THE UNIVERSITY OF CHICAGO

TESTING FOR DIFFERENCES IN POLYGENIC SCORES IN THE PRESENCE OF CONFOUNDING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY
JENNIFER GRACE BLANC

CHICAGO, ILLINOIS
DECEMBER 2024

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am very grateful to my advisor, Dr. Jeremy Berg, who has been an amazing mentor to me throughout my graduate career. All of the following work would not have been possible without his support and mentorship. Being a professor's first graduate student is a unique experience, and I am thankful Jeremy took a chance on me and supported me through both the highs and lows of this project. Jeremy is a careful and innovative scientist while being an exceedingly kind mentor who always puts his students first, and I am grateful to have trained in his lab. I would also like to thank all past and present members of the Berg lab, especially Vivaswat Shastry, who has supported me scientifically and as a friend.

I would also like to thank my committee members, Dr. Matthew Stephens, Dr. John Novembre, and Dr. Xuanyao Liu, for the invaluable feedback and support I have received. Additionally, I'd like to thank Dr. Peter Carbonetto, Dr. Andy Dahl, Dr. Matthias Steinruecken, and Dr. Maryn Carlson for their help with various stages of this project. Thank you to the entire human genetics community at The University of Chicago for fostering a collaborative and rigorous environment where I have learned to become a better, more thoughtful scientist. Also, a special thank you to Sue Levison for all her help throughout the years; I don't know what we'd do without you.

Thank you to all my former mentors, including Dr. Michael Turelli, Dr. Graham Coop, Dr. Eimear Kenny, Dr. Emily Josephs, and Dr. Gillian Belbin, without whom I never would have considered a career in genetics.

I would like to thank all of my friends and fellow graduate students who I have met here, including Dr. Kate Farris, Dr. Selene Clay, Dr. Katie Aracena, Jaeda Patton, Santiago Herrera, Astra Huang, Abhimanyu Lele, Renée Fonseca, and many others. I especially want to thank Jojo Tang, my roommate of six years, who has been there for me through my best and worst moments, Maggie Steiner, who has brought so much joy and friendship to my life, and Shreya Ramachandran, my partner in crime, who helped make Chicago feel like home.

I also want to thank all my friends from Davis and my volleyball crew for their support (and their willingness to listen to me complain) over the years.

Thank you to my boyfriend, Christopher; your support means the world to me. You believed in me at times when I didn't believe in myself and have been there both as a cheerleader and a shoulder to cry on every time I needed it. Your kindness and optimism about the future are inspiring, and I can't wait to do it all with you.

Finally, thank you to my family - my brother, Kevin, my parents, Anna and Jim, and my grandparents, Marcia, Stephen, Marilu, and Gene. Without you, none of this would be possible. Mom, it's no mystery where my love of science came from. My whole life I've seen you fight tirelessly for your students, for science education, and for social justice. You inspire me every day, and I can never thank you enough for your endless support. Dad, you have taught me to stand up for myself, treat others with kindness and understanding, and to "not sweat the small stuff." I want to thank you for all the sacrifices you have made so that Kevin and I can achieve our goals. I love you all; your support has gotten me to where I am today, and I am so excited for the next chapter.

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 and empirical analysis, 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. Specifically, for any given test, there exists a single axis of population structure in the GWAS panel that needs to be controlled in order to protect the test. Based on this result, we propose a new approach for directly estimating this axis of population structure in the GWAS panel. We then use simulations to compare the performance of this approach to the standard approach in which the principal components of the GWAS panel genotypes are used to control for stratification. Finally, we develop a hybrid approach for empirical data analysis that uses the test panel genotypes to estimate how well protected any given test is by the inclusion of principal components and apply this approach across a diverse set of tests.

CHAPTER 1

INTRODUCTION

A fundamental goal of human genetics is to understand the genetic basis of phenotypic variation. The explosion of sequencing data and the development of genome-wide association studies (GWAS) throughout the 21st century have accelerated our knowledge of the genetic architecture underlying important anthropometric, physiological, behavioral, and disease traits. By measuring the association between a genetic variant and phenotype, GWAS results provide an estimate of a variant's effect on the phenotype, averaged over the environments experienced by the individuals in that sample. Human GWAS results have shown that the majority of traits of interest are highly polygenic, with each individual variant estimated to have a small effect size[1]. The ability to estimate effect sizes has opened up the possibility of phenotypic prediction. For example, these effect estimates can be combined into polygenic scores in a separate prediction panel by taking a sum of the genotypes of individuals in that panel, weighted by the estimated effects[2]. Under the relatively strict assumptions that genetic and environmental effects combine additively, that variation in the phenotype is not correlated with variation in ancestry within the GWAS panel, and that the 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 variants. 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. The promise of polygenic scores is therefore that they provide a means for phenotypic prediction from genotype data alone.

The promise of phenotypic prediction and the ability to link complex trait variation to underlying genetic variants has transformed many areas of human genetics where the calculation of polygenic scores has already become a routine procedure. In medicine, polygenic scores have shown potential for predicting some diseases with accuracy comparable to

monogenic mutations and traditional risk factors[3]. In evolutionary biology, polygenic scores allow for the investigation into the evolution of the genetic basis of complex traits[4]. The application of polygenic scores to the growing collection of ancient DNA data, in particular, offers an exciting opportunity to better understand the evolutionary forces that have shaped current variation[5,6,7,8]. Finally, in social science, polygenic scores are used as a tool to study the relationship between genetic variation and social outcomes[9].

Although polygenic scores have become ubiquitous across human genetics literature, numerous implementation problems still exist. Here we focus on a singular known problem: the susceptibility of polygenic scores to ancestry-based confounding, also known as population stratification. As stated above, the promise of polygenic scores is that, under specific assumptions, they allow for phenotypic prediction from genotype data alone. However, the effects of individual variants are typically estimated from population samples in which the environments that individuals experience vary as a function of their social, cultural, economic, and political contexts. Differences in these factors are often correlated with differences in ancestry within population samples, and these ancestry-environment correlations can induce systematic biases in the estimated effects of individual variants (Figure 1.1A). Similar biases can also arise if genetic effects on the phenotype vary as a function of ancestry within the GWAS sample (Figure 1.1B). Ancestry stratification is a long recognized problem in the GWAS study design, and many steps have been taken to guard against its effects.

Next, I provide an overview of some of the most common approaches to control for stratification in GWAS and highlight both their successes and where they have fallen short, providing both theoretical and empirical examples. I divide existing mitigation approaches into two major categories: statistical correction and bias avoidance. Starting with statistical correction, the three most common approaches that I cover here are principal component analysis (PCA), linear mixed models (LMMs) and LD score regression. PCA is one of the earliest approaches and corrects for ancestry associated biases through inclusion of top

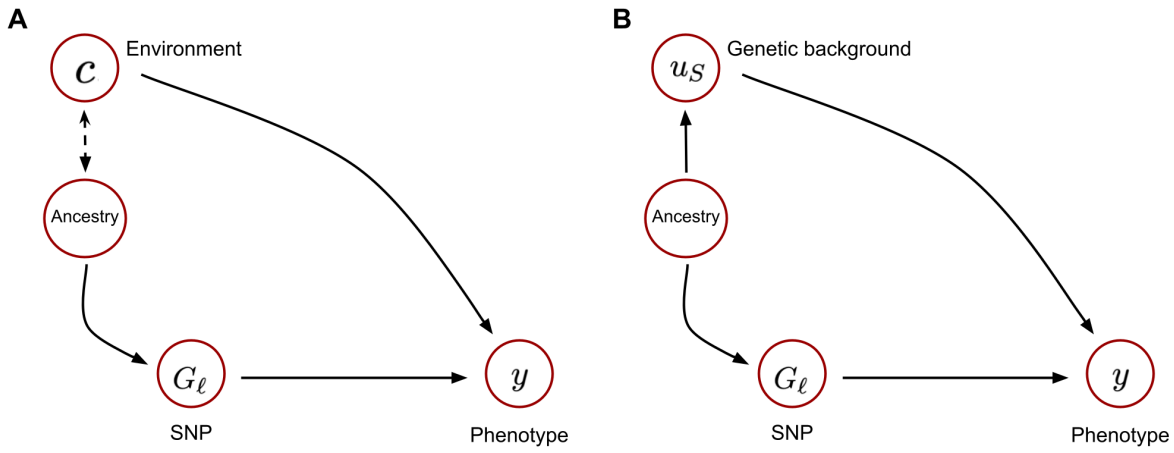

Figure 1.1: **Causal models for confounding in effect size estimates in GWAS** In GWAS, we aim to measure the association between a single SNP, G_ℓ, and phenotype, y, in a population sample. However, ancestry variation across samples can lead to differences in the allele frequency of the SNP. If this ancestry variation is correlated with environmental effects on the phenotype (A) or the causal genetic background (B), then the association between the SNP and phenotype is confounded, and the resulting effect size estimate will be biased (i.e population stratification in the GWAS literature). We note that A and B are not mutually exclusive, and both types of confounding can occur simultaneously.

genetic principal components (PCs) as fixed effects in the GWAS model[10]. It has long been appreciated that in human datasets top principal components often mirror major geographic patterns in the sample and therefore their inclusion in the GWAS model should capture and regress out major confounders[11,12,13,14,15]. Inclusion of top PCs is a very popular approach to correcting for population stratification due in part to its simplicity and interpretability. Overall, PCA combined with strict genome-wide significance thresholds has resulted in a high replication rate for significant loci[16,17]. However, because PCA tries to control for ancestry using a low-dimensional representation, more subtle confounding may not be captured and it is known to fail when distant family members are included in the GWAS panel (often called "cryptic relatedness" in the literature)[18,19,20,21]. Relatedly, it is difficult to determine the number of PCs to include in the model without making assumptions about the distribution of confounders.

A more sophisticated, albeit related, approach to controlling for confounding in association studies is the use of linear mixed models (LMMs) for GWAS[22,23]. In LMMs, relatedness between individuals is modeled via the inclusion of a random effect, the covariance of which is structured by the genetic relatedness matrix (GRM). Statistically, using an LMM with the GRM included as a random effect is equivalent to including *all* PCs as covariates, each with a normal prior[24,25,26]. In the PCA approach, including J PCs as fixed effects simply eliminates some of the dimensions along which genotypes and phenotype can be spuriously correlated. This can be thought of as modeling the genetic background in J dimensions and/or as assuming that major environmental confounders are correlated with these top axes of variation. This formulation demonstrates two limitations of PCA. First, it can result in loss of power from completely eliminating entire dimensions and second, it is unable to capture confounders, of either type, that align with lower PCs that were not included[26]. On the other hand, effect sizes estimated using LMMs eliminate variation along all PCs but in proportion to their eigenvalue. LMMs, therefore, have the benefit of modeling all axes of variation but are constrained in the amount of variation that can be captured by each.

The role of LMMs in population structure control depends on one's assumptions about the distribution of confounders. If genetic confounding is the primary concern, it is easy to conceptualize the LMM as modeling the genetic background, and LLMs are expected to decrease stratification by removing this effect[23,27,28,29]. Simulation under this scenario (i.e, with no environmental effects) confirms that LMMs that include the GRM as the sole control for population structure outperform an approach that uses only the PCs[18]. If one views LMMs as controlling primarily for environmental confounding, then the assumption is that environmental confounders have a distribution similar to the GRM[30] and contribute in proportion to their eigenvalue. It is unclear how plausible this assumption is in human datasets and in simulations with strong environmental confounding PCA outperforms LMMs[31]. One option, often done in practice, is to include top PCs as fixed effects alongside the GRM. This

hybrid approach combines the ability of fixed effects to completely eliminate axes of variation that we might expect to have strong confounding while also including some control along all axes. Therefore, if both types of confounding are of significant concern, the primary innovation of LMMs might be their ability to account for both large-scale population structure via inclusion of PCs and smaller-scale structure via inclusion of the GWAS panel GRM. Overall, the flexibility of LMMs, alongside recent algorithmic advances that have reduced the computational burden, has made LMMs the preferred approach for biobank-scale GWASs[23,27,28,29] but there are still open questions surrounding the optimal parameterization in the face of unknown confounders.

Finally, LD score regression has been an important development for both stratification control in GWAS and tests for associations between ancestry gradients and effect size estimates. LD score regression is designed to distinguish between polygenicity and confounding in GWAS summary statistics[32]. The core idea is that for highly polygenic traits there should be a positive relationship between an individual variant's LD score (i.e, the sum of r^2 values between the focal site and all other SNPs) and it's squared marginal effect size, where the slope of this relationship is proportional to the heritability of the trait. Confounding, on the other hand, will inflate effect sizes regardless of LD score such that the intercept of the regression is > 1. The intercept can then be used to re-scale the association statistics, which can be seen as a form of genomic control that is not overly conservative when the trait is in fact highly polygenic. However, re-scaling effect sizes is equivalent to moving the genome-wide significance threshold, and does not remove ancestry-associated bias in effect size estimates themselves, the key issue for polygenic scores.

In bivariate LD score regression, the products of effect sizes for two traits are regressed on LD scores. The slope of this regression is proportional to the genetic correlation between traits, while the intercept is still interpreted as capturing confounding[33]. Bivariate LD score regression has been used to test for directional selection on complex traits by replacing one

set of effect sizes with a set of site-level summary statistics that measure allele frequency change or differentiation[34,35]. A significantly positive slope with minimal inflation of the intercept is interpreted as evidence of selection acting on alleles associated with the trait. However, previous work has raised concerns about spurious inflation of the LD score slope due to background selection[36]. Overall, LD score regression is a very important tool in complex trait genetics but it is not yet clear if it produces effects sizes or selection statistics that are completely immune from stratification bias.

Alongside the development of statistical correction approaches, numerous bias avoidance strategies have emerged with the proliferation of GWAS datasets. Here I cover three approaches: family-based association tests, homogeneous GWAS panels, and evolutionarily diverged GWAS and prediction panels in the context of polygenic scores. The first class of bias mitigation strategies is family-based association tests. One common study design is within-sibship GWAS, where the difference between sibling phenotypes is regressed on the difference in genotypes. Because full siblings share the same parents and are typically raised in the same (or a very similar) environment, any systematic phenotypic differences between sibling pairs will be due to the random transmission of parental alleles to each sibling[37]. Therefore, effect size estimates from within-sibship GWASs should be immune to population stratification[38]. It should be mentioned that effect sizes from within-sibship GWASs are only guaranteed to be unbiased if the allele has the same average effect in a heterozygous vs. a homozygous parent, an assumption that could be violated by gene-by-environment interactions, gene-by-gene interactions, or LD pattern differences[39]. However, this is separate from ancestry-based stratification that I am considering in this thesis and family based association tests are considered the "gold standard" for protecting effect sizes from stratification bias. Empirically, effect size estimates from within-sibship GWASs are generally smaller and result in smaller heritability estimates[40,41,42], suggesting that population GWAS effects are partially capturing indirect effects, including stratification effects, on

the trait. While family-based association studies are a compelling way forward in regards to eliminating ancestry-based bias in polygenic scores, current datasets are limited in terms of sample size and diversity of phenotypes; therefore, limiting the type of questions that can be answered using these data. The scale of current "unrelated" population GWAS datasets is unmatched, hence ancestry-based bias is an important problem to solve if we are to take full advantage of polygenic scores as a tool.

For population GWAS, a very common mitigation strategy is to select homogeneous samples, with respect to both ancestry and the environment. Focusing on GWAS panels with comparatively little genetic diversity decreases the magnitude of potential spurious correlations between genetic variants and phenotypes. Additionally, limiting sampling to narrow geographic regions should similarly decrease variation in the distribution of environmental effects on the phenotype, though this is difficult to measure empirically. Many of the largest available GWAS datasets (e.g UK Biobank[43], Biobank Japan[44], FinnGen[45]) consist of individuals living in the same country and it is not uncommon for researchers to further restrict their GWAS and polygenic score analyses to a subset of individuals with similar ancestry[43,29].

