the

standard PCA-based approach to controlling for stratification as an effort to capture this particular axis and control for it, and discuss reasons why this approach may fail. Finally, we propose a more direct approach that uses the test panel genotypes to estimate the necessary covariate, and we show using simulations that our approach can succeed in setting where PCA fails.

2 Model

2.1 Genotypes

To model the distribution of genotypes in both panels, we assume that each individual's expected genotype at each site can be modelled 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. We begin by first developing the population level model, before extending it to individuals.

2.1.1 Population model

We assume that for each site ℓ, the vector of allele frequencies across K populations can be decomposed as

$$p_\ell = \lambda_\ell + \Delta p_{\ell,D} + \Delta p_{\ell,S} \tag{1}$$

where λ_ℓ is the allele frequency in the original ancestral population, while the deviations $\Delta p_{\ell,D}$ and $\Delta p_{\ell,S}$ capture variation in allele frequency across populations due to genetic drift and natural selection respectively. We approximate the effect of drift via a multivariate Normal model, such that $\Delta p_{\ell,D} \sim MVN\left(0, \lambda_\ell\left(1-\lambda_\ell\right)\mathbf{F}_{pop}\right)$, where $\mathbf{F}_{pop}$ describes the covariance structure imposed by the population model [42, 43, 44, 45]. Selection induces an additional deviation of $\Delta p_{\ell,S} = \alpha\beta_\ell\lambda_\ell\left(1-\lambda_\ell\right)A_{pop}$, where β_ℓ is site ℓ's effect on the phenotype of interest, α is the strength of selection on that phenotype, and A_{pop} is a vector recording the extent to which each population inherits from the ancestral population in which selection occurred.

2.1.2 Sampling individuals for GWAS and test panels

Next, we consider taking two samples, one to compose the test panel and one to compose the GWAS panel. Individuals in each panel are created as mixtures of the underlying populations. There are N test panel individuals and the deviation of their genotypes from the expected genotype in the ancestral population ($2\lambda_\ell$) is

$$X_\ell = X_{\ell,D} + X_{\ell,S} + X_{\ell,B}, \tag{2}$$

where $X_{\ell,D} = 2\mathbf{W}_X\Delta p_{\ell,D}$, $X_{\ell,S} = 2\mathbf{W}_X\Delta p_{\ell,S}$. The matrix $\mathbf{W}_X$ has dimensions $N \times K$, with rows specifying the fraction of ancestry that each of the N test panel individuals inherit from

each of the K populations. We can think of the quantity $2\lambda_\ell + X_{\ell,D} + X_{\ell,S}$ as giving a set of expected genotypes given the evolutionary history of the population, while $X_{\ell,B}$ contains the binomial sampling deviations across individuals given these expected genotypes.

Similarly, for the M GWAS panels individuals, the deviation of their genotypes can be decomposed as

$$G_\ell = G_{\ell,D} + G_{\ell,S} + G_{\ell,B}, \tag{3}$$

where $G_{\ell,D} = 2\mathbf{W}_G \Delta p_{\ell,D}$, $G_{\ell,S} = 2\mathbf{W}_G \Delta p_{\ell,S}$, W_G is an $M \times K$ matrix with rows specifying the amount of ancestry that each GWAS panel individual inherits from each of the K populations and $G_{\ell,B}$ captures the binomial sampling variance given the expected genotypes of the GWAS panel individuals (similar to above, $2\lambda_\ell + G_{\ell,D} + G_{\ell,S}$ specifies this set of individual specific expected genotypes).

2.1.3 Individual level model

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

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

where the matrix

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{XX} & \mathbf{F}_{XG} \\ \mathbf{F}_{GX} & \mathbf{F}_{GG} \end{bmatrix} \tag{5}$$

contains the within and between panel relatedness coefficients. These are in turn related to the population level covariance matrix as $\mathbf{F}_{XX} = W_X \mathbf{F}_{pop} W_X^\top$, $\mathbf{F}_{GG} = W_G \mathbf{F}_{pop} W_G^\top$, and $\mathbf{F}_{GX} = \mathbf{F}_{XG}^\top = W_G \mathbf{F}_{pop} W_X^\top$. Similarly, the genotypic deviations due to selection can be written at the individual level as $X_{\ell,S} = 2\alpha\beta_\ell\lambda_\ell\left(1-\lambda_\ell\right)A_X$ and $G_{\ell,S} = 2\alpha\beta_\ell\lambda_\ell\left(1-\lambda_\ell\right)A_G$, where $A_X = \mathbf{W}_X A_{pop}$ and $A_G = \mathbf{W}_G A_{pop}$ describe the extent to which individuals in each of the panels inherit from the selection event. Finally, $X_{\ell,B}$, $G_{\ell,B}$, $X_{\ell,D}$ and $G_{\ell,D}$ all have mean 0, so the expected genotypes are $\mathbb{E}[X_\ell] = X_{\ell,S} = 2\alpha\beta_\ell\lambda_\ell\left(1-\lambda_\ell\right)A_X$ and $\mathbb{E}[G_\ell] = G_{\ell,S} = 2\alpha\beta_\ell\lambda_\ell\left(1-\lambda_\ell\right)A_G$.

2.2 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

$$\begin{aligned} y &= \sum_{\ell}^{S} \beta_\ell G_\ell + e \\ &= u + e \end{aligned} \tag{6}$$

where $u = \sum_{\ell}^{S} \beta_\ell G_\ell$ 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 can differ in some arbitrary but unknown way across individuals. We write the distribution of environmental effects as $e \sim MVN\left(c, \sigma_e^2 \mathbf{I}\right)$, where c is the vector of expected environmental effects.

Similar to our decomposition in eq. 3, the genetic effect, u, can be broken down into the contributions from drift, selection, and sampling such that $u = u_D + u_S + u_B$. Here $u_S = \sum_{\ell}^{S} \beta_\ell G_{\ell,S}$ is the vector of expected values of the genetic contributions to the phenotype, given the ancestries of the individuals in the GWAS panels. Both u_D and u_B have expectation zero and u_D has a correlation structure determined by relatedness matrix $\mathbf{F}_{GG}$ (i.e $u_D \sim MVN(0, 2V_A \mathbf{F}_{GG})$), where $V_A = 2\sum_{\ell}^{S} \beta_\ell^2 \lambda_\ell (1-\lambda_\ell)$ is the additive genetic variance in the original ancestral population. In contrast, the entries of u_B are uncorrelated with each other. Therefore $\mathbb{E}[u] = u_S$ and $u_S + c$ gives the vector of individuals' expected phenotypes, given their ancestry and socio-environmental contexts. We assume that these are not known.

2.3 Polygenic scores

We consider a vector of mean centered polygenic scores, computed in the test panel. If the causal effects were known, then the contribution of the S causal sites included in the polygenic score would be

$$Z = \sum_{\ell}^{S} \beta_\ell X_\ell. \tag{7}$$

Under our genotype model, Z is a vector of random variables with the randomness coming from $X_{\ell,D} + X_{\ell,B}$, neutral variation in the test panel genotypes due drift and binomial sampling, respectively. Therefore, the vector of expected polygenic scores for the N individuals in the test panel is

$$\begin{aligned} \mathbb{E}\left[Z\right] &= \sum_{\ell}^{S} \beta_\ell \mathbb{E}\left[X_\ell\right] \\ &= \sum_{\ell}^{S} \beta_\ell X_{\ell,S} \\ &= \alpha V_A A_X. \end{aligned} \tag{8}$$

2.4 Polygenic score association tests

We want to test the hypothesis that the polygenic scores are associated with some test vector, T, more than is expected due to drift alone. We assume that T is measured only in the test panel, and might represent an eco-geographic variable of interest (e.g latitude [12] or an encoding of whether one lives in a particular geographic region or not [9, 46], the fraction of an individual's genome assigned to a particular "ancestry group"[18, 20], or one of the top genetic principal components

of the test panel genotype matrix [21]). We do not necessarily assume that T is equal to A_X, the axis along which selection has actually perturbed the polygenic scores.

To test for association of polygenic scores with the test vector, we consider the linear model

$$Z = qT + \varepsilon \tag{9}$$

where ε is i.i.d. Normal across individuals. A more powerful test is available by modeling the correlation structure among individuals, but the simpler i.i.d. model is sufficient for our purposes (see section S6). We assume that the test vector, T, is scaled to have a variance of one, so the slope is given by

$$q = \frac{1}{N} Z^\top T, \tag{10}$$

and its expectation by,

$$\begin{aligned} \mathbb{E}\left[q\right] &= \frac{1}{N} \mathbb{E}\left[Z\right]^\top T \\ &= \frac{1}{N} \alpha V_A A_X^\top T \end{aligned} \tag{11}$$

Under the null model, $\alpha = 0$, so $\mathbb{E}[q] = 0$, reflecting the fact that genetic drift is directionless (though we may also have $\mathbb{E}\left[q\right] = 0$ if the test vector is not aligned with the axis along which selection has perturbed the polygenic scores, i.e. if $A_X^\top T = 0$). We compare this null hypothesis to an alternative in which $\alpha \neq 0$ (assuming also that $A_X^\top T \neq 0$), reflecting the possibility that the polygenic score may have either a positive or negative association with the test vector. We refer to this as a "polygenic score association test".

We can also re-frame this test as a statement about the association between the effect sizes 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 [12]. Similar to the decomposition of the test panel genotypes above, we can decompose the genotype contrasts as

$$r_\ell = r_{\ell,D} + r_{\ell,S} + r_{\ell,B}, \tag{12}$$

which capture contributions from drift, selection, and sampling, respectively. Writing β and r for the vectors of effect sizes and genotype contrasts across loci, we can write the test statistic as

$$q = \frac{1}{N} \beta^\top r. \tag{13}$$

The equivalence between eq. 10 and eq. 13 will be useful for illuminating several of our results.