Large homogeneous GWAS panels combined with the use of PCs and/or LMMs is often considered the "silver standard" for controlling for population stratification in GWAS and have been largely successful in minimizing the number of false positive single variant associations[46]. However, effect size estimates can still exhibit slight residual stratification biases that are not large enough to significantly alter the false discovery rates for individual variants. These biases can be compounded when aggregating across loci in polygenic scores leading to confounded predictions in which the ancestry-associated effects are mistaken for genetic effects. A recent simulation study[47] used both PCA and LMMs to correct for population structure in the face of both broad and localized environmental confounders and found that both methods resulted in slightly biased effect sizes that confounded polygenic scores

built in a separate panel drawn from the same demographic model. The confounding was particularly acute for localized environmental confounders not easily captured by top PCs. Empirically, polygenic scores exhibit geographic clustering even in relatively homogeneous samples and after strict control for population stratification[48,49,50,51] and display patterns that are not robust to the choice of effect size estimates[52,53], suggesting that residual ancestry stratification may still be a problem.

This issue has been particularly apparent in the detection of directional selection acting on complex traits. In some ways, 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 frequencies induced by selection on a polygenic trait into a detectable signal[54,55,56,57]. Several research groups have developed and applied methods to detect these signals[58,59,34,60,61,62,63,64]. 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[65,66,58,59,67,68,69,60,70,71,34]; 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[36,72,62,63]. Further analysis suggested that much of the original signal could be attributed to spurious correlations between effect size estimates and patterns of allele frequency variation, presumably induced by uncorrected ancestry stratification in the original GWAS[36,72].

Recently, in the context of selection tests, Chen et al. (2020)[73] 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 Le et al. (2022)[6] and Akbari et al. (2024)[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[67], 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[74,75,76,77,78]. Less accurate polygenic scores are then less able to capture the evolution of the mean polygenic score, all else equal[77]. 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[79]. 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 and will necessitate a more precise understanding of how ancestry overlap between panels influences stratification bias.

In this thesis, I investigate confounding in polygenic scores using theory, simulations, and empirical analysis. In Chapter 2, we first model the covariance of genotypes between a GWAS and test panel in terms of an underlying population genetic model, and give 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. We show how including this singular axis of ancestry as a covariate in the GWAS removes effect size bias, such that the resulting polygenic scores are unbiased with respect to the target axis in the test panel. In Chapter 3, we evaluate ways to control for confounding along this axis in practice, including the standard PCA-based approach, as well as a new approach that uses test panel genotypes to estimate the target axis directly. Using simulations and theory, we find that the utility of each approach depends on a host of factors, including the number of independent SNPs used to compute the correction, the number of samples in the GWAS panel, and the amount of

variance in the GWAS panel explained by the target axis. Finally, building on our theoretical and simulation results, we develop a hybrid approach for empirical applications that allows researchers to use the direct approach to estimate how well protected any given polygenic score association test is by different numbers of PCs. We apply our approach to 30 different tests and 17 different phenotypes, including re-analyzing previously documented cases of polygenic selection. In Chapter 4 I summarize the progress we have made in understanding and controlling for stratification bias in polygenic score association tests and highlight future research questions related to our work.

CHAPTER 2

A POPULATION GENETIC MODEL OF BIAS IN POLYGENIC SCORES

2.1 Introduction

Since the first introduction of polygenic scores to the human genetics literature, Purcell et al. (2009)[2], ancestry-based confounding has been a concern. In that study, the authors built polygenic scores for schizophrenia and state that "We eliminated several possible confounders, with emphasis on subtle population stratification. Defining score alleles in British Isles samples and testing in target samples from Sweden, Portugal and Bulgaria, and vice versa, we observed a similar pattern of results. It is unlikely that the same substructure is overrepresented in the corresponding phenotype class when discovery and target samples are from distinct populations". This approach of using evolutionarily diverged GWAS and test panels, which we discuss in depth in Chapter 1, operates under a model where confounding due to population stratification is a function of the overlap in structure between panels.

The intuition behind this model is conceptually sound: if effect sizes become correlated with ancestry in the GWAS panel (see Figure 1.1), either due to correlated environmental effects or background genetic effects, then polygenic scores computed using those effect sizes will also become spuriously correlated with the same ancestry gradient(s), if they exist, in the prediction panel. In 2015, Robinson et al.[67] wrote down a population genetic model for stratification bias in effect size estimates. They consider a stratified GWAS sample where individuals are drawn equally from two diverged subpopulations whose mean phenotypes differ due to both an allele frequency difference at the causal focal site and a non-genetic effect. The authors show that if the non-genetic effect is ignored, then the estimated effect size is biased, and that the bias is a function of the F_{ST} between the two populations (see Equation 2.3 in their supplement). Next, they show that if those uncorrected effect sizes

are used to build polygenic scores in individuals drawn from the same two subpopulations, the scores will be biased in the direction of the non-genetic confounding effect in the GWAS panel. The Robinson et al. (2015) model clarifies how bias is produced in terms of common population quantities (e.g F_{ST}, heterozygosity, and allele frequency differences) for a simple two-population split model.

The Robinson et al. model, while very useful, is a simplified toy model with only a single axis of population structure in both the GWAS and test panels. The reality of human genetic variation is far more complex, and often GWAS and prediction panels come from different datasets, each with panel-specific and (potentially) shared structure. Additionally, there is also a growing appreciation for the fact that discrete population models, while useful in some applications, are not the most faithful representation of human genetic variation [80,81]. Combining the above observations, it is difficult to determine how diverged panels have to be to ensure that polygenic scores are immune to stratification. Overall, while the empirical approach of using evolutionarily diverged panels surely does reduce the overlap in structure, and therefore the potential for confounding, I believe that developing a more comprehensive model that clarifies precisely how continuous shared structure between panels generates bias in the distribution of polygenic scores is needed.

In this Chapter, we combine elements of both population and statistical genetic modeling to investigate how bias in the distribution of polygenic scores is produced. Specifically, we model the relatedness between individuals in the GWAS and test panels, and show how confounding in the GWAS panel translates into biased polygenic score predictions in a way that depends on the cross-panel relatedness. We then explore the procedure of testing for an association between polygenic scores and axes of ancestry variation in the test panel. We show that for any given test, there exists a single axis of ancestry variation in the GWAS panel that must align with the confounders to produce a biased test. Building on this result, we outline a set of conditions for guaranteeing an unbiased association test. Finally, we

describe a new theoretical approach for meeting these conditions, and revisit the standard PCA approach in light of our results.

2.2 Model

To model the distribution of genotypes in both panels, we assume that each individual's expected genotype at each site can be modeled as a linear combination of contributions from a potentially large number of ancestral populations, which are themselves related via an arbitrary demographic model. Natural selection, genetic drift, and random sampling each independently contribute to the distribution of genotypes across panels, and we make the approximation that these three effects can be combined linearly. In Section 2.8.1 we develop the full population model, which we then extend to individuals. Here, we present just the individual genotype model, along with our model of the phenotype.

2.2.1 Genotypes

We consider two samples of individuals, one to compose the GWAS panel and one to compose the test panel. Individuals in each panel are created as mixtures of an arbitrary number of K underlying populations, related via an arbitrary demographic model (see supplement Section 2.8.1), where a_ℓ is the ancestral allele frequency at site ℓ. There are N test panel individuals, and the vector of deviations of their genotypes from the mean genotype in the ancestral population ($2a_\ell$) is

$$X_\ell = X_{\ell,D} + X_{\ell,S} + X_{\ell,B}, \tag{2.1}$$

where $X_{\ell,D}$ and $X_{\ell,S}$ are the deviations due to drift and natural selection, respectively. We can think of the quantity $2a_\ell + X_{\ell,D} + X_{\ell,S}$ as giving a set of expected genotypes, given the evolutionary history of the populations from which the test panel individuals were sampled

from, while $X_{\ell,B}$ contains the binomial sampling deviations across individuals given these expected genotypes.

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

$$G_\ell = G_{\ell,D} + G_{\ell,S} + G_{\ell,B}, \tag{2.2}$$

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

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

$$Var\left(\begin{bmatrix} X_{\ell,D} \\ G_{\ell,D} \end{bmatrix}\right) = 4a_\ell\left(1-a_\ell\right)\mathbf{F} \tag{2.3}$$

where the matrix

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{XX} & \mathbf{F}_{XG} \\ \mathbf{F}_{GX} & \mathbf{F}_{GG} \end{bmatrix} \tag{2.4}$$

contains the within and between panel relatedness coefficients. Entries of $\mathbf{F}$ give the relatedness between pairs of individuals, given the underlying demographic model and the fraction of ancestry each individual derives from each population. Thus, the entries of $\mathbf{F}$ are directly related to the expected pairwise coalescent times between pairs of samples, given the demographic model[82].

2.2.2 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

$$\begin{aligned} y &= \sum_{\ell}^{S} \beta_\ell G_\ell + e \\ &= u + e \end{aligned} \tag{2.5}$$

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 effect can differ in some arbitrary but unknown way across individuals with no assumptions about the covariance structure. 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 Equation 2.2, the genetic effect, u, can be broken down into the contributions from drift, selection, and binomial sampling, such that $u = u_D + u_S + u_B$. Here $u_S = \sum_{\ell}^{S} \beta_\ell G_{\ell,S}$ contains fixed effects reflecting the expected genetic contributions to the phenotype, given history of selection acting on the phenotype, and given the ancestries of the individuals in the GWAS panels (see supplement Section 2.8.1). Both u_D and u_B have expectation zero, so $\mathbb{E}[u] = u_S$. The vector of individuals' expected phenotypes, given their ancestry and socio-environmental contexts, is therefore given by $u_S + c$. We assume that these are not known.

2.3 The impact of stratification bias on polygenic scores

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

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

$$Z = \sum_{\ell}^{S} \beta_\ell X_\ell. \tag{2.6}$$

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

$$\begin{aligned} \hat{\beta}_\ell \mid G_\ell, u, e &= \frac{y^\top G_\ell}{G_\ell^\top G_\ell} \\ &= \beta_\ell + \frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} + \frac{e^\top G_\ell}{G_\ell^\top G_\ell} \end{aligned} \tag{2.7}$$

where we have decomposed the genetic effect into the causal contribution from the focal site and the contribution from the background, i.e. $u = \beta_\ell G_\ell + u_{-\ell}$. This allows us to further decompose the marginal association in Equation 2.7 into the causal effect (β_ℓ), the association between the focal site and the background genetic contribution from all other

sites ($u_{-\ell}^\top G_\ell / G_\ell^\top G_\ell$), and the association with the environment ($e^\top G_\ell / G_\ell^\top G_\ell$).

The deviation of an allele's estimated effect size from its expectation depends in part on $G_{\ell,D}$, the component of variation in the GWAS panel genotypes due to genetic drift. Because $G_{\ell,D}$ can be 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 polygenic scores,

$$\mathbb{E}\left[\hat{Z} - Z\right]^\top = \mathbb{E}\left[\sum_{\ell=1}^{S} \frac{u^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top + \sum_{\ell=1}^{S} \frac{e^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top\right] \tag{2.8}$$

$$\approx \frac{S}{M}\left(\mu_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX}, \tag{2.9}$$

(see Section 2.8.3) where μ_S is the vector of expected genetic backgrounds, c is the vector of expected environmental effects, and

$$\tilde{\mathbf{F}}_{GX} = \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M}\right]$$

$$\approx \frac{\mathbf{F}_{GX}}{1+\overline{F_G}}. \tag{2.10}$$

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.

If the GWAS and test panels do not overlap in population structure, then $\tilde{\mathbf{F}}_{XG} = \mathbf{0}$, and the polygenic scores are unbiased with respect to ancestry (i.e. $\mathbb{E}\left[\hat{Z} - Z\right] = 0$), independent of the confounders, μ_S and c[2,73,6]. Stratification may still bias individual effects, but these residual biases are indistinguishable from noise from the perspective of the polygenic scores, as they are uncorrelated with all axes of population structure present in the test panel.

2.4 Bias in polygenic scores leads to biased polygenic score associations

We want to test the hypothesis that the polygenic scores are associated with some test vector, T. We assume that T is measured only in the test panel, and might represent an eco-geographic variable of interest (e.g latitude[59] or an encoding of whether one lives in a particular geographic region or not[49,83], the fraction of an individual's genome assigned to a particular "ancestry group"[58,60], or one of the top genetic principal components of the test panel genotype matrix[61]).

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

$$q = \frac{1}{N} Z^\top T. \tag{2.11}$$

A more powerful test is available by modeling the neutral correlation structure among individuals due to relatedness (see Section 2.8.7), but the simpler i.i.d. model presented here is sufficient for our purposes. Under the null model where selection has not perturbed allele frequencies in the test panel, $\mathbb{E}[q] = 0$, reflecting the fact that genetic drift is directionless.

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

$$\begin{aligned} \mathbb{E}\left[\hat{q} - q\right] &= \mathbb{E}\left[\hat{Z} - Z\right]^\top T \\ &\approx \frac{S}{NM} \left(\mu_S^\top + c^\top\right) \tilde{\mathbf{F}}_{GX} T. \end{aligned} \tag{2.12}$$

Therefore, we expect the polygenic score association test to be biased when the test vector (T)

aligns with the vector of expected phenotypes ($\mu_S + c$) in a space defined by the cross panel genetic 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 a 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 expressing the association statistic, q, in a different way. Specifically, we can reframe this test as a statement about the association between the effect sizes and a set of genotype contrasts, $r_\ell = \frac{1}{N} X_\ell^\top T$, which measure the association between the test vector and the genotypes at each site[59]. Writing β and r for the vectors of effect sizes and genotype contrasts across loci, the association test statistic can be rewritten as

$$q = \beta^\top r. \tag{2.13}$$

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

$$\begin{aligned} \mathbb{E}\left[\hat{q} - q\right] &= \frac{S}{M}\mathbb{E}\left[\left(\hat{\beta}^\top - \beta^\top\right) r\right] \\ &\approx \frac{S}{M}\left(\mu_S^\top + c^\top\right)\tilde{F}_{Gr} \end{aligned} \tag{2.14}$$

where

$$\begin{aligned} \tilde{F}_{Gr} &= \mathbb{E}\left[\frac{G_{\ell,D} r_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M}\right] \\ &= \tilde{\mathbf{F}}_{GX} T. \end{aligned} \tag{2.15}$$

Here Equation 2.14 expresses the bias entirely in terms of vectors that belong to the GWAS panel: for each GWAS panel individual m, $\tilde{F}_{Gr,m}$ measures the covariance between individual m's genotype and the genotype contrasts of the test, standardized at each site by the variance of genotypes across individuals in the GWAS panel (Equation 2.15). Thus, $\hat{q}$ is biased when

the vector of expected phenotypes ($\mu_S + c$) aligns with this vector of standardized covariances ($\tilde{F}_{Gr}$). Confounders that are orthogonal to this axis do not generate bias in the association test, even if they bias the polygenic scores along other axes.

2.5 Controlling for stratification bias in polygenic association tests

Given the above results, how can we ensure that patterns we observe in the distribution of polygenic scores are not the result of stratification bias? As discussed above, a conservative solution is to prevent bias by choosing a GWAS panel that does not have any overlap in population structure with the test panel, but this is not ideal due to the well-documented portability issues that plague polygenic scores[74,84,78], 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. Alternatively, our analysis above suggests that including $\tilde{F}_{Gr}$ as a covariate in the GWAS model is a sufficient condition for an unbiased test regardless of the pattern of confounding that exists in the GWAS panel.

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

If we include $\tilde{F}_{Gr}$ as a single fixed-effect covariate in the GWAS model, variation along $\tilde{F}_{Gr}$ can no longer be used to estimate effect sizes. As a result, $\hat{\beta}$ is uncorrelated with genotype contrasts r under the null. If there is confounding along other shared axes of ancestry variation, the polygenic scores may still be biased along other axes, as

$$\mathbb{E}\left[\hat{Z} - Z\right]^\top \approx \frac{S}{M}\left(\mu_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \tag{2.16}$$

where

$$\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \approx \mathbf{P}\tilde{\mathbf{F}}_{GX} \tag{2.17}$$

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

$$\begin{aligned} \mathbb{E}\left[\hat{q} - q\right] &\approx \frac{S}{NM}\left(\mu_S^{\top} + c^{\top}\right)\tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} T \\ &\approx 0 \end{aligned} \tag{2.18}$$

and thus the polygenic score association test is unbiased (see Sections 2.8.5 and 2.8.6).

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

A standard approach to controlling for population stratification in polygenic scores is to include the top J principal components of the GWAS panel genotype matrix as covariates in the GWAS for some suitably large value of J[10]. In our model, how does this approach relate to including $\tilde{F}_{Gr}$ as a covariate in the GWAS?

As outlined in Section 2.2.1, $\mathbf{F}_{GG}$ contains the expected within panel relatedness for the individuals in the GWAS panel, the structure of which is determined by the demographic model. If we could take the eigendecomposition of $\mathbf{F}_{GG}$ directly, the resulting PCs are what we refer to as "population" PCs. The number of population PCs that correspond to structure is entirely dependent on the population model. For example, below (Section 3.3.1), we simulate under a 4 population sequential split model (Figure 3.1), in which case there are three population PCs that reflect real underlying structure. Later (Section 3.3.2) we simulate under a symmetric equilibrium migration model on a six-by-six lattice grid (Figure 3.3), in which case there are 35 population PCs reflecting the underlying population

structure. Including these population PCs as covariates in the GWAS would be sufficient to remove all ancestry-associated bias in effect size estimates and render the resulting polygenic scores uncorrelated with any axis of ancestry variation under the null hypothesis.

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

$$\tilde{F}_{Gr} = \sum_i \eta_i U_i \tag{2.19}$$

where U_i is the i^{th} 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* population PCs are included as covariates, but the relative magnitude of the contributions from different PCs is fixed, and we estimate only a single slope that scales the contributions from all the PCs jointly, i.e.,

$$y = G_\ell \beta_\ell + \left(\sum_i \eta_i U_i \right) \omega + e. \tag{2.20}$$

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

A second interpretation of the PC correction approach is that it operates on a hypothesis that the major axes of confounding in a given GWAS panel (i.e., μ_S and c in our notation) can be captured by the included PCs[30]. If this condition is met, effect size estimates will be unbiased with respect to all axes of ancestry variation, whether they exist within a given test panel or not, and therefore any polygenic score association test that uses these effect size estimates will be unbiased with respect to ancestry as well. Combining this interpretation

with the results from above, population PCs should successfully eliminate bias in polygenic score association tests if the J PCs included in the GWAS either capture the confounding effects on the phenotype, eliminating all effect size bias, or if they capture $\tilde{F}_{Gr}$, ensuring that effect size bias relevant to the test is removed.

2.6 Conclusion

Ancestry stratification in polygenic scores has been recognized as a potential implementation problem since their inception to the human genetics literature[2]. Intuitively, stratification bias in polygenic scores should depend on the overlap in population structure between the GWAS and prediction panels. However, to the best of our knowledge, no existing model has jointly analyzed the GWAS and polygenic score procedures in a comprehensive way that reflects the complex and continuous nature of human genetic variation. In this Chapter, we developed this model and showed that the bias in an individual polygenic score is proportional to the dot product of the vector of expected confounders and the kinship coefficients between the test panel individual and every individual in the GWAS panel. This expression (Equation 2.9) illustrates how bias in polygenic scores is generated and provides insight into paths that minimize potential biases.

It is worth noting some of the assumptions we make throughout this Chapter. First, I have emphasized the need for models that do not rely on discrete, panmictic populations. However, our full model (see Section 2.8.1) does model an individual's ancestry as a mixture of K ancestral populations that are themselves related via an arbitrary demographic model. Under our model, it is clear that the expected covariance between pairs of individuals, F_{ij}, depends on how similar their ancestry proportions are. We chose to include this underlying population model because it provides a concrete and familiar interpretation of $\mathbf{F}$. However, none of our key results about the distribution of polygenic scores (see Equation 2.9) or the bias in the association test statistics (see Equation 2.12) rely on the assumption that

there is an underlying population model. Rather, they follow from the assumption that the covariance structure of the drift component of the genotypes is proportional to $\mathbf{F}$. If the expected covariance matrix is known, or can be written down for a different type of generative models (e.g., a model in which individuals in the sample are related via a known pedigree, or a continuous model in which the expected covariance is a function of the spatial separation among individuals), all our results follow from there, and assumptions about discrete populations are not needed.

Our choice to model individual genotypes as mixtures of underlying populations is related to our decision to explicitly model the genotypes as a linear combination of the contributions of selection, drift, and binomial sampling. Specifically, we chose to frame our polygenic association test as testing for a perturbation of test panel genotypes that aligns with the test vector, where the perturbation is due to a selection event that occurred in an ancestral population. Again, we make this choice because models of selection acting on a polygenic trait are well-studied[59,54,55,85,86], we were able to define a clear null and alternative hypothesis, and we have expertise in polygenic adaptation tests. However, our key results still apply when perturbations in the polygenic scores are due to other non-neutral processes (e.g, phenotype biased migration, assortative mating on ancestry, phenotype dependent ascertainment). In practice, $\hat{q}$ is usually compared to an empirical null constructed by breaking the relationship between effect sizes and genotypes, either by resampling frequency-matched SNPs or randomly flipping the sign of the effect sizes[59,87,6,35]. Thus, these types of polygenic score association tests, as presented here, are unable to distinguish between causal mechanisms causing perturbations in real data anyway.

In terms of controlling for bias, the results of the theoretical work in this Chapter illuminate several paths to mitigating the impact of confounding. I began this Chapter by drawing attention to the fact that concern about confounding in polygenic scores, and the strategy of using diverged panels as a way to diminish its effects, is common in the polygenic

score literature. Our work suggests that $\tilde{\mathbf{F}}_{GX} = \mathbf{0}$ is the condition that needs to be met in order to ensure unbiased polygenic scores. For polygenic score association tests, only a single axis of variation in the GWAS panel, $\tilde{F}_{Gr}$, needs to be controlled for to eliminate bias in $\hat{q}$. In Section 2.5.1, we used properties of multiple regression to show that including $\tilde{F}_{Gr}$ itself as a covariate in the GWAS model removes any confounders along this axis directly. Additionally, in Section 2.5.2, we outline the conditions that need to be met for J population PCs to remove any bias in $\hat{q}$.

In the realm of theoretical models, it is not too difficult to come up with demographic scenarios where there is no overlap in structure between GWAS and prediction panels, making $\tilde{\mathbf{F}}_{GX} = \mathbf{0}$. Any tree-based model where individuals from each panel don't share any internal branch length would satisfy this condition (see Figure 3.1C for an example). Similarly, in cases of overlapping population structure, controlling for $\tilde{F}_{Gr}$ via population PCs would amount to including the correct linear combination of population PCs that records structure along shared branches relevant to the test (PC 1 in Figure 3.1A is an example of this).

However, $\tilde{\mathbf{F}}_{GX}$, $\tilde{F}_{Gr}$, and population PCs are theoretical quantities. In reality, we have to take the insights gained from our model and figure out how to correct for bias in practice when we must rely on estimating quantities of interest. The reason I make a distinction between "population" PCs, which come from taking the eigen decomposition of $\mathbf{F}_{GG}$ directly, and "sample" PCs which, come from taking the eigen decomposition of the observed genotypic covariance matrix, is that the success of PCA, as an approach to capturing $\tilde{F}_{Gr}$, will depend on the accuracy of sample PCs as estimators of the population PCs. In the next Chapter, we explore the factors that affect the accuracy of sample PCs, develop an approach that uses test panel genotypes to estimate $\tilde{F}_{Gr}$ directly, and propose a test for independence between an ancestry gradient in the test panel and structure in the GWAS panel. The theoretical results from this Chapter were vital for my work in the next Chapter and helped to highlight both the challenges and opportunities in developing tests for differences in polygenic scores

in the presence of confounding.

2.7 Author contributions

Jennifer Blanc: Conceptualization, methodology, formal analysis, investigation, writing - original draft, writing - review and editing, funding acquisition

Jeremy Berg: Conceptualization, methodology, formal analysis, investigation, writing - original draft, writing - review and editing, supervision, project administration, funding acquisition

2.8 Extended model and results

2.8.1 Full Model

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.

Population model

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

$$p_\ell = a_\ell + \Delta p_{\ell,D} + \Delta p_{\ell,S} \tag{2.21}$$

where a_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 $\Delta p_{\ell,D} \sim MVN\left(0, a_\ell\left(1-a_\ell\right)\mathbf{F}_{pop}\right)$, where $\mathbf{F}_{pop}$ describes the covariance structure imposed by the population model[88,89,90,91]. Selection induces an additional deviation of $\Delta p_{\ell,S} = \eta\beta_\ell a_\ell\left(1-a_\ell\right)A_{pop}$, where β_ℓ is site ℓ's effect (in standardized units) on the phenotype of interest, η 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. Under the null, $\eta = 0$. Under the alternative, $\eta \sim N\left(0, \sigma_\eta^2\right)$, so that

$$\Delta p_{\ell,D} + \Delta p_{\ell,S} \sim MVN\left(0, a_\ell\left(1-a_\ell\right)\left(\mathbf{F}_{pop} + \sigma_\eta^2\beta_\ell^2 a_\ell\left(1-a_\ell\right)A_{pop}A_{pop}^\top\right)\right) \tag{2.22}$$

where $A_{pop}A_{pop}^\top$ is a rank one matrix accounting for the impact of selection on the joint distribution of allele frequencies. We assume the trait is sufficiently polygenic and/or that selection is sufficiently weak that

$$\beta_\ell^2 a_\ell\left(1-a_\ell\right) \ll \frac{1}{\sigma_\eta^2} \quad \text{for all } \ell, \tag{2.23}$$

so that for any individual variant, the effects of drift dominate over those of selection (i.e. $\Delta p_{\ell,D} >> \Delta p_{\ell,S}$ for all ℓ). This condition is met, for example, if there are at least hundreds of loci and σ_η^2 is on the order of one.

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 ($2a_\ell$) is

$$X_\ell = X_{\ell,D} + X_{\ell,S} + X_{\ell,B}, \tag{2.24}$$

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 $2a_\ell + X_{\ell,D} + X_{\ell,S}$ 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 panel individuals, the deviation of their genotypes can be decomposed as

$$G_\ell = G_{\ell,D} + G_{\ell,S} + G_{\ell,B}, \tag{2.25}$$

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, $2a_\ell + G_{\ell,D} + G_{\ell,S}$ specifies this set of individual specific expected genotypes).

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

$$Var\left(\begin{bmatrix} X_{\ell,D} \\ G_{\ell,D} \end{bmatrix}\right) = 4a_\ell \left(1 - a_\ell\right) \mathbf{F} \tag{2.26}$$

where the matrix

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{XX} & \mathbf{F}_{XG} \\ \mathbf{F}_{GX} & \mathbf{F}_{GG} \end{bmatrix} \tag{2.27}$$

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 $X_{\ell,S} = 2\eta\beta_\ell a_\ell (1 - a_\ell) A_X$ and $G_{\ell,S} = 2\eta\beta_\ell a_\ell (1 - a_\ell) A_G$, 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.

The joint covariance structure of the individual genotypes is therefore

$$Var\left(\begin{bmatrix} X_\ell \\ G_\ell \end{bmatrix} \right) = 4a_\ell (1 - a_\ell) \left(\mathbf{F} + \sigma_\eta^2 \beta_\ell^2 a_\ell (1 - a_\ell) \mathbf{A} + \mathbf{B}_\ell \right) \tag{2.28}$$

where

$$\mathbf{A} = \begin{bmatrix} A_X A_X^\top & A_X A_G^\top \\ A_G A_X^\top & A_G A_G^\top \end{bmatrix} \tag{2.29}$$

and $\mathbf{B}_\ell$ is a diagonal matrix with entries accounting for overdispersion of the binomial sampling step due to evolutionary variation of the underlying allele frequencies (i.e., from drift and selection). The exact details of this matrix are not important other than that it is diagonal; i.e., it does not contribute to covariance among individuals.

Phenotypes

As in Section 2.2.2, we assume that the vector of mean-centered phenotypes for the M individuals in the GWAS panel can be written as

$$\begin{aligned} y &= \sum_{\ell}^{S} \beta_\ell G_\ell + e \\ &= u + e \end{aligned} \tag{2.30}$$

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. The distribution of environmental effects is $e \sim MVN\left(c, \sigma_e^2 \mathbf{I}\right)$, where c is the vector of expected environmental effects. We consider c to be fixed (as opposed to random).

The genetic effect, u, can be broken down into the contributions from drift, selection, and sampling. In this case, $u_S = \sum_{\ell}^{S} \beta_\ell G_{\ell,S}$ 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 $V_A = 2\sum_{\ell}^{S} \beta_\ell^2 a_\ell \left(1 - a_\ell\right)$ is the additive genetic variance in the original ancestral population. The full covariance structure of u is therefore

$$Var(u) = 2V_A\mathbf{F} + \sigma_\eta^2 \left(V_A^2 + C_A^2\right) \mathbf{A} + 2V_A\mathbf{B}_u \tag{2.31}$$

where $C_A = \sigma_\eta^2 \sum_{\ell \neq \ell'} \beta_\ell^2 \beta_{\ell'}^2 a_\ell \left(1 - a_\ell\right) a_{\ell'} \left(1 - a_{\ell'}\right) A_G A_G^\top$ and $\mathbf{B}_u = \sum_\ell \mathbf{B}_\ell$

We are ultimately interested in the behavior of polygenic scores given that some stratification effect exists in the phenotype. To this end, from here forward we will generally condition on the strength of selection on the phenotype, η, treating it as a fixed effect along-

side c. Doing so gives us the condition expectations and covariance structures:

$$\mathbb{E}\left[y \mid \eta, c\right] = u_S + c \tag{2.32}$$

and

$$Var\left(y \mid \eta, c\right) = 2V_A\left(\mathbf{F} + \mathbf{B}_u\right) + \sigma_e^2\mathbf{I}. \tag{2.33}$$

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

$$Z = \sum_{\ell}^{S} \beta_\ell X_\ell. \tag{2.34}$$

We continue now to condition on the effects of selection. Doing so, 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 to drift and binomial sampling, respectively. Therefore, the vector of expected polygenic scores for the N individuals in the test panel is

$$\begin{aligned}
\mathbb{E}\left[Z \mid \eta\right] &= \sum_{\ell}^{S} \beta_\ell \mathbb{E}\left[X_\ell \mid \eta\right] \\
&= \sum_{\ell}^{S} \beta_\ell X_{\ell,S} \\
&= \eta V_A A_X.
\end{aligned} \tag{2.35}$$

Here the expected polygenic scores depend on the strength of selection on the phenotype, the ancestral additive genetic variance, and the extent to which each individual in the test

panel inherits from the selection event.