Thus far we have assumed that causal effects are known and therefore the only source of randomness in q comes from the genotypes in the test panel. However, in practice effects sizes need to be estimated (see below) and we instead consider polygenic scores computed using marginal effect size estimates, $\hat{Z} = \sum_\ell^S \hat{\beta}_\ell X_\ell$, and use the test statistic

$$\begin{aligned} \hat{q} &= \frac{1}{N} \hat{Z}^\top T \\ &= \frac{1}{N} \hat{\beta}^\top r, \end{aligned} \tag{14}$$

which will allow us to characterize the bias, $\mathbb{E}[\hat{q} - q]$, in our polygenic score association test.

3 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 the test vector. We first consider the case where no attempt is made to correct for population structure. Motivated by these results, we propose a new correction procedure to remove bias associated with the test vector, and consider several details associated with the implementation 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

$$\begin{aligned}\hat{\beta}_\ell \mid G_\ell, u, e &= \frac{y^\top G_\ell}{G_\ell^\top G_\ell} \\ &= \beta_\ell + \frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} + \frac{e^\top G_\ell}{G_\ell^\top G_\ell}\end{aligned} \tag{15}$$

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 (15) into the causal effect, the association between the focal site and the background genetic contribution from all other sites, and the association with the environment. Under the null hypothesis (i.e. $\alpha = 0$), variation in G_ℓ is uncorrelated with both $u_{-\ell}$ and e, so an allele's expected marginal effect is unbiased over evolutionary replicates (see also Robinson et al. (2015) [27] supplement eq. 2.3). Under the alternative hypothesis ($\alpha \neq 0$), the expected marginal effect is biased away from zero due to the association between the focal site and expected genetic background and may be biased in either direction depending on the vector of expected environmental effects (see section S1).

3.2 Stratification bias induces ancestry associated bias in polygenic scores

The deviation of an allele's estimated effect size from its expectation depends on $G_{\ell,D} + G_{\ell,B}$, the neutral evolutionary and sampling variation around expected genotypes in the GWAS panel, respectively. Because $G_{\ell,D}$ is correlated with $X_{\ell,D}$ (deviations due to drift in 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 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 potential bias in polygenic score association tests. The vector of expected polygenic scores in the test panel is

$$\mathbb{E}\left[\hat{Z}\right]^\top \approx \mathbb{E}\left[Z\right] + \frac{S}{M}\left(\mu_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX} \tag{16}$$

(see section S2) where $\mathbb{E}[Z] = \alpha V_A A_X$ (see eq. 8), $\mu_S = \alpha V_A A_G$ is the vector expected genetic backgrounds, c is the vector of expected environmental effects, and

$$\begin{aligned}\tilde{\mathbf{F}}_{GX} &= \mathbb{E}\left[\frac{G_{\ell,D}X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top(G_{\ell,D}+G_{\ell,B})/M}\right] \\ &\approx \frac{\mathbf{F}_{GX}}{1+\overline{F_G}}.\end{aligned} \tag{17}$$

Here $\overline{F_G} = \frac{1}{M}\sum_{m=1}^{M} F_{mm}$ is the average level of self relatedness in the GWAS panel and $\tilde{\mathbf{F}}_{GX}$ is the expected cross-panel genetic relatedness matrix computed on standardized genotypes, which is approximately equal to $\frac{\mathbf{F}_{GX}}{(1+\overline{F_G})}$ if $\overline{F_G}$ is small.

This result 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, 35], as $\mathbb{E}\left[\hat{Z} - Z\right] = 0$ if $\tilde{\mathbf{F}}_{XG} = \mathbf{0}$. Conditional on a given set of genotypes observed in the GWAS panel, stratification may still bias individual effects, (i.e. $\mathbb{E}\left[\hat{\beta}_\ell | G_\ell\right] \approx \beta_\ell + \frac{\mu_{S,-\ell}^\top G_\ell}{G_\ell^\top G_\ell} + \frac{c^\top G_\ell}{G_\ell^\top G_\ell}$) 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.

3.3 Bias in polygenic scores leads to biased polygenic score associations

The bias in the polygenic score association test statistic ($\hat{q}$) follows straightforwardly from the bias in the polygenic scores,

$$\mathbb{E}[\hat{q} - q] \approx \frac{S}{NM}\left(\mu_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX}T. \tag{18}$$

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 similarity matrix ($\tilde{\mathbf{F}}_{XG}$). 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 $\tilde{\mathbf{F}}_{XG} = \mathbf{0}$, we need only to ensure that certain linear combination of the entries of $\tilde{\mathbf{F}}_{XG}$ are equal to zero, i.e. that $\tilde{\mathbf{F}}_{GX}T = 0$.

We can gain further intuition by recalling that testing for an association between the polygenic scores and the test vector is equivalent to testing for an association between the effect size estimates and the genotype contrasts (see eq. 13). This allows us to rewrite the bias in $\hat{q}$ as

$$\begin{aligned}\mathbb{E}[\hat{q} - q] &= \frac{S}{NM}\mathbb{E}\left[\left(\hat{\beta}^\top - \beta^\top\right)r\right] \\ &\approx \frac{S}{NM}\left(\mu_S^\top + c^\top\right)\tilde{F}_{Gr}\end{aligned} \tag{19}$$

where

$$\tilde{F}_{Gr} = \mathbb{E}\left[\frac{G_{\ell,D} r_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M}\right] \tag{20}$$

$$= \tilde{\mathbf{F}}_{GX} T,$$

$$\approx \frac{W_G \mathbf{F}_{pop} T_{pop}}{1+\overline{F_G}}, \tag{21}$$

where $\mathbf{F}_{pop} T_{pop}$ is a vector of length K, and $T_{pop} = W_X^\top T$ expresses the test vector in terms of contrasts between the underlying K populations.

We gain two main insights here. First, eq. 19 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 (eq. 20). 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 which are orthogonal to this axis do not generate bias in the association test, even if they bias the polygenic scores along other axes. Second, eq. 21 shows how the axis captured by $\tilde{F}_{Gr}$ in the GWAS panel is related to the underlying populations in our model, and to the pattern of population structure among them. The vector $T_{pop} = W_X^\top T$ has length K and captures the axis of the test in terms of the underlying populations, while $\mathbf{F}_{pop} T_{pop}$ similarly has length K, and captures the extent to which genetic drift along the path to each population is associated with the axis of population structure identified by the test vector. $\mathbf{F}_{pop} T_{pop}$ therefore describes the axis of confounding in terms of populations. Multiplying this vector by W_G then rotates this axis into the individual space of the GWAS panel to give $\tilde{F}_{Gr}$.

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 [36, 47, 40], 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 an option.

Alternately, our analysis above suggests that including $\tilde{F}_{Gr}$ as a covariate in the GWAS model will result in an unbiased test no matter what the pattern of confounding is. In this case, the OLS estimated marginal effect at site ℓ, conditional on G_ℓ, u and e, is

$$\hat{\beta}'_\ell \mid G_\ell, u, e = \beta_\ell + \frac{u_{-\ell}^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} + \frac{e^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} \tag{22}$$

where $\mathbf{P} = \left(\mathbf{I} - \frac{1}{\|\tilde{F}_{Gr}\|}\tilde{F}_{Gr}\tilde{F}_{Gr}^{\top}\right)$ prevents variation along the axis specified by $\tilde{F}_{Gr}$ from being used in the estimation of 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

$$\mathbb{E}\left[\hat{Z}\right]^{\top} \approx \mathbb{E}\left[Z\right] + \frac{S}{M}\left(\mu_S^{\top} + c^{\top}\right)\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}} \tag{23}$$

where

$$\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}} \approx \mathbf{P}\tilde{\mathbf{F}}_{GX} \tag{24}$$

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

$$\begin{aligned}\mathbb{E}\left[\hat{q} - q\right] &\approx \frac{S}{NM}\left(\mu_S^{\top} + c^{\top}\right)\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}}T \\ &\approx 0\end{aligned} \tag{25}$$

and the polygenic score association test is unbiased (see S4 and S5).

3.4.2 Relationship to PCA

A standard approach to controlling for population stratification in polygenic scores is to include the top J principle components of the GWAS panel genotype matrix as covariates in the GWAS, for some suitably large value of J [3]. How does our approach relate to the this standard approach?

In our model, $\mathbf{F}_{GG}$ contains the within panel relatedness for the individuals in the GWAS panel, where the entries themselves are a linear combination of the population level covariance matrix, $\mathbf{F}_{pop}$, the structure of which is determined by the demographic model. If we could take the eigendecomposition of $\mathbf{F}_{GG}$ directly, then the number of PCs that correspond to structure is entirely dependent on the population model. For example, in section 3.5.1 we simulate under a 4 population sequential split model (Figure 1), in which case there are three PCs that reflect real underlying structure. In section 3.5.2 we use a symmetric equilibrium migration model on a six-by-six lattice grid (Figure 3), in which case there are 35 PCs reflecting underlying population structure. Including these PCs is sufficient to remove all ancestry-associated bias in effect sizes estimates and render polygenic scores uncorrelated with any axis of ancestry variation under the null hypothesis.

To see how our approach relates to the PCA correction approach, we can write $\tilde{F}_{Gr}$ as a linear combination of GWAS panel PCs,

$$\tilde{F}_{Gr} = \sum_i \eta_i U_i \tag{26}$$

where U_i is the ith PC of $\mathbf{F}_{GG}$ and the weights are given by $\eta_i = Cov(U_i, \tilde{F}_{Gr})$. Estimating the marginal associations with $\tilde{F}_{Gr}$ as a covariate can therefore be understood as fitting a model in which *all* 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.

$$y = G_\ell \beta_\ell + \left(\sum_i \eta_i U_i \right) \omega + e. \tag{27}$$

As a corollary, if we perform a polygenic score association test using GWAS effect size estimates in which the top J PCs of $\mathbf{F}_{GG}$ 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 $\tilde{F}_{Gr}$ is captured by those J top PCs, i.e. that $\eta_i \approx 0$ for $i > J$.