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 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

$$Z = qT + \varepsilon \tag{2.36}$$

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 2.8.7). We assume that the test vector, T, is scaled to have a variance of one, so the slope is given by

$$q = \frac{1}{N} Z^\top T, \tag{2.37}$$

and its conditional expectation by,

$$\begin{aligned} \mathbb{E}\left[q \mid \eta\right] &= \frac{1}{N} \mathbb{E}\left[Z \mid \eta\right]^\top T \\ &= \frac{1}{N} \eta V_A A_X^\top T \end{aligned} \tag{2.38}$$

Under the null model, $\eta = 0$, 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 $\eta \neq 0$ (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 also re-frame this test as a statement about the association between the effect sizes and a set of genotype contrasts, $r_\ell = \frac{1}{N} X_\ell^\top T$, which measure the association between the test vector and the genotypes at each site[59]. Similar to the decomposition of the test panel genotypes above, we can decompose the genotype contrasts as

$$r_\ell = r_{\ell,D} + r_{\ell,S} + r_{\ell,B}, \tag{2.39}$$

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

$$q = \beta^\top r. \tag{2.40}$$

2.8.2 Expectation of marginal effects

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

$$\hat{\beta}_\ell \mid G_\ell, u, e = \frac{y^\top G_\ell}{G_\ell^\top G_\ell} \tag{2.41}$$

$$= \beta_\ell + \frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} + \frac{e^\top G_\ell}{G_\ell^\top G_\ell} \tag{2.42}$$

(Equation 2.7). Now, taking the expectation over the randomness due to drift and sampling in the GWAS panel genotypes and the total genetic effect, and over the random component

of the environmental effect on the phenotype, the expected effect size estimate at site ℓ is

$$\mathbb{E}\left[\hat{\beta}_\ell \mid \eta, c\right] = \beta_\ell + \mathbb{E}\left[\frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} \mid \eta\right] + \mathbb{E}\left[\frac{e^\top G_\ell}{G_\ell^\top G_\ell} \mid c\right] \tag{2.43}$$

$$= \beta_\ell + \mathbb{E}\left[u_{-\ell}^\top \mid \eta\right]\mathbb{E}\left[\frac{G_\ell}{G_\ell^\top G_\ell} \mid \eta\right] + \mathbb{E}\left[e^\top \mid c\right]\mathbb{E}\left[\frac{G_\ell}{G_\ell^\top G_\ell} \mid \eta\right] \tag{2.44}$$

$$= \beta_\ell + u_{S,-\ell}^\top \mathbb{E}\left[\frac{G_\ell}{G_\ell^\top G_\ell} \mid \eta\right] + c^\top \mathbb{E}\left[\frac{G_\ell}{G_\ell^\top G_\ell} \mid \eta\right] \tag{2.45}$$

$$\approx \beta_\ell + u_{S,-\ell}^\top \frac{\mathbb{E}\left[G_\ell \mid \eta\right]}{\mathbb{E}\left[G_\ell^\top G_\ell \mid \eta\right]} + c^\top \frac{\mathbb{E}\left[G_\ell \mid \eta\right]}{\mathbb{E}\left[G_\ell^\top G_\ell \mid \eta\right]} \tag{2.46}$$

$$\approx \beta_\ell + u_{S,-\ell}^\top \frac{G_{\ell,S}}{4a_\ell(1-a_\ell)(1+\overline{F}_G)} + c^\top \frac{G_{\ell,S}}{4a_\ell(1-a_\ell)(1+\overline{F}_G)} \tag{2.47}$$

$$\approx \beta_\ell + \eta A_G^\top V_{A,-\ell} \frac{\eta\beta_\ell A_G}{2(1+\overline{F}_G)} + c^\top \frac{\eta A_G}{2(1+\overline{F}_G)} \tag{2.48}$$

$$\approx \beta_\ell + \frac{\eta^2\beta_\ell V_{A,-\ell}}{2(1+\overline{F}_G)} + \frac{\eta c^\top A_G}{2(1+\overline{F}_G)} \tag{2.49}$$

$$\approx \beta_\ell \left(1 + \frac{1}{2}\frac{\eta^2 V_{A,-\ell}}{1+\overline{F}_G}\right) + \frac{1}{2}\frac{\eta c^\top A_G}{1+\overline{F}_G} \tag{2.50}$$

where $1+\overline{F_G} = 1+\frac{1}{M}\sum_{m=1}^{M} f_{mm}$ is the average level of self relatedness in the GWAS panel. The approximation at line 2.46 comes from approximating the expectation of the ratio as the ratio of expectations. There is an additional approximation in line 2.47 that comes from assuming that $\mathbb{E}\left[G_\ell^\top G_\ell \mid \eta\right] = 4a_\ell(1-a_\ell)(1+\eta^2\beta_\ell^2 a_\ell(1-a_\ell)+\overline{F}_G) \approx 4a_\ell(1-a_\ell)(1+\overline{F}_G)$, consistent with our assumption that the effects of selection are small at the level of individual loci, relative to the effects of drift and sampling.

Therefore, consistent with results from[67], the expected marginal effect estimate for a given site is expected to be equal to the causal effect if the null holds (i.e. if $\eta = 0$). In contrast, under our alternative model, the focal site and the genetic background have a positive empirical covariance as they are both influenced by selection. As a result, the expected marginal effect is biased away from zero (i.e., further in the direction of the causal

effect), and may be biased in either direction depending on whether the environmental confounder is positively or negatively associated with the axis along which selection has perturbed the focal allele. For the following analyses concerning bias in the distribution of polygenic scores and in polygenic score association tests (see Section 2.8.3 and 2.8.4) we assume that the above biases in marginal effect estimates induced by selection are small relative to those induced by shared drift between the GWAS and test panels[92,54,86,93,59].

2.8.3 The expected polygenic scores

Given the marginal effect estimates, the vector of N polygenic scores in the test panel is given by,

$$\hat{Z}^\top \mid \mathbf{G}, \mathbf{X}, u, e = \sum_{\ell=1}^{S} \hat{\beta}_\ell X_\ell^\top \tag{2.51}$$

$$= \sum_{\ell=1}^{S} \beta_\ell X_\ell + \sum_{\ell=1}^{S} \frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top + \frac{e^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top . \tag{2.52}$$

We are interested in the expected polygenic scores in settings where both genetic and environmental confounders exist. We can write this expectation as

$$\mathbb{E}\left[\hat{Z}^\top \mid \eta, c\right] = \sum_{\ell=1}^{S} \mathbb{E}\left[\beta_\ell X_\ell \mid \eta\right] + \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top \mid \eta\right] + \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{e^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top \mid \eta, c\right] . \tag{2.53}$$

The first sum is the expected "true" polygenic scores given the non-neutral effects on the test panel genotypes ($\sum_{\ell=1}^{S} \mathbb{E}\left[\beta_\ell X_\ell \mid \eta\right] = \sum_{\ell=1}^{S} \beta_\ell X_{\ell,S} = \mathbb{E}[Z]^\top = \eta V_A A_X$) while the second sum captures contributions to the polygenic scores due to associations between the genotypes at individual focal sites and the genetic component of the phenotype from all other sites ($u_{-\ell} = \sum_{\ell' \neq \ell} \beta_{\ell'} G_{\ell'}$). The third captures contributions due to associations between the genotypes and the environmental component. We consider each of the second two in turn,

beginning with the environment, which is more straightforward.

For each site ℓ, we can write

$$\mathbb{E}\left[\frac{e^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top \mid \eta, c\right] \approx \mathbb{E}\left[\frac{e^\top G_{\ell,D}}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})} X_{\ell,D}^\top \mid c\right] \tag{2.54}$$

$$\approx \mathbb{E}\left[e^\top \mid c\right] \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})}\right] \tag{2.55}$$

$$\approx \frac{c^\top}{M} \mathbb{E}\left[\frac{G_\ell X_\ell^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M}\right] \tag{2.56}$$

$$\approx \frac{c^\top}{M} \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M}\right] = \frac{1}{M} c^\top \tilde{\mathbf{F}}_{GX} \tag{2.57}$$

$$\approx \frac{c^\top}{M} \frac{\mathbb{E}\left[G_{\ell,D} X_{\ell,D}^\top\right]}{\mathbb{E}\left[(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M\right]} = \frac{1}{M} \frac{c^\top \mathbf{F}_{GX}}{\left(1+\overline{F_G}\right)}. \tag{2.58}$$

The approximation in line 2.54 arises from assuming that $G_{\ell,S}$ and $X_{\ell,S}$ are small compared to $G_{\ell,D}$ and $X_{\ell,D}$, and can therefore be ignored (see 2.8.2). Alternately, 2.54 is exact under the null that $\eta = 0$. In line 2.58, we make a ratio of expectations approximation. $\tilde{\mathbf{F}}_{GX}$ is the expected kinship matrix computed on standardized genotypes, which is approximately equal to $\frac{\mathbf{F}_{GX}}{(1+\overline{F_G})}$, with the approximation being better if $\overline{F_G}$ is small. Summing across sites, we have

$$\sum_{\ell=1}^{S} \mathbb{E}\left[\frac{e^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top \mid \eta, c\right] \approx \frac{S}{M} c^\top \tilde{\mathbf{F}}_{GX}. \tag{2.59}$$

where S is the number of causal sites.

For the contributions to the polygenic score arising from associations between genotypes

and the genetic component of the phenotype, we have

$$\mathbb{E}\left[\frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top \mid \eta\right] \approx \mathbb{E}\left[\frac{u_{-\ell}^\top G_{\ell,D}}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})} X_{\ell,D}^\top \mid \eta\right] \tag{2.60}$$

$$\approx \mathbb{E}\left[u_{-\ell}^\top \mid \eta\right] \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})}\right] \tag{2.61}$$

$$\approx \frac{u_{S,-\ell}^\top}{M} \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M}\right] \tag{2.62}$$

$$\approx \frac{u_{S,-\ell}^\top}{M} \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M}\right] = \frac{1}{M} u_{S,-\ell}^\top \tilde{\mathbf{F}}_{GX} \tag{2.63}$$

$$\approx \frac{u_{S,-\ell}^\top}{M} \frac{\mathbb{E}\left[G_{\ell,D} X_{\ell,D}^\top\right]}{\mathbb{E}\left[(G_{\ell,D}+G_{\ell,B})^\top (G_{\ell,D}+G_{\ell,B})/M\right]} = \frac{1}{M} \frac{u_{S,-\ell}^\top \mathbf{F}_{GX}}{\left(1+\overline{F_G}\right)}, \tag{2.64}$$

where $u_{S,-\ell} = \sum_{\ell' \neq \ell} \beta_{\ell'} G_{S,\ell'} = u_S - \beta_\ell G_{\ell,S}$. All sites in our model are unlinked, allowing us to treat $u_{-\ell}$ as independent from G_ℓ.

Summing across sites, we have

$$\sum_{\ell=1}^{S} \mathbb{E}\left[\frac{u_{-\ell}^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top\right] \approx \frac{1}{M} \sum_{\ell=1}^{S} u_{S,-\ell}^\top \tilde{\mathbf{F}}_{GX} \tag{2.65}$$

$$\approx \frac{1}{M} \sum_{\ell=1}^{S} \left(u_S^\top - \beta_\ell G_{\ell,S}^\top\right) \tilde{\mathbf{F}}_{GX} \tag{2.66}$$

$$\approx \frac{1}{M} \left(S u_S^\top - u_S^\top\right) \tilde{\mathbf{F}}_{GX} \tag{2.67}$$

$$\approx \frac{S-1}{M} u_S^\top \tilde{\mathbf{F}}_{GX} \tag{2.68}$$

$$\approx \frac{S}{M} u_S^\top \tilde{\mathbf{F}}_{GX}. \tag{2.69}$$

Putting genetic and environmental contributions together, the expected polygenic scores are

approximately

$$\mathbb{E}\left[\hat{Z}^\top\right] = \sum_{\ell=1}^{S} \mathbb{E}\left[\beta_\ell X_\ell\right] + \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{u^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top\right] + \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{e^\top G_\ell}{G_\ell^\top G_\ell} X_\ell^\top\right] \tag{2.70}$$

$$\approx \mathbb{E}\left[Z\right]^\top + \frac{S}{M}\left(u_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX}, \tag{2.71}$$

where we have made the conditioning implicit in our notation for the sake of consistency with main text. We continue in this vein for the rest of this Section: all expectations from here on should be interpreted as conditional on η and c.

2.8.4 Polygenic score association test bias using marginal effects

The expected value of our test statistic, $\hat{q}$, follows straightforwardly from the expected values of the polygenic scores. We have

$$\mathbb{E}\left[\hat{q}\right] = \frac{1}{N}\mathbb{E}\left[\hat{Z}^\top\right] T \tag{2.72}$$

$$\approx \mathbb{E}\left[q\right] + \frac{S}{NM}\left(\mu_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX} T = \mathbb{E}\left[q\right] + \frac{S}{NM}\left(\mu_S^\top + c^\top\right)\tilde{F}_{Gr}. \tag{2.73}$$

The bias in $\hat{q}$ is therefore approximately equal to

$$\mathbb{E}\left[\hat{q} - q\right] \approx \mathbb{E}[\hat{q}] - \mathbb{E}\left[q\right] \tag{2.74}$$

$$\approx \frac{S}{NM}\left(\mu_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX} T = \frac{S}{NM}\left(\mu_S^\top + c^\top\right)\tilde{F}_{Gr} \tag{2.75}$$

where

$$\tilde{F}_{Gr} = \mathbb{E}\left[\frac{G_{\ell,D} r_{\ell,D}^{\top}}{(G_{\ell,D}+G_{\ell,B})^{\top}(G_{\ell,D}+G_{\ell,B})/M}\right] \tag{2.76}$$

$$= \tilde{\mathbf{F}}_{GX} T \tag{2.77}$$

$$\approx \frac{W_G \mathbf{F}_{pop} T_{pop}}{1+\overline{F_G}}. \tag{2.78}$$

Here line 2.78 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 $\tilde{F}_{Gr}$.

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

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

$$\hat{\beta}'_{\ell} \mid G_{\ell}, u, e = \beta_{\ell} + \frac{u_{-\ell}^{\top} \mathbf{P} G_{\ell}}{G_{\ell}^{\top} \mathbf{P} G_{\ell}} + \frac{e^{\top} \mathbf{P} G_{\ell}}{G_{\ell}^{\top} \mathbf{P} G_{\ell}} \tag{2.79}$$

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

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

$$Z^\top \mid \mathbf{G}, u, e = \sum_{\ell=1}^{S} \hat{\beta}'_\ell X_\ell^\top \tag{2.80}$$

$$= \sum_{\ell=1}^{S} \beta_\ell X_\ell + \sum_{\ell=1}^{S} \frac{u^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} X_\ell^\top + \sum_{\ell=1}^{S} \frac{e^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} X_\ell^\top \tag{2.81}$$

so the expected polygenic scores can be written as

$$\mathbb{E}\left[Z^\top\right] = \sum_{\ell=1}^{S} \mathbb{E}\left[\beta_\ell X_\ell\right] + \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{u^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} X_\ell^\top\right] + \sum_{\ell=1}^{S} \mathbb{E}\left[\frac{e^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} X_\ell^\top\right]. \tag{2.82}$$

Following the same steps as above (see 2.54, 2.60, and 2.65), we take the expectation of each of the above terms separately. The first term is again the expected "true" polygenic scores ($\sum_{\ell=1}^{S} \mathbb{E}\left[\beta_\ell X_\ell\right] = \sum_{\ell=1}^{S} \beta_\ell X_{\ell,S} = \mathbb{E}[Z]^\top$). Then we consider the association between the environment and the residual genotypes,

$$\mathbb{E}\left[\frac{e^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} X_\ell^\top\right] \approx \frac{c^\top}{M} \mathbf{P} \mathbb{E}\left[\frac{G_{\ell,D} X_{\ell,D}^\top}{(G_{\ell,D}+G_{\ell,B})^\top \mathbf{P} (G_{\ell,D}+G_{\ell,B})/M}\right] \tag{2.83}$$

$$\approx \frac{c^\top}{M} \mathbf{P} \tilde{\mathbf{F}}_{GX} \tag{2.84}$$

$$\approx \frac{c^\top}{M} \left(\tilde{\mathbf{F}}_{GX} - \frac{1}{\|\tilde{F}_{Gr}\|} \tilde{F}_{Gr} \tilde{F}_{Gr}^\top \tilde{\mathbf{F}}_{GX}\right) \tag{2.85}$$

$$\approx \frac{c^\top}{M} \tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \tag{2.86}$$

where we make an additional approximation by ignoring the $\mathbf{P}$ in the denominator, which is reasonable so long as $\tilde{F}_{Gr}$ explains only a small fraction of the variance in the GWAS panel genotypes. Summing of the S causal sites in the polygenic score,

$$\sum_{\ell=1}^{S} \mathbb{E}\left[\frac{e^\top \mathbf{P} G_\ell}{G_\ell^\top \mathbf{P} G_\ell} X_\ell\right] \approx \frac{S}{M} c^\top \tilde{\mathbf{F}}_{GX}^{\perp \tilde{F}_{Gr}} \tag{2.87}$$

Turning to the contribution to the polygenic score arising from the association between focal site and the residual genetic background, the expectation is

$$\mathbb{E}\left[\frac{u_{-\ell}^{\top}\mathbf{P}G_{\ell}}{G_{\ell}^{\top}\mathbf{P}G_{\ell}}X_{\ell}^{\top}\right] \approx \frac{u_{S,-\ell}^{\top}}{M}\mathbf{P}\mathbb{E}\left[\frac{G_{\ell,D}X_{\ell,D}^{\top}}{(G_{\ell,D}+G_{\ell,B})^{\top}\mathbf{P}(G_{\ell,D}+G_{\ell,B})/M}\right] \tag{2.88}$$

$$\approx \frac{u_{S,-\ell}^{\top}}{M}\left(\tilde{\mathbf{F}}_{GX} - \frac{1}{\|\tilde{F}_{Gr}\|}\tilde{F}_{Gr}\tilde{F}_{Gr}^{\top}\tilde{\mathbf{F}}_{GX}\right) \tag{2.89}$$

$$\approx \frac{u_{S,-\ell}^{\top}}{M}\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}} \tag{2.90}$$

where we again make the approximation of ignoring $\mathbf{P}$ in the denominator. Summing across the S sites,

$$\sum_{\ell=1}^{S}\mathbb{E}\left[\frac{u^{\top}\mathbf{P}G_{\ell}}{G_{\ell}^{\top}\mathbf{P}G_{\ell}}X_{\ell}\right] \approx \frac{1}{M}\sum_{\ell=1}^{S}u_{S,-\ell}^{\top}\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}} \tag{2.91}$$

$$\approx \frac{S}{M}u_{S}^{\top}\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}} \tag{2.92}$$

Putting the true polygenic scores and the residual genetic and environmental contributions together, the expected polygenic scores after controlling for $\tilde{F}_{Gr}$ are approximately,

$$\mathbb{E}\left[\hat{Z}\right] \approx \mathbb{E}\left[Z\right] + \frac{S}{M}\left(\mu_{S}^{\top} + c^{\top}\right)\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}}. \tag{2.93}$$

2.8.6 Polygenic score association test bias controlling for $\tilde{\mathbf{F}}_{Gr}$

Following from the distribution of polygenic scores after controlling for $\tilde{\mathbf{F}}_{Gr}$, the expected association test statistic is,

$$\mathbb{E}\left[\hat{q}\right] \approx \mathbb{E}\left[q\right] + \frac{S}{M}\left(\mu_S^\top + c^\top\right)\tilde{\mathbf{F}}_{GX}^{\perp\tilde{F}_{Gr}}T \tag{2.94}$$

$$\approx \mathbb{E}\left[q\right] + \frac{S}{M}\left(\mu_S^\top + c^\top\right)\left(\tilde{\mathbf{F}}_{GX} - \frac{1}{\|\tilde{F}_{Gr}\|}\tilde{F}_{Gr}\tilde{F}_{Gr}^\top\tilde{\mathbf{F}}_{GX}\right)T \tag{2.95}$$

$$\approx \mathbb{E}\left[q\right] + \frac{S}{M}\left(\mu_S^\top + c^\top\right)\left(\tilde{F}_{Gr} - \frac{1}{\|\tilde{F}_{Gr}\|}\tilde{F}_{Gr}\tilde{F}_{Gr}^\top\tilde{F}_{Gr}\right) \tag{2.96}$$

$$\approx \mathbb{E}\left[q\right] + \frac{S}{M}\left(\mu_S^\top + c^\top\right)\left(\tilde{F}_{Gr} - \tilde{F}_{Gr}\right) \tag{2.97}$$

$$\approx \mathbb{E}\left[q\right] \tag{2.98}$$

such that the expected bias is

$$\mathbb{E}\left[\hat{q} - q\right] \approx 0. \tag{2.99}$$

Here we highlight a reasonable assumption we make regarding our ability to detect true signal when controlling for $\tilde{F}_{Gr}$. As discussed in the main text, including $\tilde{F}_{Gr}$ as a covariate removes any correlation between $\hat{\beta}$ and r under the null hypothesis where $\eta = 0$. Notably, under the alternative hypothesis, our ability to detect signal is preserved as we use variation along other axes in the GWAS panel to estimate effect sizes which in turn can still be correlated with r_S in the test panel. We therefore rely on the assumption that for site ℓ there is enough remaining variation in G_ℓ after regressing out $\tilde{F}_{Gr}$ to estimate $\hat{\beta}_\ell$. If all variation in G_ℓ lies along $\tilde{F}_{Gr}$, $y^\top\mathbf{P}G_\ell \approx 0$ and we will be unable to estimate $\hat{\beta}_\ell$. For example, if $\tilde{F}_{Gr}$ represented individuals on opposite sides of a population split this would only apply to sites with fixed differences. This is unlikely to be a concern in practice as F_{ST} is low in human populations[94] and most variants will not be perfectly correlated with

a single axis of structure. Therefore, it is reasonable to assume that it will still be possible to accurately estimate β_ℓ, albeit with slightly larger standard error.

2.8.7 Modeling the correlation structure among test panel individuals

In Section 2.4, we test for the association between polygenic scores and the test vector using the linear model,

$$Z = qT + \varepsilon \tag{2.100}$$

where ε is i.i.d Normal across individuals. If T is scaled to have a variance of 1, the slope is given by,

$$q = \frac{1}{N} Z^\top T \tag{2.101}$$

which we use as our test statistic of interest.

However, a more powerful test is available by modeling the correlation structure among individuals in the test panel. Under our genotypic model the expected pattern of covariance of genotypes at site among individuals in the test panel is,

$$Var\left(X_\ell\right) = \mathbb{E}[X_\ell X_\ell^\top] = 4a_\ell\left(1-a_\ell\right)\left(\mathbf{F}^*_{XX}\right), \tag{2.102}$$

where $\mathbf{F}^*_{XX} = \mathbf{F}_{XX} + \mathbf{B}_X$. $\mathbf{F}_{XX}$ contains the within panel relatedness coefficients and $\mathbf{B}_X$ is a diagonal matrix with entries specifying the amount of additional variance for each individual due to sampling.

Now, because the polygenic scores are sums across a large number of sites, under the null model the distribution of standardized polygenic scores is approximately multivariate

Normal

$$\frac{Z}{\sqrt{2V_A}} \sim MVN\left(0, \mathbf{F}^*_{XX}\right) \tag{2.103}$$

where $V_A = 2\sum_\ell^S \beta_\ell^2 a_\ell(1-a_\ell)$ is the additive genetic variance of the polygenic scores in the original ancestral population.

To test for evidence of an association, we compare this null model to an alternative in which polygenic scores are associated with T,

$$\frac{Z}{\sqrt{2V_A}} \sim MVN\left(T\gamma, \mathbf{F}^*_{XX}\right), \tag{2.104}$$

where γ is the slope. Here we assume that T is mean centered and scaled such that $\mathbf{F}^{*-\frac{1}{2}}_{XX}T$ has a variance of one. With this scaling, the generalized least squares estimate of γ is given by $\hat{\gamma} = {}^{T^\top}\mathbf{F}^{*-1}_{XX}Z/(N\sqrt{2V_A})$, where the $\mathbf{F}^{*-1}_{XX}$ term accounts for the evolutionary non-independence of the individuals in the panel. Under the null hypothesis, $\hat{\gamma} \sim N\left(0, \frac{1}{N}\right)$.

To remove this dependence on the sample size, we let the test statistic be $q_x = \sqrt{N}\hat{\gamma}$, so that

$$q_x = \frac{Z^\top \mathbf{F}^{*-1}_{XX}T}{\sqrt{2NV_A}} \tag{2.105}$$

$$= \frac{1}{\sqrt{2NV_A}}\sum_{\ell=1}^{S}\beta_\ell X_\ell^\top \mathbf{F}^{*-1}_{XX}T \tag{2.106}$$

$$= \frac{1}{\sqrt{2NV_A}}\beta r_x. \tag{2.107}$$

Under the null $q_x \sim N(0,1)$. This test statistic is related to Berg and Coop's Q_X statistic[59,71] in a straight-forward way. For a test of selection along a single axis of ancestry variation, $Q_X = q_x^2$. More generally, the Q_X statistic combines multiple tests along different axes of ancestry variation, accounting for non-independence among axes and providing the

appropriate multiple testing penalty.

As discussed above, q_x is proportional to the generalized least squares estimate of γ whereas q in the main text is proportional to the ordinary least squares estimate. Therefore, q_x accounts for covariance and heteroscedasticity in test panel genotypes caused by within test panel population structure. This results in a more powerful test, in precisely the same way that linear mixed models increase the power of association tests in GWAS by accounting for covariance in the phenotypes due to genetic relatedness.

To apply our correction procedure while accounting for the non-independence among test panel individuals, we can compute $r_{x,\ell} = X_\ell^\top \hat{\mathbf{F}}_{XX}^{*-1} T$ and plug r_x into Equation 3.2 in the main text instead of r_ℓ. This version of $\hat{F}_{Gr}$ can again be included as a covariate in the GWAS and those corrected effect sizes, in conjunction with r_x can then be used to compute q_x. Overall, we would expect this procedure to result in slightly stronger protection against stratification bias, as the $r_{x,\ell}$ will have a slightly lower variance than the r_ℓ, and thus lead to a more accurate estimated $\hat{F}_{Gr}$, though we have not systematically investigated this.

CHAPTER 3

CONTROLLING FOR STRATIFICATION BIAS IN PRACTICE

3.1 Introduction

One promise of polygenic scores is that they allow for phenotypic prediction from genotype data alone. To fulfill this promise, we expect consistent predictions that are robust to choice of effect size estimate. As discussed in Chapter 2, bias in polygenic score predictions is a function of the relationship between GWAS and target samples. If confounding is well controlled for in the GWAS, for example by PCs or LMMs (see Chapter 1), then we do not expect relationships between polygenic scores and ancestry gradients in the test panel to change as a function of the ancestry composition of the GWAS panel.

Empirically, two recent studies, Refoyo-Martínez et al. (2021)[87] and Kerminen et al. (2019)[53], found that patterns in polygenic scores differences were not robust to choice of effect estimates. Kerminen et al. tested three different height polygenic scores, each with effect sizes estimated in a GWAS panel with a different genetic distance from the target panel of Finnish individuals, and predicted three significantly different values for the difference in height between the eastern and western Finish subgroups. Additionally, they observed geographic structure in polygenic scores for five other traits, two of which predicted implausibly large differences in phenotypes between eastern and western subgroups. Similarly, Refoyo-Martínez et al. used effect sizes from seven different height GWASs to compute mean polygenic scores for populations in the 1kGP. They observed not only significant variation in the magnitude of polygenic score differences between populations but also variation in the direction of differences. Because both studies use the same set of test panel individuals, the observed inconsistencies must come from the effect size estimates.

Ancestry stratification is not the only explanation for the above observations. For example, the most extreme geographic clustering, usually seen in smaller meta-analysis GWASs,

might reflect true biological differences and larger biobank GWASs are over correcting for ancestry and regressing out real signals. Although if this were the case, we would expect to see consistent predictions across different meta-analyses, which is not the case[87]. Alternatively, the difference in signals could be driven by population specific variants that are or are not included in GWASs done in different ancestries. Population specific effect sizes (i.e $G \times G$ or $G \times E$ in the GWAS panel) could also generate a similar pattern. While none of these explanations can be explicitly ruled out, most complex traits are highly polygenic and the difference in signal is unlikely to be driven by a small number of population specific variants. Moreover, there is little evidence for widespread heterogeneity in effect sizes across ancestry backgrounds[95,96].

There is evidence that ancestry stratification is partially responsible for the observed inconsistencies. Stratification bias is a phenomenon that affects effect size estimates genome-wide, not just trait-associated ones. Kerminen et al. found that constructing polygenic scores from non-associated SNPs ($p > 0.5$) also produced an east/west difference in Finland, a pattern that is not expected if the signal were driven by real biological differences, but is consistent with uncorrected ancestry stratification. Refoyo-Martínez et al. also showed that effect sizes from different cohorts are differentially correlated with the top 20 PC loadings from the test panel. Sohail et al. (2019)[72] makes the same observation comparing effect sizes estimated in GIANT to those estimated in the UKBB. Both studies observe larger correlations with effect sizes from smaller, more diverse meta-analyses (GIANT and PAGE) than they do in larger, more homogeneous biobanks (BBJ and UKBB). Again, these observations are consistent with cohort-specific uncorrected stratification bias generating relationships between top PCs and *all* effect size estimates.

Outside of the aforementioned results, other studies have found that polygenic scores exhibit geographic clustering[49,50,51,97] even after strict control for stratification. Again, while these observations could be due to real biological differences across the geographic region,

it is also expected if effect sizes suffer from residual ancestry-based stratification. In the Mendelian randomization literature, several studies have found evidence for residual stratification in effect size estimates[98,38,40] and as discussed in Chapter 1, effect sizes estimated using family data are typically smaller and result in lower heritability estimates[40,41,42], something we expect if population effect sizes suffer from uncorrected stratification bias.

In this Chapter we build on results from Chapter 2 to explore strategies for controlling for stratification bias in polygenic score association tests in practice. In the previous Chapter, we used a model-based approach to identify a single axis of ancestry variation in the GWAS panels, $\tilde{F}_{Gr}$, that needs to be controlled for to guarantee an unbiased test between polygenic scores built in the test panel, $\hat{Z}$, and an ancestry gradient of interest, T. In this Chapter, we first explore two different approaches to controlling for $\tilde{F}_{Gr}$, estimating the axis directly using the test panel contrasts or capturing it using sample principal components of the GWAS panel. We then apply both approaches to a range of simulated scenarios. We find that the utility of the direct approach depends on the error in our estimator, a function of the overlap in structure between panels. On the other hand, the error in J sample PCs depends only on the pattern of structure in the GWAS panel. In simulations where there is no true genetic signal and complete overlap in structure, the direct estimator is able to equal or outperform sample PCs. When the variation explained by the shared axis of structure decreases, the direct estimator is very noisy and struggles to protect the test.

In real data, where we are trying to both reduce confounding and maximize power of the polygenic score, we find sample PCs to always be the preferred approach. Because our direct estimator only controls for structure along a single axis of variation in the GWAS panel, confounding along other axes, while not contributing to bias in $\hat{q}$, adds additional noise and results in the ascertainment of SNPs that are less tightly coupled with causal variants. The inclusion of top sample PCs controls for stratification along multiple axes, resulting in more powerful polygenic scores. However, as previously discussed, one downside

of the PCA approach is that there is no way to test if the number of PCs included was sufficient to protect a polygenic score association test from stratification bias. Therefore, we combine both approaches and use our direct estimator as a way to check what fraction of the variation in ancestry along the axis identified by the direct estimator is captured by J sample PCs. We then apply this approach to 30 different polygenic association tests across four different GWAS panels and identify signals of polygenic score differentiation that are well protected from confounding.

3.2 $\hat{F}_{Gr}$ and sample PCs as estimators of $\tilde{F}_{Gr}$

In the previous Chapter we defined the conditions under which including $\tilde{F}_{Gr}$ or the top J population PCs as fixed covariates removes stratification bias and leads to an unbiased association test. However, both $\tilde{F}_{Gr}$ and U are theoretical quantities that depend on the population model, which we do not observe in practice. Instead, we must estimate these quantities, $\hat{F}_{Gr}$ and $\hat{U}$, with error, from sample genotype data.

3.2.1 Sample principal components

The sample PCs, $\hat{U}$, can be computed by taking the eigen decomposition of the empirical genetic covariance matrix, or the singular value decomposition of the genotype matrix. Existing results from random matrix theory allow us to obtain some understanding of the accuracy of $\hat{U}$ as an estimator of U. Specifically, in many GWASs the number of individuals in the GWAS panel, M, is roughly on the same order as the number of SNPs, L. In this setting, the accuracy of the sample eigenvector $\hat{U}_j$ depends on the corresponding population eigenvalue (λ_j) and the ratio of the number of individuals to the number of SNPs in the GWAS panel (M/L). As shown first by Patterson et al. (2006) in the context of genetics[99] (see also Baik et al. (2005)[100]), PCA exhibits a phase change behavior in which a given sample PC is only expected to align with the population PC if the corresponding population

eigenvalue is greater than a threshold value of $1+\sqrt{\frac{M}{L}}$. Below this threshold, the sample PC is orthogonal to the population PC. It is worth noting that increasing the number of individuals in the GWAS panel, M, also changes the population eigenvalue such that population PCs are generally more detectable with larger sample sizes.