In practice, however, we do not directly observe $\mathbf{F}_{GG}$ (and therefore cannot directly observe the PCs of the underlying data generating process, U). Rather, we observe a finite sample, and must estimate these quantities from the sample genotype data [3]. The accuracy of these "sample" PCs, $\hat{U}$, as compared to U, depend on the level of structure in the GWAS panel, as well as the number of individuals and SNPs. Specifically, as shown first by Patterson et al. (2006) in the context of genetics [48], PCA exhibits a phase change behavior in which a given principal component will only be detectable if the corresponding eigenvalue is greater than a threshold value that depends on the ratio of the number of SNPs to the number of individuals. Perhaps less well appreciated however is the fact that even when the corresponding eigenvalue exceeds this threshold, the angle between the sample PC and the true PC of the underlying population model may still be substantially less than one, particularly if the relevant eigenvalue does not far exceed the detection threshold [49, 50]. Thus, when PCA is applied to smaller GWAS samples, either of these two related phenomena may make it difficult to accurately approximate $\tilde{F}_{Gr}$ as a linear combination of the top sample PCs, leading to a failure to fully account for stratification bias in polygenic score association tests.

A second 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. $\mu_S + c$ in our notation) can be captured by the included PCs [51]. If this condition is met, effect size estimates are unbiased with respect to all axes of ancestry variation and therefore any association test that utilizes them will be unbiased as well. Combining this interpretation with results from above, PCA should successfully eliminate bias in polygenic score association tests if the J sample PCs included in the GWAS either capture the confounding effects on the phenotype, eliminating all effect size bias, or capture $\tilde{F}_{Gr}$, ensuring that effect size bias relevant to the test is removed.

3.4.3 Implementing the bias correction

Given these difficulties with the PCA approach, we propose an alternative, which is to obtain a direct estimate of $\tilde{F}_{Gr}$ using the relevant test panel genotype contrasts. Given the test panel genotype contrasts and GWAS panel genotypes from L sites genome-wide we obtain an estimate, $\hat{F}_{Gr}$, as

$$\hat{F}_{Gr} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{G_\ell r_\ell}{G_\ell^\top G_\ell / M}. \tag{28}$$

We then estimate marginal effects in the GWAS under the model

$$y = G_\ell \beta_\ell + \hat{F}_{Gr} \omega + \varepsilon, \tag{29}$$

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

3.4.4 Predicting GWAS panel genotypes

Above, we derived our proposed method in the context of a population genetic model. The form of eq. 28 suggests another useful interpretation, that of controlling for a prediction of the GWAS panel genotypes. To see this, suppose first that we standardize the GWAS panel genotypes by their variance so that $\tilde{G}_\ell = \frac{G_\ell}{\sqrt{G_\ell^\top G_\ell / M}}$ and we similarly standardize the test panel contrasts so that $\tilde{r}_\ell = H \frac{r_\ell}{\sqrt{G_\ell^\top G_\ell / M}}$, where H is a constant shared across loci, and chosen so that $\tilde{r}^\top \tilde{r} = L$. Now, consider the matrix regression model

$$\tilde{\mathbf{G}} = \rho \tilde{r}^\top + \varepsilon. \tag{30}$$

In this model, $\tilde{\mathbf{G}}$ is an $M \times L$ matrix containing the standardized genotypes ($\tilde{G}_\ell$) as columns, ρ 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 an $M \times L$ matrix of residuals. The OLS estimate of ρ is

$$\hat{\rho} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{G_\ell r_\ell}{G_\ell^\top G_\ell / M}, \tag{31}$$

identical to the expression for $\hat{F}_{Gr}$ in eq. 28 (i.e. $\hat{\rho} = \hat{F}_{Gr}$). Thus, for each site ℓ, a linear prediction of $\tilde{G}_\ell$ given $\tilde{r}_\ell$ is equal to the product of $\hat{\rho}$ (i.e. $\hat{F}_{Gr}$) and $\tilde{r}_\ell$:

$$\mathbb{E}\left[\tilde{G}_\ell \mid \tilde{r}_\ell\right] = \hat{\rho} \tilde{r}_\ell. \tag{32}$$

Therefore, we can intuitively interpret our bias correction method as including a prediction of the GWAS panel genotypes given the test panel genotype contrast as a covariate when estimating the effect at site ℓ.

This formulation also helps to highlight why our approach can succeed in settings where PCA fails (a point we demonstrate further in simulations below). Both our predictor ($\tilde{r}$) and our responses ($\tilde{G}_\ell$) are scaled to have a variance of one, so the entries of $\hat{F}_{Gr}$ vary around their true values with a standard deviation of approximately $L^{-1/2}$. The error in $\hat{F}_{Gr}$ therefore does not depend on the size of the GWAS panel, in contrast to PCA. Including $\hat{F}_{Gr}$ as a covariate when estimating effects should therefore provide protection against false associations between polygenic scores and ancestry so long as the entries of $\tilde{F}_{Gr}$ vary on a scale that is significantly greater than the scale of this estimation error, i.e. if $\sqrt{\frac{1}{M}\sum_m \tilde{F}_{Gr,m}^2} \gg L^{-1/2}$. Intuitively, it is easier to protect against stratification bias along a given axis of population structure if we specify that axis directly (even if we do so in a different panel) than if we use an unsupervised approach such as PCA to discover and estimate several axes of population structure and hope that the relevant axis is captured by these estimates.

3.4.5 Downward bias with true signal

Our approach proposes to use 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. Shouldn't this remove the signal we are trying to detect? While the answer is yes, at least 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.

To see why, we can rewrite the regression eq. 29 in terms of a sum of the contribution from our focal site and all other sites as

$$y = G_\ell \beta_\ell + \left(\frac{L-1}{L} \hat{F}_{Gr,-\ell} + \frac{1}{L} \tilde{G}_\ell \tilde{r}_\ell \right) \omega + e \tag{33}$$

where $\hat{F}_{Gr,-\ell}$ is the estimate of $\tilde{F}_{Gr}$ that one would obtain using all sites other than site ℓ. Thus, because our $\hat{F}_{Gr}$ includes a contribution from the focal site, controlling for it induces a slight bias in the estimated effect size, the sign of which depends on the sign of $\tilde{r}_\ell$. If $\tilde{r}_\ell$ is positive, $\hat{\beta}_\ell$ has a slight negative bias, whereas if $\tilde{r}_\ell$ is negative, the bias will be positive. Similar effects are noted elsewhere in the statistical genetics literature, for example in correcting for PCs of gene expression data [52], and linear mixed models [53], where it has been termed "proximal contamination". Assuming that the variance of $\tilde{r}_\ell$ across sites included in the score is similar to those used to compute the correction, 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.

$$\mathbb{E}\left[\hat{q} \mid q\right] \approx q \left(1 - \frac{S}{L}\right). \tag{34}$$

Notably, controlling for PCs of the GWAS panel genotype matrix will induce a similar effect if the axis of structure corresponding to the test is also present in the GWAS panel and gets included in the GWAS model as a covariate. In either case, the downward bias should be small so long as sites used to compute the polygenic score are only a small subset of those used to estimate $\hat{F}_{Gr}$. While this picture is somewhat complicated by the existence of linkage disequilibrium in real populations, 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 [54], suggesting that even for a polygenic score that included, for example, 10,000 SNPs, the downward bias should be no more than 2%. An important caveat is that some methods for computing polygenic scores allow for all or at least a substantial fraction of all SNPs genome wide to make non-zero contributions to the polygenic score. Further concern 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 [53, 5] or via iterative approaches that first aim to ascertain SNPs using a genome-wide estimate of $\hat{F}_{Gr}$ before re-estimating effects using an estimate of $\hat{F}_{Gr}$ computed from sites not in strong LD with any of the ascertained sites. Understanding the behavior of tests for polygenic score-ancestry associations in the context of these methods will require careful attention to the methods' assumptions.

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 (labeled A, B, C and D), which are related to one another by an evenly balanced population phylogeny (Figure 1). 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 consider 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 eq. 15, the marginal effect size estimate for site ℓ is

$$\begin{aligned} \hat{\beta}_\ell \mid G_\ell, e &= \frac{G_\ell^\top e}{G_\ell^\top G_\ell} \\ &= \frac{1}{2}\frac{\Delta_{AB}\left(\hat{p}_{A,\ell} - \hat{p}_{B,\ell}\right)}{G_\ell^\top G_\ell / M} + \frac{G_\ell^\top \varepsilon}{G_\ell^\top G_\ell} \end{aligned} \tag{35}$$

where $\hat{p}_{A,\ell}$ and $\hat{p}_{B,\ell}$ are the observed sample allele frequencies for population A and B at site ℓ (see also equation 2.3 in the supplement of [27]).

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

$$\begin{aligned} \mathbb{E}\left[\hat{q} - q\right] &= \Delta_{AB} \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{(\hat{p}_{A,\ell} - \hat{p}_{B,\ell})(\hat{p}_{C,\ell} - \hat{p}_{D,\ell})}{G_\ell^\top G_\ell / M}\right] \\ &= \Delta_{AB} S \tilde{F}_4(A, B; C, D) \end{aligned} \tag{36}$$

where $\tilde{F}_4(A, B; C, D)$ is a version of Patterson's F_4 statistic [55, 56], standardized by the genotypic variance in the GWAS panel, which measures the amount of genetic drift common to population 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 Figure 1A, where shared drift on the internal branch generates such a correlation, yielding a positive value for $\tilde{F}_4(A, B; C, D)$, but not for the model in Figure 1B, where there is no shared internal branch, $\tilde{F}_4(A, B; C, D) = 0$.