However, even when the corresponding eigenvalue exceeds this threshold, the angle between the sample PC and the population PC may still be substantially less than one, particularly if the relevant eigenvalue does not far exceed the detection threshold [101,102]. Specifically, the squared correlation between the population PC and the sample PC is approximately

$$\left(U_j^\top \hat{U}_j\right)^2 \approx \begin{cases} \frac{1-\frac{M}{L}/(\lambda_j-1)^2}{1+\frac{M}{L}/(\lambda_j-1)}, & \lambda_j > 1+\sqrt{\frac{M}{L}} \\ 0, & \lambda_j \in [1, 1+\sqrt{\frac{M}{L}}] \end{cases} \tag{3.1}$$

(see Johnstone and Paul (2018) [101] for details). Thus even in cases where $\tilde{F}_{Gr}$ is fully captured by the top J population PCs, either of these two related phenomena may make it difficult to accurately approximate $\tilde{F}_{Gr}$ as a linear combination of the top J sample PCs, leading to a failure to fully account for stratification bias in polygenic score association tests.

3.2.2 Estimating $\tilde{F}_{Gr}$ directly using test panel genotypes

Given this limitation of PCA, it's natural to ask whether other estimators of $\tilde{F}_{Gr}$ might perform better. One choice, suggested by our theoretical results, is a direct estimator that utilizes the relevant test panel genotype contrasts. Given the test panel genotype contrasts (r_ℓ) and GWAS panel genotypes (G_ℓ), we can obtain a direct estimator of $\tilde{F}_{Gr}$ as

$$\hat{F}_{Gr} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{G_\ell r_\ell}{G_\ell^\top G_\ell / M}. \tag{3.2}$$

Then, if $\hat{F}_{Gr}$ is a sufficiently accurate estimator of $\tilde{F}_{Gr}$, we should be able to render

a given polygenic score association test unbiased by estimating marginal effects under the model

$$y = G_\ell \beta_\ell + \hat{F}_{Gr}\omega + \varepsilon, \tag{3.3}$$

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

We can expect this method to be successful when the variance of the error component of $\hat{F}_{Gr}$ is small relative to the variance of the entries of $\tilde{F}_{Gr\cdot}$. The variance of $\tilde{F}_{Gr}$ will be greater when the amount of overlap in population structure between the two panels along this specific axis is greater. We can think about the variance of the error component in terms of a linear model that tries to predict the GWAS panel genotypes using the test panel genotype contrasts. If we write $\tilde{G}_{i\cdot}$ to denote the vectors of genotypes for GWAS individual i and $\tilde{r}$ for the test panel genotype contrasts, each standardized by the variance in the GWAS panel, then we can fit the linear model

$$\tilde{G}_{i\cdot} = \tilde{r}\tilde{F}_{Gr,i} + e. \tag{3.4}$$

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

This approach proposes 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. One concern is that this procedure might remove the signal we are trying to detect. In supplemental Section 3.8.1 we show that while this is true for naive applications, the effect will be small so long as the number of SNPs used to compute the correction is large relative to the number included in the polygenic score

(i.e $S \ll L$). Notably, controlling for sample PCs of the GWAS panel genotype matrix will induce a similar effect if the sample PCs capture $\tilde{F}_{Gr}$. We confirm via simulations (see Figure 3.18) that downward bias in $\hat{q}$ when including $\hat{F}_{Gr}$ or sample PCs is minimal when $S \ll L$. 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[103,23] 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.

3.3 Applications

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

3.3.1 Toy model

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

We first consider a toy model with four populations (labeled A, B, C and D), which are related to one another by an evenly balanced population phylogeny (Figure 3.1). 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 3.1A), and another where A and B are sister (Figure 3.1C).

For simplicity, we consider a purely environmental phenotype (i.e. $h^2 = 0$) with a difference in mean between populations A and B is equal to Δ_{AB} (Figure 3.1B). Following

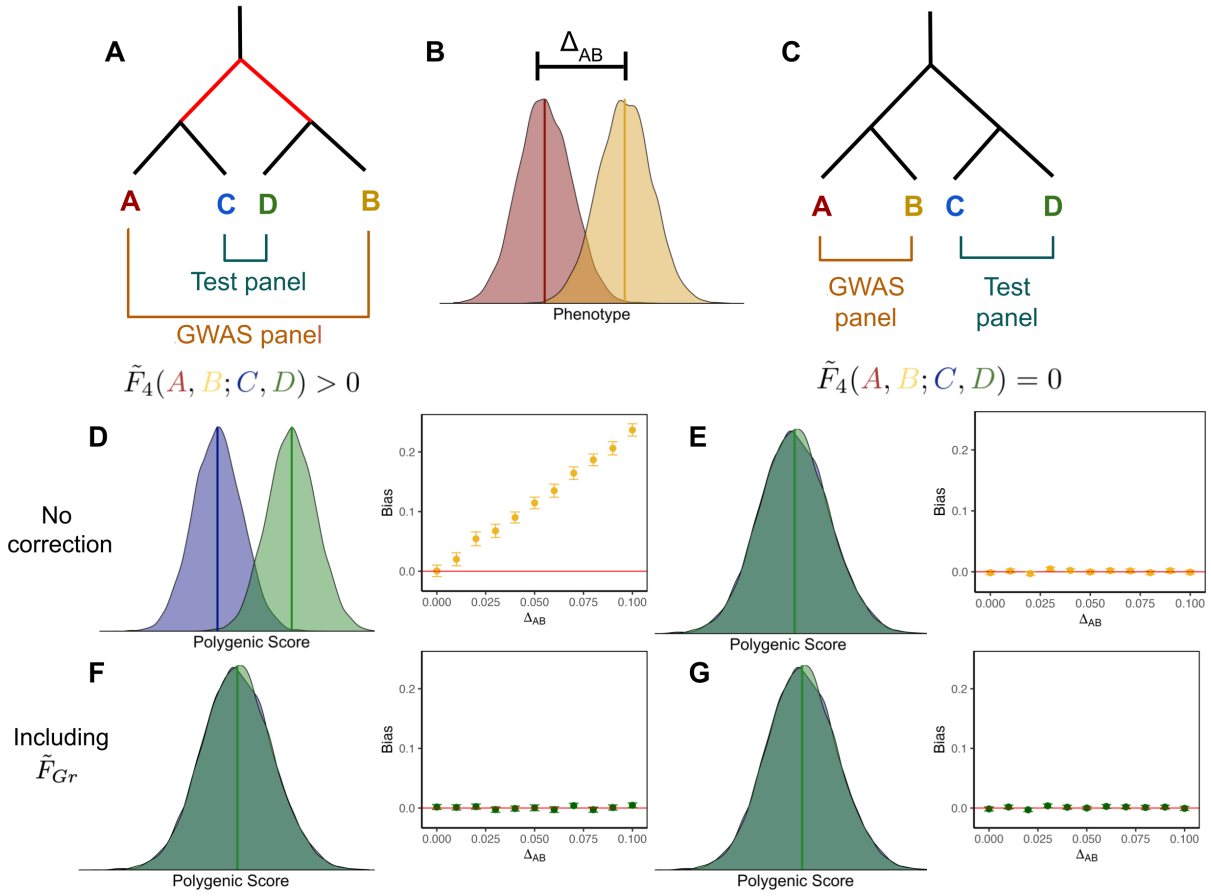

Figure 3.1: **Schematic of two different panel configurations. The effect of stratification depends on the overlapping structure between the GWAS and test panels.** (A, C) Two different topologies used to create the GWAS and test panels. (B) Stratification was modeled in the GWAS panel by drawing an 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 bias in $\hat{q}$ increases with the magnitude of stratification in the GWAS. (E) However, when there is no overlapping structure between panels, there is no expected difference in mean polygenic scores between C and D and $\hat{q}$ remains unbiased regardless of the magnitude of stratification. (F, G) Including $\tilde{F}_{Gr}$ as a covariate in the GWAS controls for stratification, eliminating bias in $\hat{q}$ regardless of Δ_{AB} or the overlapping structure between GWAS and test panels.

from Equation 2.7, the marginal effect size estimate for site ℓ is

$$\begin{aligned}\hat{\beta}_\ell \mid G_\ell, e &= \frac{G_\ell^\top e}{G_\ell^\top G_\ell} \\ &= \frac{1}{2}\frac{\Delta_{AB}\left(\hat{p}_{A,\ell} - \hat{p}_{B,\ell}\right)}{G_\ell^\top G_\ell / M} + \frac{G_\ell^\top \varepsilon}{G_\ell^\top G_\ell}\end{aligned} \tag{3.5}$$

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 Robinson et al. (2015)[67]).

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

$$\begin{aligned}\mathbb{E}\left[\hat{q} - q\right] &= \Delta_{AB}\sum_{\ell=1}^{S}\mathbb{E}\left[\frac{(\hat{p}_{A,\ell} - \hat{p}_{B,\ell})(\hat{p}_{C,\ell} - \hat{p}_{D,\ell})}{G_\ell^\top G_\ell / M}\right] \\ &= \Delta_{AB} S \tilde{F}_4(A, B; C, D)\end{aligned} \tag{3.6}$$

where $\tilde{F}_4(A, B; C, D)$ is a version of Patterson's F_4 statistic[104,105], standardized by the genotypic variance in the GWAS panel, which measures the amount of genetic drift common to populations A and B that is also shared by populations C and D. Writing the bias in terms of this modified F_4 statistic helps illustrate the role of cross panel population structure in driving stratification bias in polygenic scores. The effect estimate at site ℓ is a linear function of $\hat{p}_{A,\ell} - \hat{p}_{B,\ell}$, so the test will be biased if $\hat{p}_{A,\ell} - \hat{p}_{B,\ell}$ is correlated with $\hat{p}_{C,\ell} - \hat{p}_{D,\ell}$. This is true for the demographic model in Figure 3.1A, 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 3.1C, where there is no shared internal branch and $\tilde{F}_4(A, B; C, D) = 0$.

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

panel. The results are consistent with our theoretical expectations: the test statistic is biased for the topology with $\tilde{F}_4(A, B; C, D) > 0$ (Figure 3.1D), but unbiased when $\tilde{F}_4(A, B; C, D) = 0$ (Figure 3.1E).

Given the population model, $\tilde{\mathbf{F}}_{XG} = \mathbf{0}$ for the unconfounded topology, making $\tilde{F}_{Gr}$ a vector of zeros. Therefore, re-running the GWAS including $\tilde{F}_{Gr}$ does not change the outcome of the already unbiased test (Figure 3.1G). 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 and including it as a covariate in the GWAS eliminates bias for the confounded topology (Figure 3.1F).

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

As we outlined above, in practice, $\tilde{F}_{Gr}$ cannot be observed directly, and must be estimated with error from the data. To illustrate the impact of this estimation error on the performance of both estimators in a simple, well understood case, we performed simulations using three different versions of our toy model in which we vary the amount of overlap in population structure between the test and GWAS panels. Specifically, given that $\tilde{F}_{Gr}$ is known in this toy model, we can compute the error in either estimator as one minus the squared correlation between $\tilde{F}_{Gr}$ and the corresponding estimator. We take all of these vectors to be standardized, so this is simply

$$\text{Error} = 1 - \left(\hat{x}^\top \tilde{F}_{Gr}\right)^2 \tag{3.7}$$

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

For each simulation, we estimated $\hat{F}_{Gr}$ as in Equation 3.2, using L genome-wide SNPs with a frequency of greater than 1% in both the test and GWAS panels. For PCA, we

computed sample PCs via singular value decomposition of the genotype matrix using the same set of SNPs that were used to compute $\hat{F}_{Gr}$, and we then take $\hat{U}_1$ (i.e. the first sample PC) as the PCA based estimator of $\tilde{F}_{Gr}$[82]. In all of these simulations, we hold the GWAS and test panel sample sizes constant at $N, M = 1,000$ and varied the number of SNPs (L) as a way to vary the accuracy of the estimators. We simulated 100 replicates for each topology, and plot the resulting averages across these replicates in Figure 3.2.

First, we simulated a scenario of complete overlap, in which there is a single population split and individuals in both the GWAS and test panels are independently drawn as 50:50 mixtures of the two population on either side of the split (Figure 3.2A). When the GWAS sample size (M) is on the same order as the number of SNPs (L), the direct estimator ($\hat{F}_{Gr}$) has a smaller error than the first PC ($\hat{U}_1$) (Figure 3.2B), and as a consequence reduces the bias by a larger amount (Figure 3.2C). Intuitively, the direct estimator singles out the relevant axis of population structure because we have identified it ourselves in the test panel, whereas PCA has to find this axis "on its own" in the high dimension GWAS panel genotype data, and thus pays an additional cost. In contrast, when $M \ll L$, PCA no longer has to pay this additional cost, and its performance improves to match that of the direct estimator.

We next simulated under the same toy model of partial overlap in population structure between test and GWAS panels that we considered above (Figure 3.2D). This results in an increase in the error of the direct estimator relative to the complete overlap case because the genotype contrasts measured in the test panel are less informative about the relevant axis of structure in the GWAS panel. In contrast, the error in $\hat{U}_1$ is unchanged, as the amount of structure in the GWAS panel is the same as in Figure 3.2A. Notably, in this case the direct estimator still outperforms PCA when $M/L \approx 1$, but PCA performs better as M/L decreases.

Finally, in Figure 3.2G we reduced the overlap in population structure even further, which leads PCA to uniformly outperform the direct estimator. Intuitively, because the overlap in population structure is so small, the direct estimator requires a very large number of SNPs to

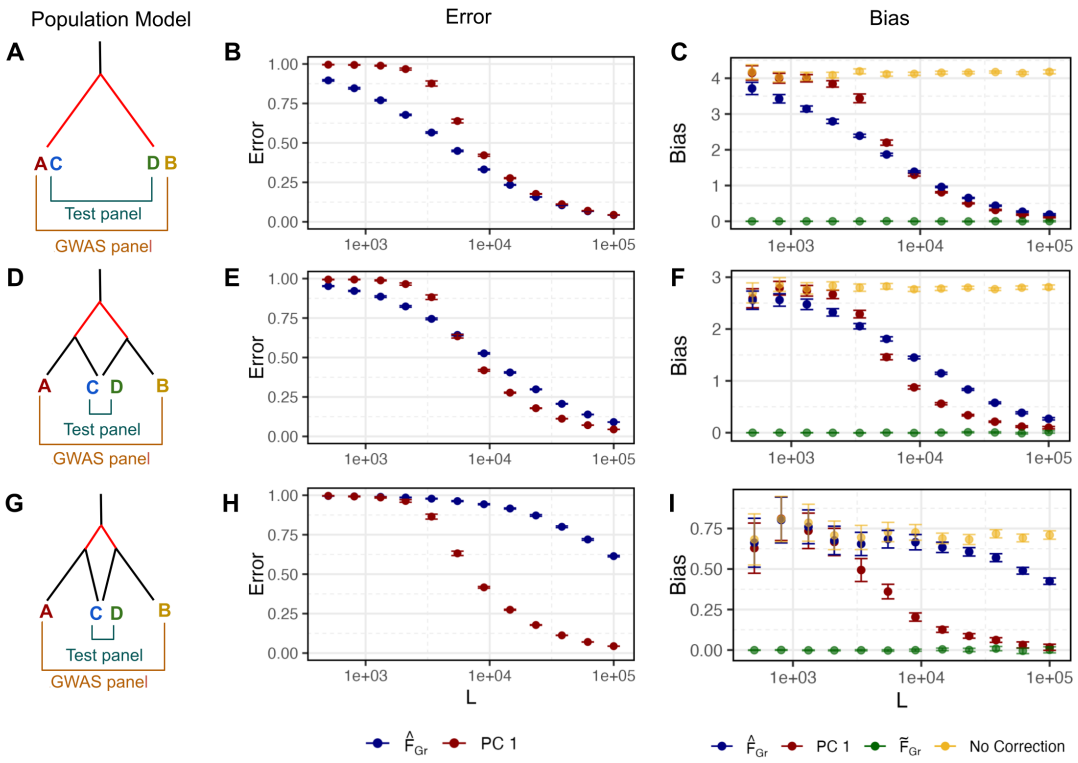

Figure 3.2: **Error in estimators of $\tilde{F}_{Gr}$ depends on the number of SNPs used to compute them.**

(A) We simulated a population model with a single split and sampled an equal proportion of individuals from each population to make a GWAS and test panel. (D,C) Here we simulated population models with two splits and sampled individuals in the overlapping structure configuration. (B, E, H) As $\tilde{F}_{Gr}$ is known for these population models, we computed the error in $\hat{U}_1$ and $\hat{F}_{Gr}$ as estimators of $\tilde{F}_{Gr}$ using eq. 3.7. For both estimators, error decreased as the number of SNPs increased. We hold the number of GWAS panel individuals constant at $M = 1,000$ so as L increases the ratio of $\frac{M}{L}$ decreases. The error in $\hat{U}_1$ does not depend on the population model as the depth of the deepest split is constant across models. Error in $\hat{F}_{Gr}$ increases as overlap between panels decreases. (C, F, I) Bias in $\hat{q}$ computed from using the estimators as covariates in the GWAS follows from the error in the estimators themselves.

produce an accurate estimate. We also note that in general across all of these simulations, while the magnitude of the reduction in bias closely tracks the error in the estimator of population structure, the reduction is slightly larger than expected for $\hat{U}_1$ (Figure 3.19).