To test this prediction, we simulated 100 replicates of four populations related by this topology where the F_{ST} between pairs of populations on the same side of the tree is approximately 0.005 and approximately 0.01 for populations on opposite sides. 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, ascertained SNPs, and then used these SNPs to construct polygenic scores and test for a difference in mean polygenic scores 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).

Given the population model, $\tilde{\mathbf{F}}_{XG} = \mathbf{0}$ for the unconfounded topology, making $\tilde{F}_{Gr}$ is a vector of zeros. For the confounded topology, the structure in $\tilde{\mathbf{F}}_{XG}$ 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. In our simulations we estimate $\hat{F}_{Gr}$, as in eq. 28, using all genome-wide SNPs with a frequency of greater than 5% in both the test and GWAS panels. The resulting vector is, as expected, tightly correlated with population label for the 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 in the GWAS, 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 wanted 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 confounded topology 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 to generate a correlation between the effect and the frequency differences between populations C and D. This procedure generates an A_{pop} with non-zero and opposite sign entries for populations C and D and is conceptually equivalent adding a selection event on the internal branch in the population phylogeny. We then 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 real samples that 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 effects 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 a genome-wide estimate of $\hat{F}_{Gr}$ will have a minimal effect on power to detect real signals as long as $S << L$.

3.5.2 Grid Simulations

To further explore our approach 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, building off of a framework developed by Zaidi and Mathieson (2020) [57]. We sampled an equal number

of individuals per deme to comprise both the GWAS and test panels and then simulate several different distributions 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 along the diagonal of the grid, and one in which only a single deme has an elevated mean, a scenario which previous work has shown to be difficult to correct for with standard tools [58, 57]. 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, the confounder, c, is a linear function of an individual's position on the latitudinal axis (Figure 3A). When we estimate effect sizes with no correction for population structure, the spatial distribution of the resulting polygenic scores 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 correlated with both latitude and longitude in the test panel and an association test that is biased along both axes (Figure 3B). When we compute $\hat{F}_{Gr}$ using latitude as the test vector, the resulting effect sizes are uncorrelated with latitudinal genotype contrasts, but remain susceptible to bias along other axes (e.g. longitude). This example highlights the targeted nature of this method, as using effect sizes from a GWAS including $\hat{F}_{Gr}$ does not remove all bias, but does make the association test using those effect sizes for the pre-specified test vector unbiased.

In the first two examples, the standard approach of including the top $J = 10$ principal components of the GWAS panel genotype matrix as covariates in the GWAS successfully control for confounding along the latitudinal axis, albeit for potentially different reasons. As outlined in Section 3.4.2, PCA is successful when either the confounder or $\tilde{F}_{Gr}$ falls in the span of the sample PCs included in the GWAS. In Figure 4 we compute the average correlation between all $M - 1$ PCs and either the confounder or our estimate $\hat{F}_{Gr}$. For the first example, both the confounder and $\hat{F}_{Gr}$ reflect broad geographic structure and are easily captured by the first two PCs. In the second example, however, the confounder lies along a more complicated geographic axis, which is not fully captured by the top 10 PCs (Figure 4B, left column). As a result, the polygenic scores retain some amount of confounding along the diagonal (Figure 3B, fourth column). However, because of the simple relationship between geography and migration patterns in our simulations and the overlapping structure between the two panels, the axis of structure relevant for the latitudinal association test in the test panel is the latitudinal axis in the GWAS panel (Figure S5). This pattern *is* well described by the top 10 PCs (Figure 4B, right column), so PCA protects the test in this scenario despite the residual confounding. The same is true for the test of association with longitude.

In the third example, we simulate an increased environmental effect in a single deme, which induces a more complex spatial pattern in the uncorrected polygenic scores (Figure 3C). We then take 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 compute $\hat{F}_{Gr}$ using these contrasts. In this case, the spatial distribution of $\hat{F}_{Gr}$ in the GWAS panel is more complicated, reflecting the genetic distance of each deme from the focal deme of the test, given the equilibrium migration model (Figure S5). Including only $\hat{F}_{Gr}$ in the GWAS protects the test from confounding, but in doing so induces a

residual pattern in the polygenic scores that is largely a mirror image of the original, uncorrected pattern of confounding. In contrast, including the top 10 PCs reduces much of the variation in the polygenic scores across the landscape, but is insufficient to protect the association test. Combining both approaches protects the test without inducing a residual pattern in the polygenic scores (Figure S6). We revisit this observation in the Section 4.

Here, PCA alone fails because neither the confounder nor $\tilde{F}_{Gr}$ are fully captured by the included PCs (Figure 4C). In fact, for this example we had to include 1000 PCs to reduce false positive rates to nominal level (Figure S3), and doing so regressed out 75% of all available genetic variation. This may initially be surprising: as there are 36 demes, we should need only 35 linear axes to explain any given pattern of variation across these demes. At least in principle, we could compute the expected pairwise coalescent times for individuals sampled from different demes, and use these to compute the expected genetic relationship matrix given the demographic model. Then, taking the eigendecomposition of this matrix would yield 35 axes with non-zero eigenvalues [59], and it would be possible to capture any confounder/protect any test by controlling for these 35 PCs. However, as we discussed above, PCs computed from the sample genotype data can be poor estimates of the PCs of the underlying population process, and below a certain threshold, may not be detectable at all [48, 49]. In this setting, adding additional sample PCs merely amounts to controlling for additional random orthogonal vectors. While a sufficiently large number of random vectors is ultimately able to capture the confounder/test well enough to control the false positive rate, this comes at the cost of also regressing out most of the variation we would like to use to estimate effect sizes, and so is not a viable option.

Finally, we 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 nearly 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

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 [32, 33, 40, 61]. 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. By using the test panel and GWAS panel genotypes jointly, this approach removes guesswork in choosing the number of top PCs to include in the GWAS and can succeed in situations where PCA fails.

Another advantage of our approach is related to the tradeoff between statistical power and robustness in polygenic selection analyses. Researchers motivated by the difficulties with stratification bias and height [32, 33] have taken 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, 35, 62] or using effect sizes from smaller family-based GWASs

[35, 63], but also choices such as focusing only on genome wide significant hits, or throwing away all effect size information except for the sign [35, 18]. Our approach provides a stronger guarantee against bias due to ancestry stratification which will allow researchers to use less conservative methods for computing polygenic scores that do not trade away so much statistical power.

Our results are also relevant for analyses that use GWAS summary statistics and coalescent methods to test for signals of directional polygenic selection [19, 23, 24, 64]. One way of thinking about these coalescent approaches is to recognize 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 immunize 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 [24, 64], 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 frame our model in terms of detecting an association between polygenic scores and an ancestry gradient of interest due selection acting on genotypes in the test panel. We focus on this scenario because there is a well-defined null hypothesis and residual stratification is known problem. However, none of our results about the source of bias rely on the signal being generated by selection. Other non-neutral process, such as assortative mating along ancestry gradients, phenotype dependent migration, or phenotype dependent ascertainment could generate signals similar to those generated by natural selection. In other settings, researchers may simply be interested in accurately estimating the association between a polygenic score and a given ancestry gradient, whether or not such an association could plausibly have been generated by neutral evolutionary processes or not. Our results apply in this scenario as well.

One disadvantage of our approach is that it requires that the test panel genotypes are available and the test axis specified before carrying out the GWAS. Thus, the benefits 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 dataset [65, 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 spurious inflation of the slope in this context, possibly due to background selection [32]. A more complete analysis of bivariate LD score regression applied in this context would be valuable.

We also note several additional 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 [66] and compute different corrections for different frequency/age bins similar to the suggestion by [57] 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 [67]. 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 [68]. Finally, the interpretation of associations between polygenic scores and axes of ancestry variation remains challenging in general, as differences in polygenic scores do not necessarily translate into differences in phenotype [13], and associations between polygenic scores and patterns of ancestry variation may have counterintuitive evolutionary explanations [69, 39, 38]. 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* [70] 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 within panel F_{ST} [**?**] was 0.01, whereas in the configuration in Figure 1C the average F_{ST} was 0.005.

The second model, modified from [57], 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 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 generated three different phenotypic gradients where the largest phenotypic shift was always equal to Δ. To generate a latitudinal gradient (Figure 3A) we added $\frac{\Delta}{5}$ to y_i for individuals in row 1, $2\frac{\Delta}{5}$ for individuals in row 2, etc. For Figure 3B we generated a gradient along the diagonal by adding $\frac{\Delta}{5}$ to the phenotype for individuals in deme (1,1), $2\frac{\Delta}{5}$ for individuals in deme (2,2), etc. For Figure 3C we shifted the phenotype of individuals in deme (1,4) by Δ. For all grid simulations we varied Δ from 0 to 2 standard deviations (Figure S4) .

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 [57] to simulate a neutral trait with $h^2 = 0.3$. We first randomly select 1 causal variant per chromosome and sample it's effect sizes from $\beta_\ell \sim N(0, \sigma_i^2[p_\ell(1-p_\ell)]^\alpha)$, where σ_i^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 $\sigma_g^2 = \sigma_i^2 \sum_{\ell=1}^{200} [2p_\ell(1-p_\ell)]^{\alpha+1} = 0.3$.

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 use θ ranging from $0.5 - 0.62$. To calculate $\overline{\Delta}_{CD}$, the true difference in polygenic score between test panel populations, we take 100 simulations with the same θ and average the difference $Z_D - Z_C$. Z_D and Z_C are the population average polygenic scores computed using the true causal effects in D and C, respectively. 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 Δ_{AB} to $e_{i,k}$ for individuals in population B.