3.3.2 Grid simulations

To further explore stratification bias 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)[47]. We sampled an equal number of individuals per deme to comprise both the GWAS and test panels, with total sample sizes $N, M = 1,440$. We then simulated several different distributions of purely environmental phenotypes across the GWAS panel individuals. We considered three different scenarios for the distribution of phenotypes. For each scenario, we estimated effect sizes, ascertained associated sites, and tested for an association between polygenic score and latitude, longitude, or membership in the single confounded deme, depending on the example. In these simulations $\tilde{F}_{Gr}$ is unknown and so we compared $\hat{F}_{Gr}$ and the top 10 sample PCs as estimators of $\tilde{F}_{Gr}$, using the same set of $L = 20,000$ SNPs that are found at a frequency greater than 1% in both panels for both estimators.

For the first example, the confounder, c, is a linear function of an individual's position on the latitudinal axis (Figure 3.3A). When we estimated effect sizes with no correction for population structure, the spatial distribution of the resulting polygenic scores reflected the distribution of the environmental confounder. Consequently, an association test using latitude as the test vector is biased. However, including $\hat{F}_{Gr}$ or the top 10 sample PCs as covariates in the GWAS model is sufficient to ensure that effect sizes that are unbiased with respect to the latitudinal genotype contrasts in the test panel, so the resulting association test is unbiased.

In the second example, we simulated confounding along the diagonal, resulting in uncor-

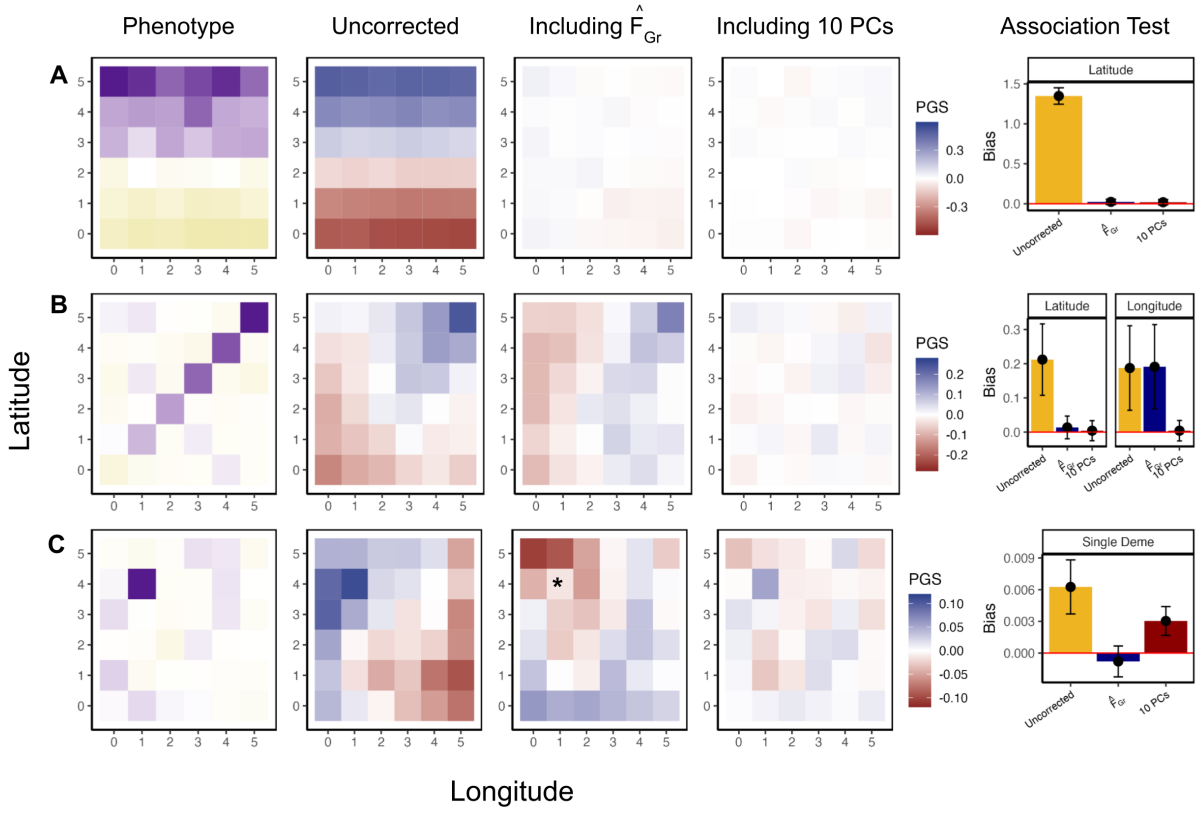

Figure 3.3: **Stratification bias in more complex demographic scenarios.** GWAS and test panel individuals were simulated using a stepping-stone model with continuous migration. In the GWAS panel, the phenotype is non-heritable and stratified along either latitude (A), the diagonal (B), or in a single deme (C). When effect sizes were 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) which is also reflected in the association test bias (fifth column). We also compare our approach to including the top 10 PCs (fourth column) which successfully protects the test in A and B but remains biased for C.

rected 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 3.3B). When we computed $\hat{F}_{Gr}$ 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 approach, 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 (when $\hat{F}_{Gr}$ is well estimated). Including 10 sample PCs protects both the latitudinal and longitudinal association tests.

In the third example, we simulated an increased environmental effect in a single deme, a scenario which induces a more complex spatial pattern in the uncorrected polygenic scores (Figure 3.3C), and which previous work has shown to be difficult to correct for with standard tools[106,47]. We then took the test vector to be an indicator for whether the test panel individuals were sampled from the deme with the environmental effect, or not, and computed $\hat{F}_{Gr}$ using these contrasts. In this scenario, including $\hat{F}_{Gr}$ as a covariate in the GWAS results in an unbiased test statistic. In contrast, the top ten sample PCs did not.

Quantifying error in population structure estimators

Next, we wanted to better understand the role error in our population structure estimators plays in these simulations. In contrast to the four population toy model, it is not straightforward to compute $\tilde{F}_{Gr}$ given our underlying demographic model, particularly for the case of testing a single deme against all others. As a result, we cannot directly measure the error in $\hat{F}_{Gr}$ or sample PCs as estimators of $\tilde{F}_{Gr}$. Instead we use the fact that under this demographic model individuals within a deme are exchangeable, and therefore have the same values of both $\tilde{F}_{Gr}$ and population PCs. This allows us to estimate the error in $\hat{F}_{Gr}$ by computing the fraction of the total variance in $\hat{F}_{Gr}$ that can be attributed to variance of individual values within demes and to variance of deme means across replicates (see 3.7.1).

For the PCs the relationship between the order of the underlying population PCs and the order of the sample PCs may differ across replicates due to the noisiness of the sample PCs, so it is not obvious how to compute the variance of the deme means across replicates. We therefore use only the within deme variances, so our estimates of the error for the PCs are technically estimates of a lower bound on the error (see 3.7.1). However, we note that for our estimation of the error in $\hat{F}_{Gr}$, we found that the variance within demes was by far the larger contributor, so we expect this to be a relatively tight bound. We then vary the number of SNPs used to compute our estimators of population structure from $L = 20,000$ down to $L = 2,000$, and observe how differences in the estimated error of our population structure estimators translate to differences in the amount of bias in the polygenic score association test statistic.

In Figure 3.3A and Figure 3.3B, $\tilde{F}_{Gr}$ corresponds to latitude, so we expect it to be captured by the top two population PCs[107]. For $L = 20,000$ (the number of SNPs used in Figure 3.3), we estimated the lower bound on the error in sample PCs 1 and 2 to be 0.011. Across the range of L values we tested, the estimated bound was no greater than 0.053 (Figure 3.4A) and including 10 PCs consistently removes bias in $\hat{q}$ (Figure 3.4B). Similarly, we estimated the error in $\hat{F}_{Gr}$ for latitude to be 0.012 when $L = 20,000$ with a maximum of 0.059 when $L = 2,000$. Although these estimates are nearly identical to the values we observe for the first two PCs, the bias in $\hat{q}$ is slightly higher (Figure 3.4B). We observed a similar result in the 4 population toy model (3.19), so this may be the same phenomenon, or it may be that PCs 3-10 are capturing some of the residual latitudinal signal that is not captured by the first two.

Next, we explored the role of error in our population structure estimators for the more difficult single deme test/confounder case (Figure 3.3C). We again computed the error in $\hat{F}_{Gr}$ as we vary L, with estimates ranging from 0.04 to 0.18 as L decreases (Figure 3.4A). For larger values of L, the error was small enough that confidence intervals on the bias

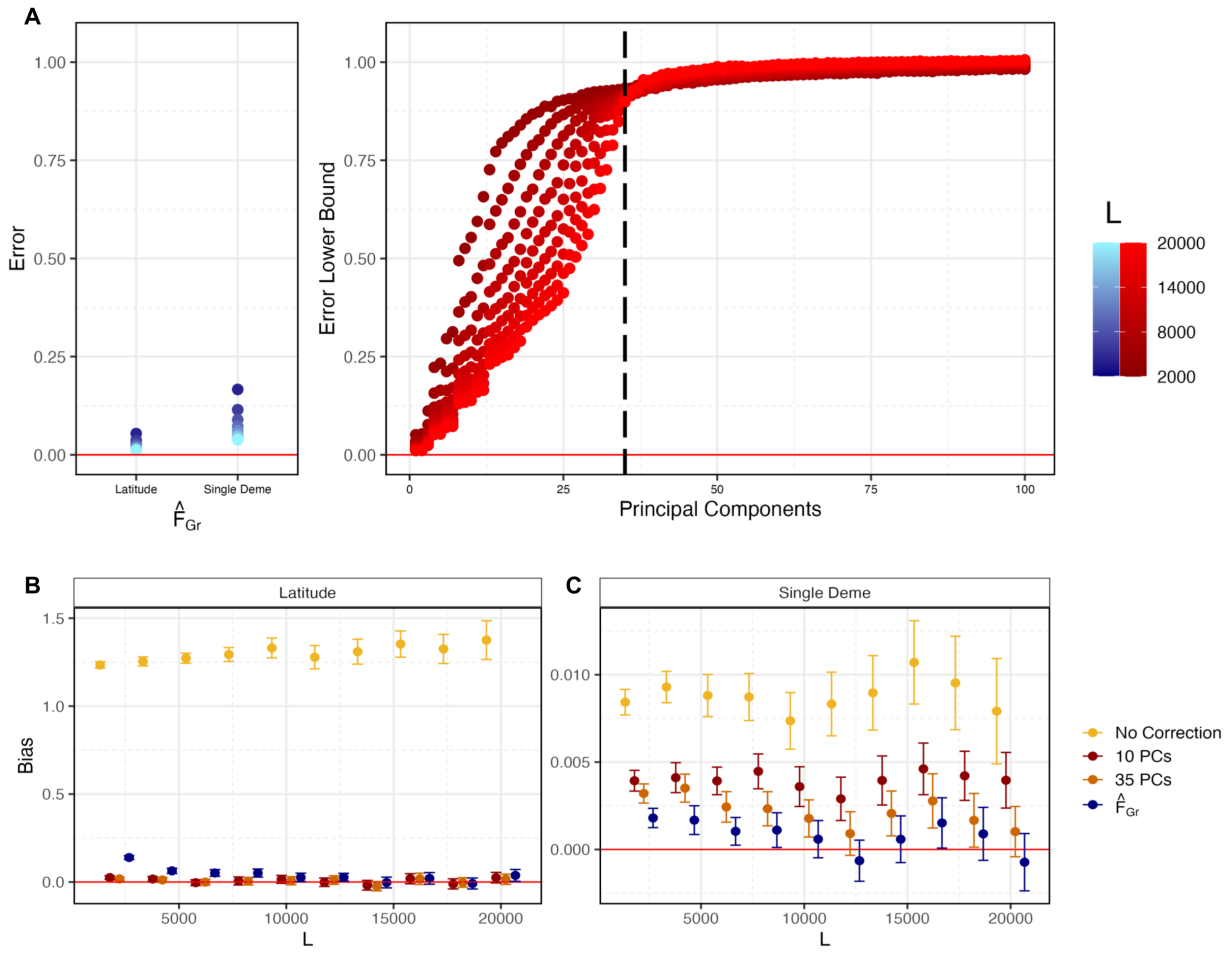

Figure 3.4: **Quantifying error in the estimates of $\hat{F}_{Gr}$ and sample PCs for the six-by-six stepping stone demographic model.**

(A) Given the stepping stone demographic model used in Figure 3.3, individuals within a deme are exchangeable and have the same $\tilde{F}_{Gr}$ and population PC value. Therefore, we used variation within demes to estimate the error in $\hat{F}_{Gr}$ and a lower bound for the error in sample PCs for different values of L (we hold $M = 1,400$). The dashed vertical line indicates PC 35, the last population PC we expect to capture real structure. (B) When latitude is the test vector, both sample PCs and $\hat{F}_{Gr}$ are well estimated and bias in $\hat{q}$ is reduced. (C) When a single deme indicator variable is the test vector, higher PCs are needed to capture $\tilde{F}_{Gr}$. These sample PCs are not well estimated and residual bias remains when 35 PCs are used for most values of L.

overlapped zero, but this was not true when we reduced L so that the error was larger (Figure 3.4C). Above, with $L = 20,000$, we found that 10 PCs were not sufficient to remove the bias. This could either be because $\tilde{F}_{Gr}$ is not captured by the top 10 population PCs *or* it could be that $\tilde{F}_{Gr}$ can be captured by 10 population PCs, but the sample PCs are too noisy as estimates of the population PCs. Given that there are 36 demes in our simulations and that individuals within demes are exchangeable, only the top 35 population PCs capture real population structure, while the rest correspond to sampling variance. As a result, if the sample PCs are sufficiently well estimated, then only 35 should be required to remove the bias. In practice, we find that using 35 PCs for larger values of L, the bias is closer to zero than it is with 10 PCs, but the confidence intervals still do not always overlap zero, and the bias is generally greater than it is when we use our direct estimator, $\hat{F}_{Gr}$ (Figure 3.4C). As expected, the performance with 35 sample PCs decreases further with an increase in the error, but is always intermediate between 10 PCs and $\hat{F}_{Gr}$. All of this is consistent with the observation that the error in the higher sample PCs (i.e. 11-35), is very high across the range of L values we explored (Figure 3.4A).