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

We computed a single covariate $\hat{F}_{Gr}$ for each polygenic association test. 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 $r = \mathbf{X}^\top T$ using the plink2 [71] function `--glm`. Finally we compute $\hat{F}_{Gr}$ (see eq. 28) using `--sscore` in plink2, taking care to standardize by the variance in the GWAS panel genotypes.

5.1.3 GWAS and Polygenic Scores

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. $y = \beta_\ell G_\ell + \omega \hat{F}_{Gr} + \epsilon$
3. $y = \beta_\ell G_\ell + \omega \mathrm{ID} + \epsilon$
4. $y = \beta_\ell G_\ell + \omega_1 \mathrm{PC}_1 + ... + \omega_k \mathrm{PC}_k + \epsilon$

where ID is an indicator of population membership for the discrete 4 population model. This set-up allowed us to compare our covariate $\hat{F}_{Gr}$ to a GWAS done with no correction and one done with perfect correction for the discrete model. 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 [71] 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, including all 200 SNPs along with the covariates.

Finally, we construct polygenic scores for the individuals in the test panel as $\hat{Z}_i = \sum_{\ell=1}^{200} \hat{\beta}_\ell X_\ell$ where $\hat{\beta}_\ell$ is the estimated effect size from the joint model and X_ℓ is the mean centered genotype value for the ℓth variant.

5.1.4 Polygenic Score Association Test

For each replicate we first compute the test statistic $\hat{q} = \frac{1}{N}\hat{Z}^\top T$ by multiplying the vector of polygenic scores for individuals in the test panel by the test vector.

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,

$$P = 2 * \frac{\sum_i^N I(|\hat{q}_i| > |\hat{q}_{obs}|)}{K}, \tag{37}$$

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, Jeff Spence, Arjun Biddanda and Yuval Simons 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. 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.

13. 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.

14. 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.

15. 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.

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

17. 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.

18. 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.

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. https://doi.org/10.1016/j.ajhg.2020.12.005.

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. 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`.

36. 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.

37. 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):3865. https://doi.org/10.1038/s41467-020-17719-y.

38. 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.

39. 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.

40. Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature. 2023; p. 1–8. https://doi.org/10.1038/s41586-023-06079-4.

41. 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. https://doi.org/10.1038/s41588-019-0379-x.

42. Cavalli-Sforza LL, Barrai I, Edwards AWF. Analysis of Human Evolution Under Random Genetic Drift. Cold Spring Harbor Symposia on Quantitative Biology. 1964;29:9–20. https://doi.org/10.1101/SQB.1964.029.01.006.

43. Nicholson G, Smith AV, Jónsson F, Gústafsson Stefánsson K, Donnelly P. Assessing Population Differentiation and Isolation from Single-Nucleotide Polymorphism Data. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2002;64(4):695–715. https://doi.org/10.1111/1467-9868.00357.

44. Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics. 2010;185(4):1411–1423. https://doi.org/10.1534/genetics.110.114819.

45. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genetics. 2012;8(11):e1002967. https://doi.org/10.1371/journal.pgen.1002967.

46. 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.

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. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLOS Genetics. 2006;2(12):e190. https://doi.org/10.1371/journal.pgen.0020190.

49. Johnstone IM, Paul D. PCA in high dimensions: An orientation. Proceedings of the IEEE. 2018;106(8):1277–1292.

50. Bloemendal A, Chen C. PCA and stratification in GWAS / A primer on random matrix theory;. `https://www.youtube.com/watch?v=B7ub92OLw1g`.

51. Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. Nature Reviews Genetics. 2013;14(1):1–2. https://doi.org/10.1038/nrg3382.

52. 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.

53. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. Nature Methods. 2012;9(6):525–526. https://doi.org/10.1038/nmeth.2037.

54. 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.

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

56. 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.

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

58. 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.

59. McVean G. A Genealogical Interpretation of Principal Components Analysis. PLOS Genetics. 2009;5(10):e1000686. https://doi.org/10.1371/journal.pgen.1000686.

60. Xu S, Chen M, Feng T, Zhan L, Zhou L, Yu G. Use ggbreak to effectively utilize plotting space to deal with large datasets and outliers. Frontiers in Genetics. 2021;12:2122.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology. 2016;12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

71. 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.

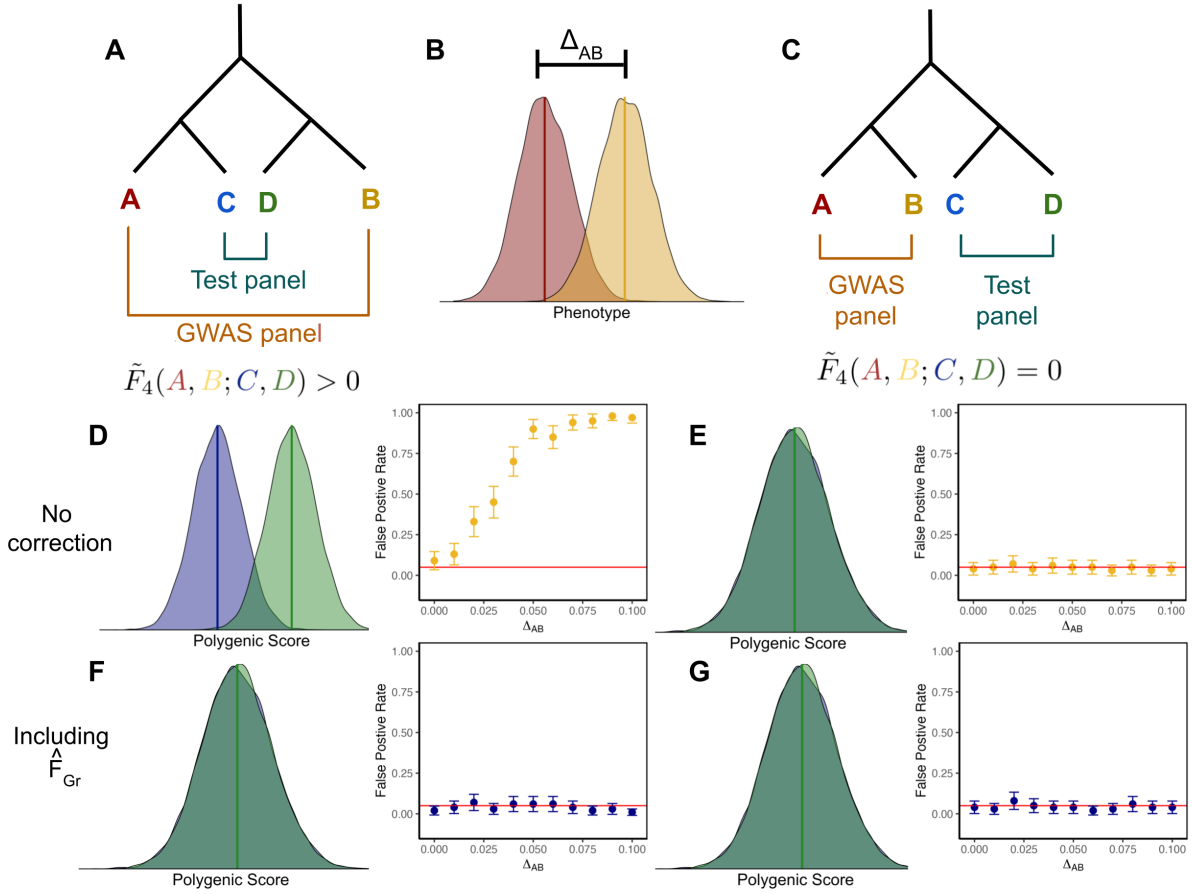

Figure 1: **Schematic of two different panel configurations. The effect of stratification depends on the overlapping structure between the GWAS and test panels.** (A, C) Two different topologies used to create the GWAS and test panels. (B) Stratification was modeled in the GWAS panel by drawing an individual's 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 no expected difference in mean polygenic scores between C and D and the false positive rate remains near 5% regardless of the magnitude of stratification. (F, G) Including $\hat{F}_{Gr}$ as a covariate in the GWAS controls for stratification, keeping the false positive rate 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}.

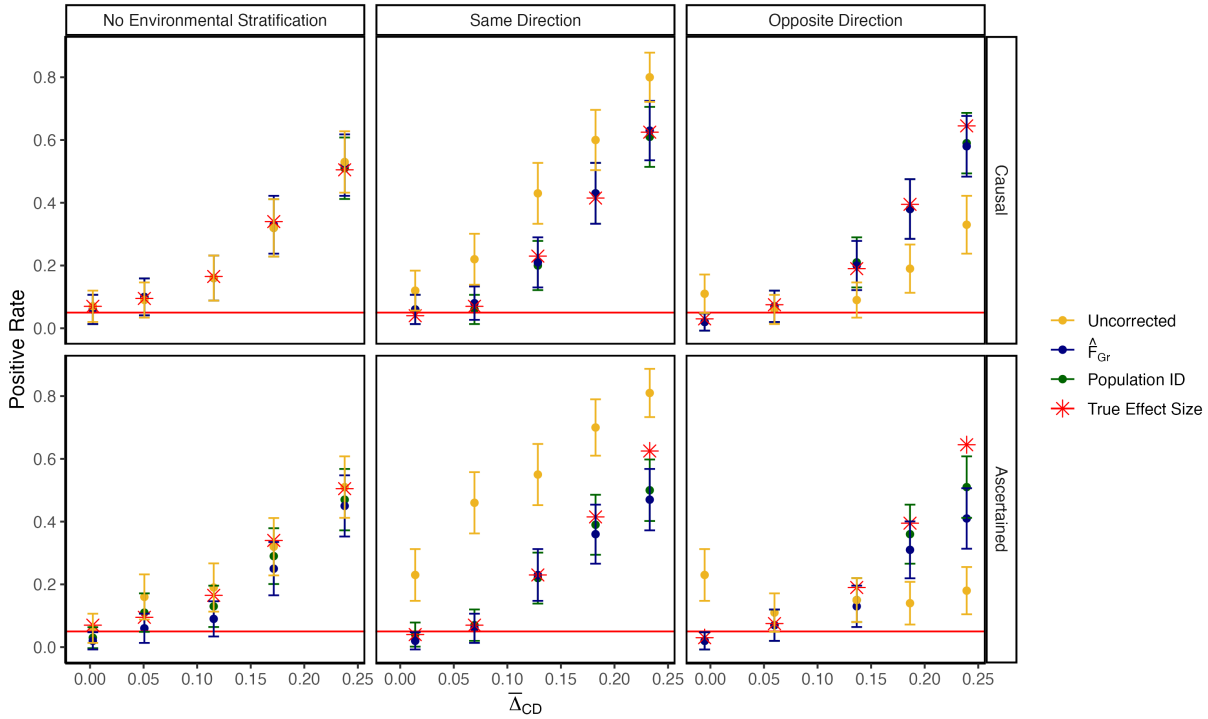

Figure 2: **Including $\hat{F}_{Gr}$ as a covariate in the GWAS model does not regress out true signal.** Here we generate a true difference in average polygenic scores by flipping the sign of a proportion of causal effect sizes to align with allele frequency contrasts, $p_{D,\ell} - p_{C,\ell}$, in 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.

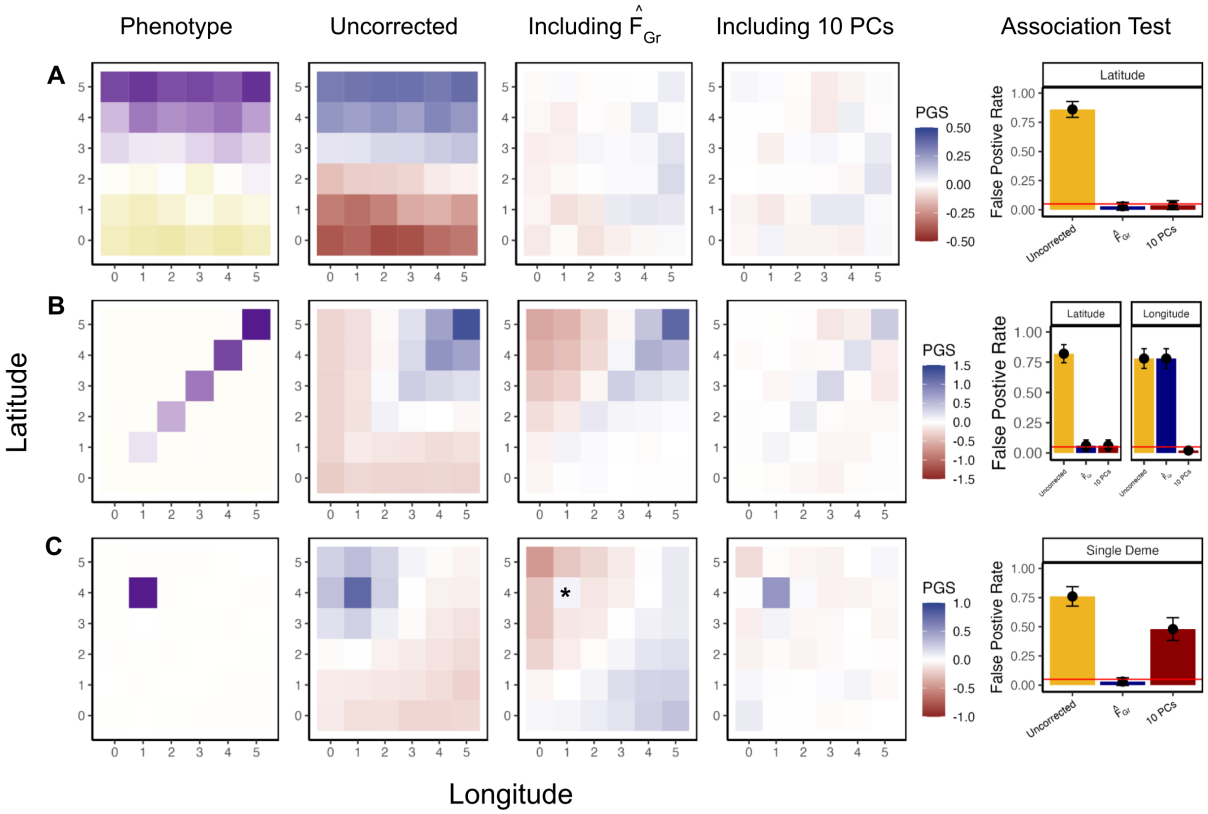

Figure 3: **Including $\hat{F}_{Gr}$ as a covariate in the GWAS model controls for stratification bias specific to the association test of interest.** GWAS and test panel individuals were simulated using a stepping-stone model with continuous migration. In the GWAS panel, the phenotype is non-heritable and stratified along either latitude (A), the diagonal (B), or in a single deme (C). 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 $\hat{F}_{Gr}$ (test vector is latitude for A and B, belonging to * deme for C) 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 (fifth column). We also compare our approach to including the top 10 PCs (fourth column). For (A) the magnitude of confounding Δ is 0.2 standard deviations while $\Delta = 1$ for (B) and (C) (see Methods).

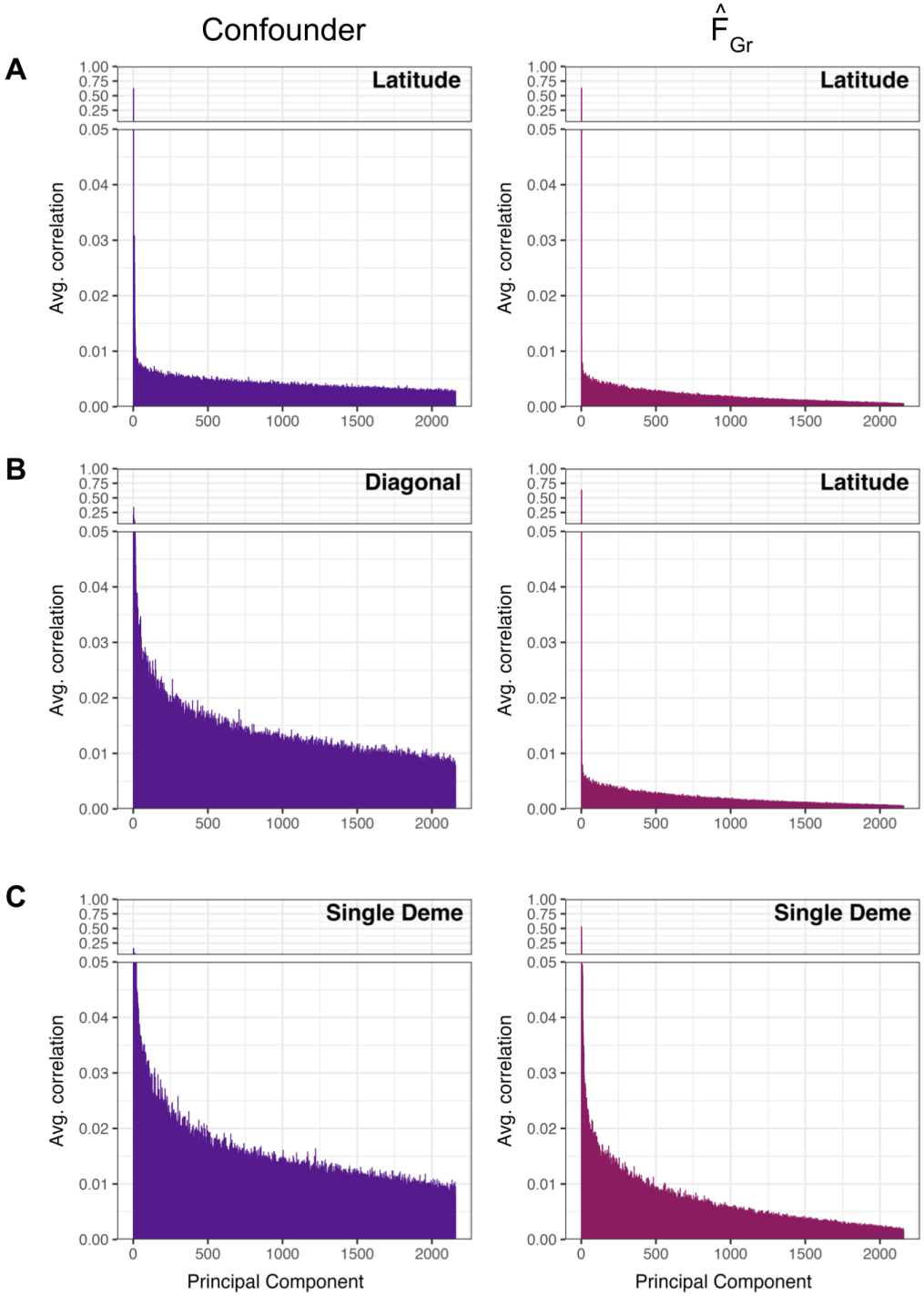

Figure 4: **Different patterns of confounding and $\hat{F}_{Gr}$ load onto different GWAS panel principal components.** For the three possible combinations of confounding and polygenic score association tests in Figure 3, we plot the average correlation across replicates between the confounder, $\hat{F}_{Gr}$, and the GWAS panel principal components. To better visualize differences, we cut the y-axis at 0.05 using [60] and enlarge the bottom portion of the plot. In (A) both the confounder and $\hat{F}_{Gr}$ represent variation along latitude and are well captured by the first two PCs. For (B) the confounder varies along the diagonal and these individual deme level differences are correlated with higher PCs. In contrast, the test vector is still latitude and $\hat{F}_{Gr}$ is again well captured by PCs 1 and 2. Finally, in (C), both the confounder and the test vector represent membership in a single deme and therefore the confounder and $\hat{F}_{Gr}$ are correlated with higher PCs that either capture finer structure in the sample or represent random orthogonal vectors.

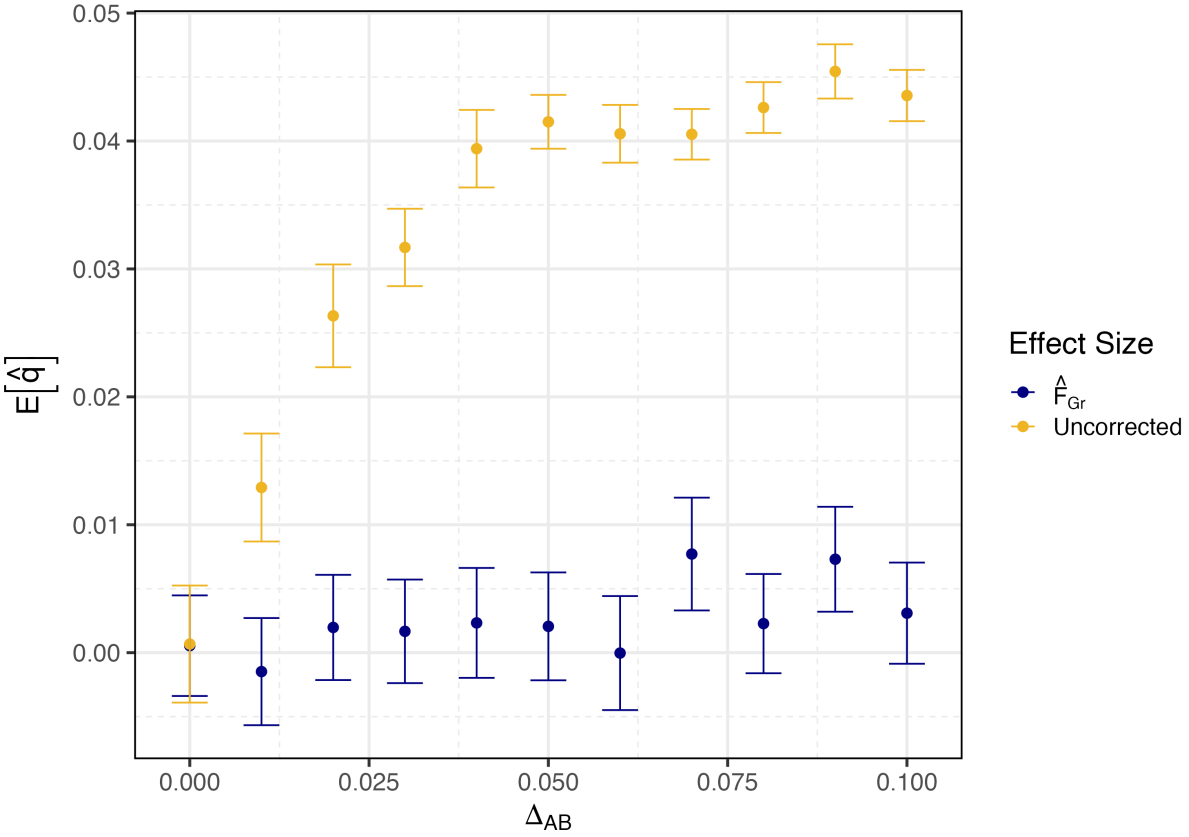

Figure S1: **Including $\hat{F}_{Gr}$ as a covariate removes bias in $\hat{q}$.** In our simulations of the 4 population toy model with overlapping topology (Figure 1A), increasing the magnitude of the confounder increases the average value of $\hat{q}$ when stratification is not controlled for. Including $\hat{F}_{Gr}$ as a covariate in the GWAS model reduces bias in the test statistic.

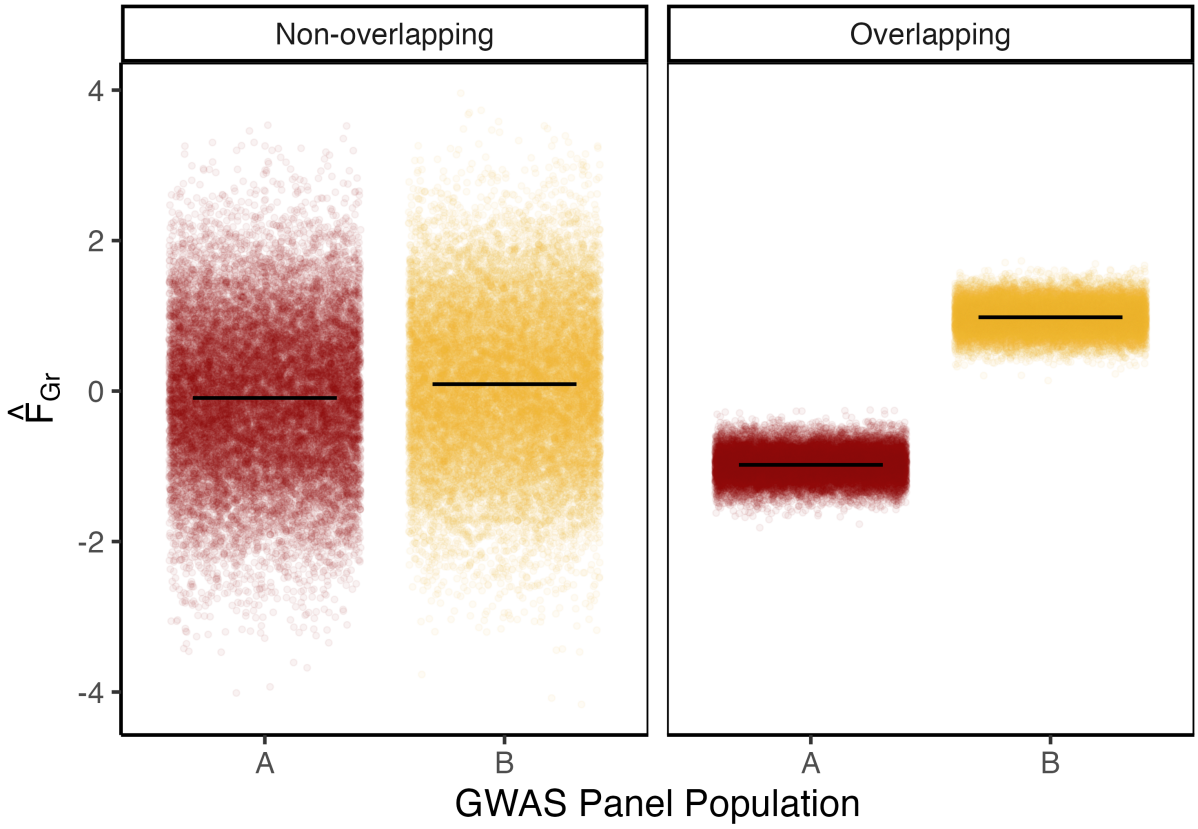

Figure S2: $\hat{F}_{Gr}$ **separates populations in the GWAS panel when test and GWAS panels have overlapping structure.** In the 4 population toy model, when the topology is non-overlapping (Figure 1C), $\hat{F}_{Gr}$, shown here for a single replicate, is purely noise and does not reflect population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

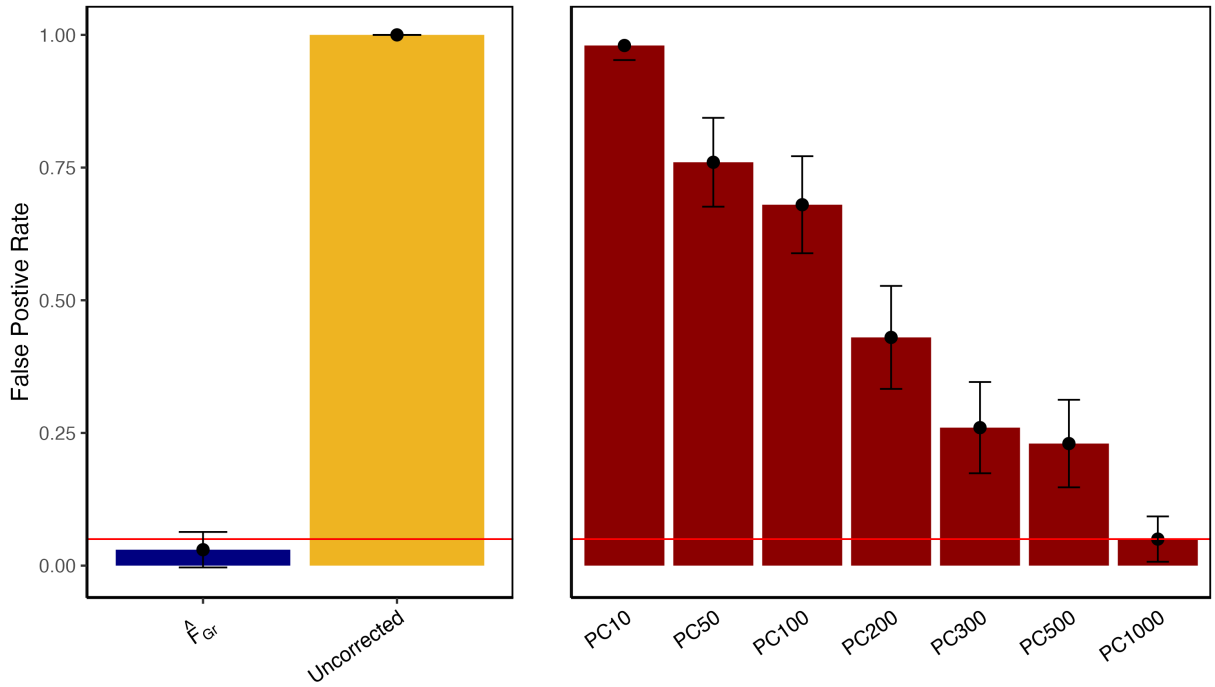

Figure S3: $\hat{F}_{Gr}$ **controls for stratification for a single deme association test.** Here we simulate a 2 standard deviation confounder for a single deme in the GWAS panel (Figure 3C) and conduct an association test for the difference in polygenic score between the same single deme and the rest of the grid in the test panel. Including $\hat{F}_{Gr}$ as a covariate in the GWAS model controls for stratification and the association test using those effect sizes is unbiased (left panel). In the right panel we vary the number of principal components of the GWAS panel included as covariates. It takes 1,000 PCs to return the false positive rate to nominal levels.

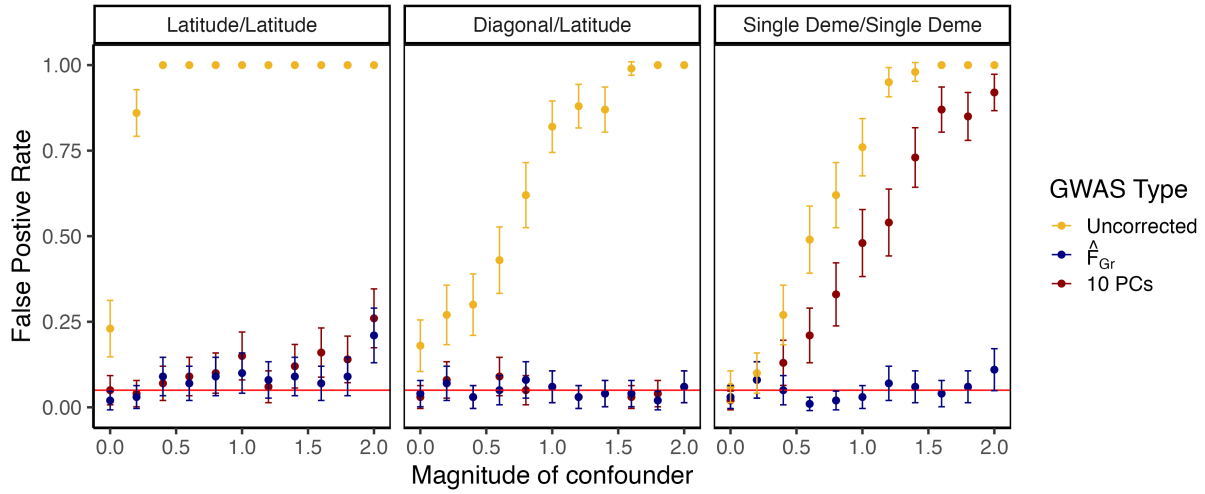

Figure S4: $\hat{F}_{Gr}$ **controls for stratification for a range confounding effects.** For the three simulation scenarios in Figure 3 we plot the false positive rate for the given association test for a range of confounding effects where the magnitude of the confounder is the difference between the most extreme phenotypic shifts in units of standard deviations (see Methods). For all but the strongest magnitude of confounding in the Latitude/Latitude situation, our approach protects the test from stratification bias. In two confounder/test scenarios (left and right columns) the standard approach of including the top 10 GWAS panel principal components begins to fail with large enough confounding.

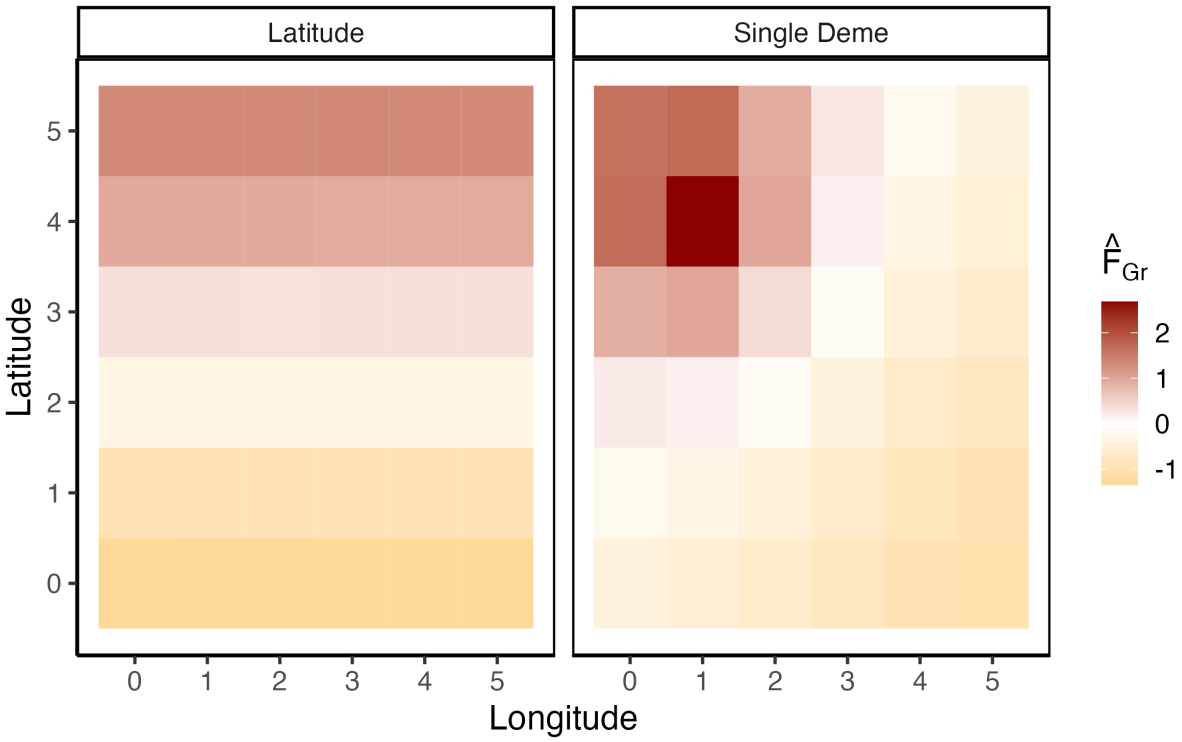

Figure S5: $\hat{F}_{Gr}$ **predicts** T **if measured in the GWAS panel.** For both of the test vectors in Figure 3 we plotted the average $\hat{F}_{Gr}$ per deme across 100 replicates. Because the test and GWAS panel are sampled from the same demographic model and have completely overlapping structure, $\hat{F}_{Gr}$ reflects the spatial pattern of T.

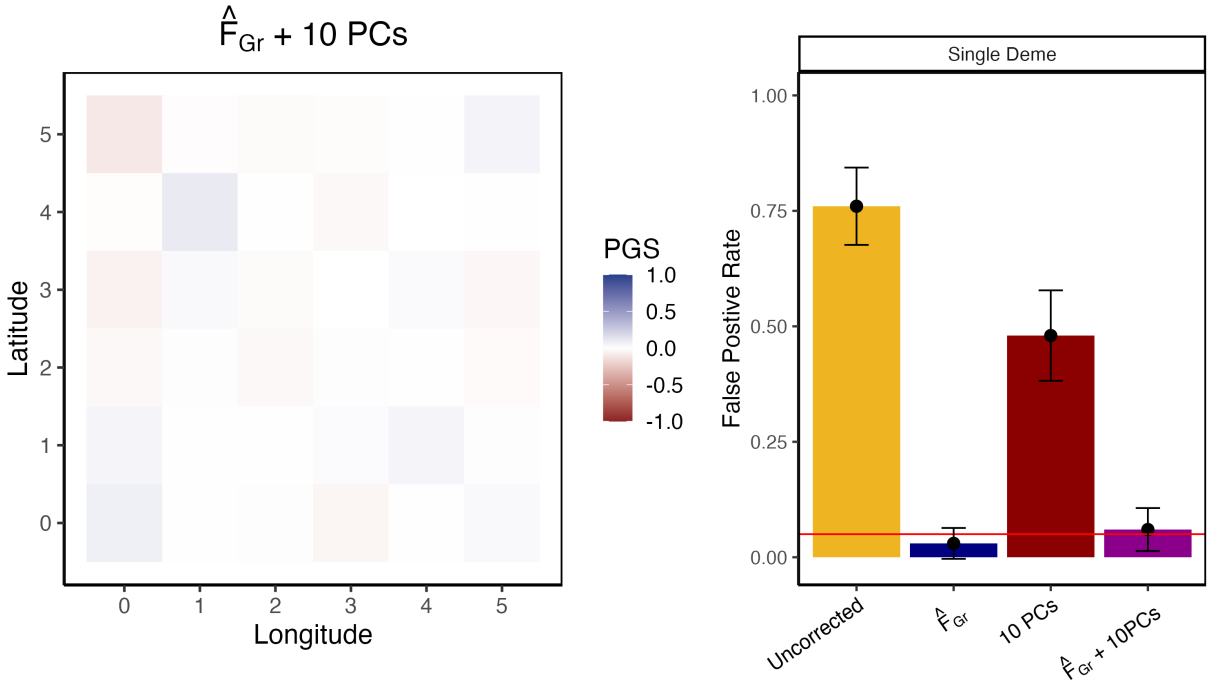

Figure S6: **Including $\hat{F}_{Gr}$ and 10 PCs protects the test and removes spatial patterns in the PGSs.** For the single deme confounder and test in Figure 3C we included both $\hat{F}_{Gr}$ and the top 10 PCs as covariates in the GWAS model. When only $\hat{F}_{Gr}$ is included the test is protected, but a residual pattern in the polygenic scores induced (Figure 3C, third column). Including both $\hat{F}_{Gr}$ and the top 10 PCs continues to protect the test (right) and removes the residual pattern (left).