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

Finally, to the extent that the PCs did succeed in removing bias in our simulations, we wanted to understand whether it was because they successfully captured the confounder or because they captured the relevant axis of structure for the test (see Section 2.5.2). To this end, for each of the three grid scenarios in the $L = 20,000$ case, we computed the cumulative proportion of variance in the confounder, c, that could be explained by the first J sample PCs, for J up to 100 (Figure 3.5). We found that while the confounding axis was well captured by sample PCs 1 and 2 for latitude (Figure 3.5A), it was not well captured by the top 10, 35, or indeed 100 PCs, for the diagonal (Figure 3.5C) or single deme confounders (Figure 3.5E). In contrast, if we take our estimator, $\hat{F}_{Gr}$, as a proxy for $\tilde{F}_{Gr}$, we find that

the PCs explain a considerably higher fraction of the variance. For the first two cases, the test axis is latitude, so this is unsurprising. However, this is true even for the single deme case, and results from the fact that relatedness among adjacent demes leads to a smoothing effect (Figure 3.20), which makes $\tilde{F}_{Gr}$ easier for the PCs to capture.

3.3.3 Example data analysis in UK Biobank

As an example of an application to real data, we estimated $\hat{F}_{Gr}$ and quantified the error for two polygenic score association tests using data from the UK Biobank[43]. Specifically, we tested for an association between standing height and birth location in the UK along either the north-south or east-west axis. First, we sampled individuals to comprise both the test and GWAS panels. Given the subtlety of the structure along our chosen test axes, our goal in constructing the test panel was primarily to have a large enough sample size that the genotype contrasts along these axes would be well-estimated. To this end, we selected individuals who chose "`White British`" in the UK Biobank's ethnic identity questionnaire and cluster near the centroid in genetic PCA space (i.e they are members of the "White British ancestry subset" as defined by the UK Biobank[43]) to comprise the test panel ($N = 399,801$). We then used each individual's place of birth in the UK, recorded in terms of northings and eastings, to create our two test vectors. Our goal in constructing the GWAS panel was to have a complex but partial overlap in population structure between GWAS and test panels, as might be expected in practice if the GWAS and test panels are drawn from different sources. To this end, we selected individuals who self-identified as "`White`" but were not included in the "White British ancestry subset" to comprise the GWAS panel ($M = 50,136$). This set is a mixture of individuals who self-identified as "`White British`" but who did not cluster close enough to the centroid in PC space to be included in the "White British ancestry subset", together with individuals who self-identified as "`Irish`" or as "`Any other white background`".

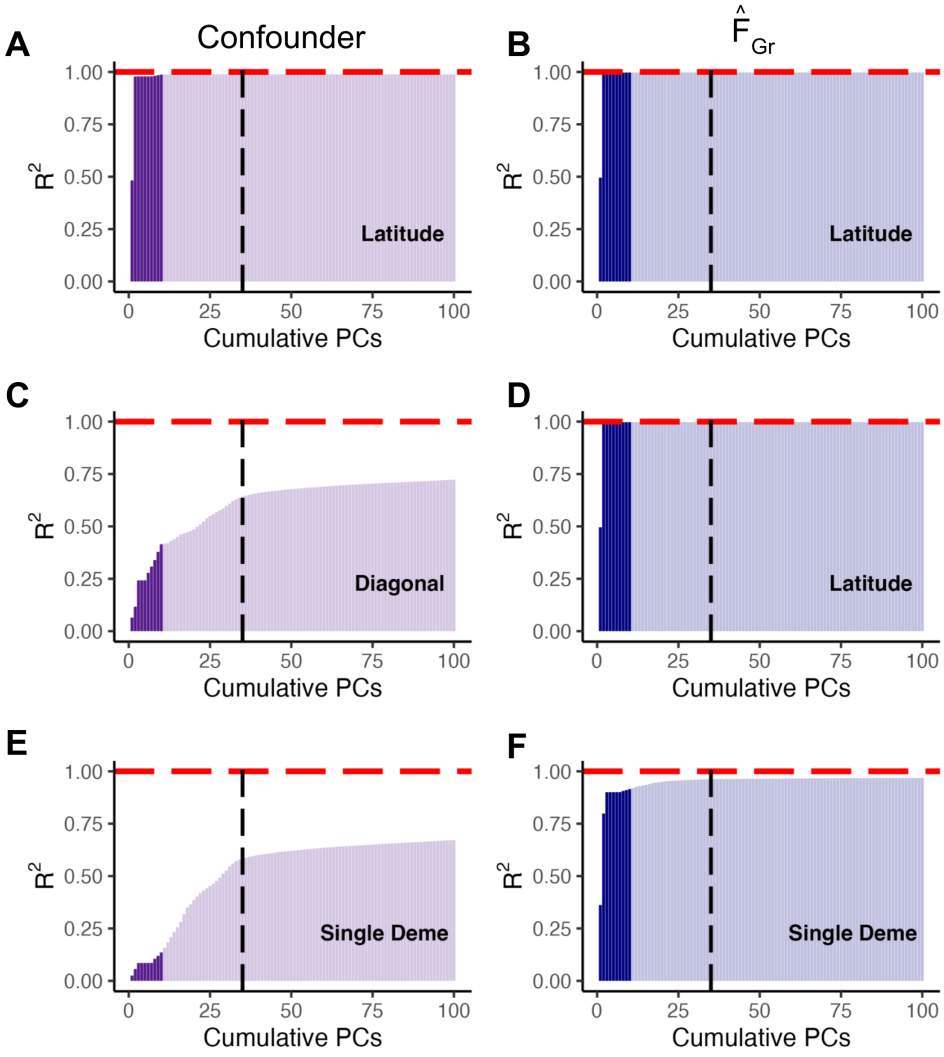

Figure 3.5: **Different patterns of confounding and $\hat{F}_{Gr}$ are captured by different GWAS panel sample PCs.**

For the three possible combinations of confounding and polygenic score association tests in Figure 3.3, we plot the variance in either the confounder or $\hat{F}_{Gr}$ explained by cumulative GWAS panel sample PCs, with the top 10 PCs highlighted in a darker color. As $\tilde{F}_{Gr}$ is unknown for this model, we estimated the error in $\hat{F}_{Gr}$ as 0.011 and 0.04 for latitude and the single deme, respectively, and therefore assume it is a decent proxy for $\tilde{F}_{Gr}$. In (A), both the confounder and $\hat{F}_{Gr}$ (and therefore $\tilde{F}_{Gr}$) 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 not well captured by top sample PCs. In contrast, the test vector is still latitude and $\hat{F}_{Gr}$ is again well captured by PCs 1 and 2. Finally, in (C), both the confounder and the test vector represent membership in a single deme and therefore not as well captured by top sample PCs.

For each test vector, we computed $\hat{F}_{Gr}$ as in Equation 3.2 with different numbers of SNPs ($L = 1,000$ to $L = 1,000,000$) that are found at greater than 1% frequency in the total sample of individuals who self-identified as "`White`". We then quantified the error in estimates of $\hat{F}_{Gr}$ by using a block jackknife across chromosomes to estimate the error for each individual, and then averaging across all individuals (see Section 3.7.2). As expected, error decreased as we increased the number of SNPs (Figure 3.6A) with 0.027 and 0.032 being the smallest error achieved for the east and north test vectors, respectively.

We then computed $\hat{q}$ for standing height, estimating effect sizes and ascertaining SNPs in the GWAS panel by taking at most one SNP per approximately independent LD block[?], and only taking SNPs for which $p < 10^{-5}$ (see Section 3.7.2). When no population structure covariates were included in the GWAS, we found a highly significant positive value of $\hat{q}$ for the east-west axis ($\hat{q} = 8.7$, $p = 3.1 \times 10^{-18}$) and a highly significant negative value of $\hat{q}$ for the north-south axis ($\hat{q} = -5.4$, $p = 7.7 \times 10^{-8}$). When we include $\hat{F}_{Gr}$, the value of $\hat{q}$ drops significantly even when $\hat{F}_{Gr}$ is poorly estimated (e.g. at $L = 1,000$) and ultimately changes sign so that when $\hat{F}_{Gr}$ is estimated with $L = 1,000,000$, $\hat{q} = -1.9$ ($p = 0.06$) and $\hat{q} = 2.4$ ($p = 0.02$) for east and north, respectively. If we take our significance threshold to be $\alpha = 0.05$ and apply a Bonferroni correction to account for the fact that we perform two tests, then we would declare $\hat{q}$ significantly different from zero if $p < 0.05/2 = 0.025$. With these choices, the east-west test is not significant, while the north-south test is just past the threshold. We also repeated this analysis while including only the top 40 sample PCs of the GWAS panel genotype matrix (and not $\hat{F}_{Gr}$), as well as both 40 PCs and our best estimated $\hat{F}_{Gr}$. While doing so changed the numerical value of $\hat{q}$, in neither case is it significantly different from the value we obtain when using only the best estimated $\hat{F}_{Gr}$, nor is it significantly different from zero.

We also note that including the top 40 PCs significantly reduced the number of SNPs reaching our p-value threshold for inclusion. When we use only $\hat{F}_{Gr}$ to control for structure,

we find 1694 and 1693 SNPs below the p-value threshold for east and north, respectively, whereas when we only include 40 PCs we find 310 for both axes. Including both 40 PCs and $\hat{F}_{Gr}$ resulted in 306 and 310 SNPs below the threshold for east and north, respectively. We expect that this is because, when we include only $\hat{F}_{Gr}$ as a control for stratification, there can still be considerable bias along other axes, which will still lead to inflated χ^2 association statistics, even if the bias along our chosen test axis is substantially reduced. This observation suggests that using $\hat{F}_{Gr}$ in isolation is not optimal for empirical applications as stratification along other axes in the GWAS panel makes ascertaining causal SNPs (or tightly linked tagged SNPs) difficult.

Additionally, the fact that there is little change between the value of $\hat{q}$ when including 40 PCs vs. including 40 PCs plus $\hat{F}_{Gr}$ suggests that these 40 PCs are doing an adequate job at capturing $\tilde{F}_{Gr}$. In the next Section, we return to this idea and present a procedure for estimating the fraction of $\tilde{F}_{Gr}$ captured by J sample PCs in empirical data, providing a means to check if the test of choice is susceptible to potential confounding in a given GWAS panel when including a given number of sample PCs.

3.4 Testing for differences in polygenic scores in empirical data

In this Section, we fully turn our attention to empirical analyses with the goal of developing a procedure that allows researchers to estimate how well protected their polygenic score association test is from stratification bias. First we introduce the statistic, $\hat{H}$, the average proportion of variance in individual GWAS panel genotypes that is explained by a set of allele frequency contrasts, r . Under the null hypothesis of no overlapping structure, $\mathbb{E}\left[\hat{H}\right] \approx \frac{1}{L}$. If the null hypothesis is rejected, the test is susceptible to potential confounding in the GWAS panel. For susceptible tests, we introduce $R^2_{J/\hat{\gamma}_{F_{Gr}}}$, an estimate of the proportion of $\tilde{F}_{Gr}$ that is captured by J sample PCs with values closer to one indicating better protection. For tests that either fail to reject the null hypothesis of no overlap or

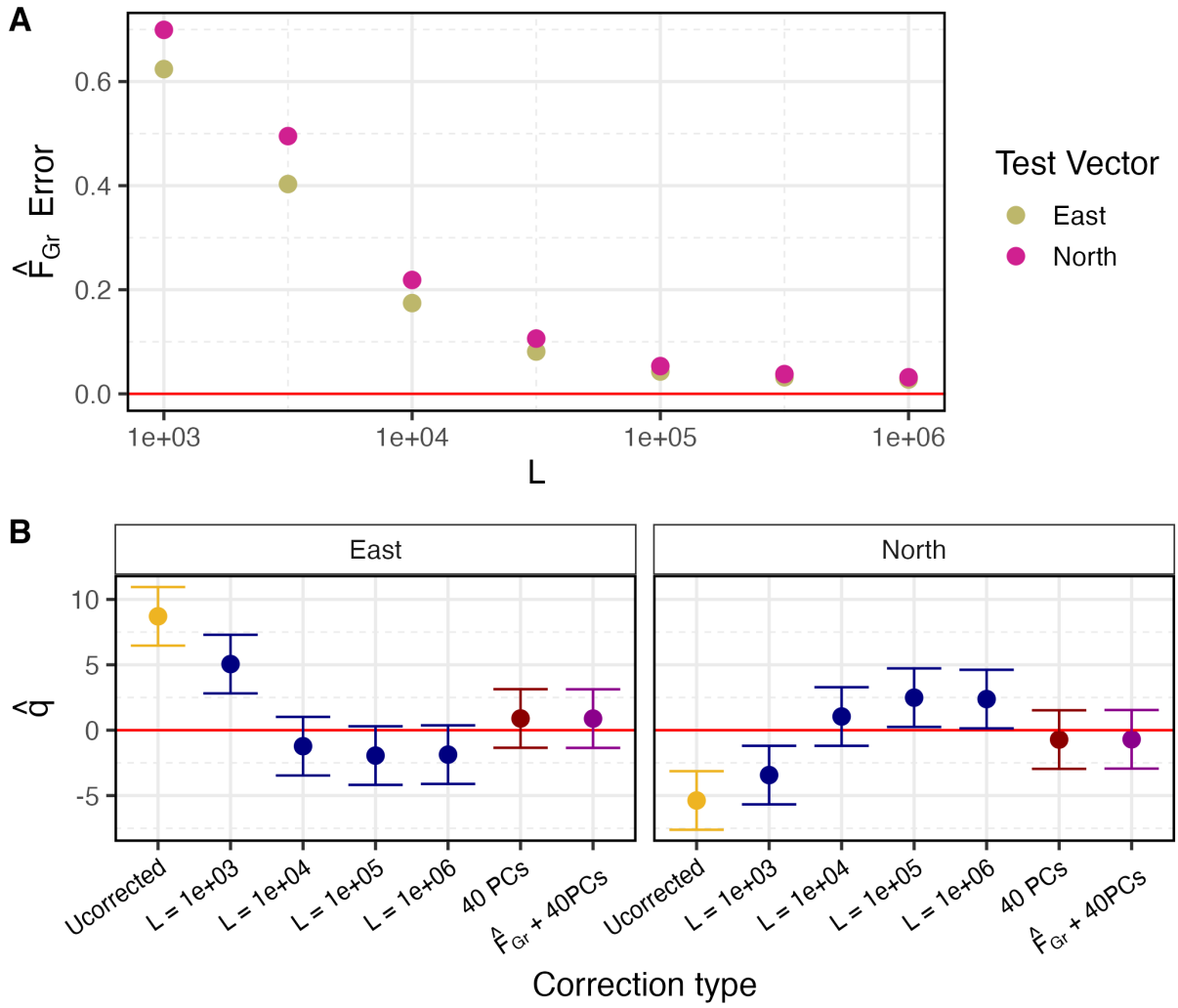

Figure 3.6: **Quantifying error in $\hat{F}_{Gr}$ for two polygenic score association tests in the UK Biobank**

(A) For a sample of $N = 399,801$ individuals in the UK Biobank we used their birth location as recorded in northing and easting coordinates as two separate test vectors and computed $\hat{F}_{Gr}$ for a distinct GWAS panel of $M = 50,136$ individuals. We varied the number of SNPs used to compute $\hat{F}_{Gr}$ and, as expected, increasing the number of SNPs decreased the estimated error in $\hat{F}_{Gr}$. (B) Using each estimate of $\hat{F}_{Gr}$ as a covariate, we conducted GWASs for standing height and ascertained sets of SNPs used to compute $\hat{q}$ in the test panel (confidence intervals are shown after applying a Bonferroni correction to account for the two tests we perform). We compared these results to an "uncorrected" $\hat{q}$ computed using effect sizes estimated in a GWAS with no covariates included and two additional estimates of $\hat{q}$, one where 40 sample PCs were included in the GWAS and one where 40 sample PCs and the estimate of $\hat{F}_{Gr}$ for $L = 1,000,000$ were included.

have $R^2_J/\hat{\gamma}_{F_{Gr}}$ values close to one, we have high confidence that $\hat{q}$ is unbiased. We apply our procedure to 30 sets of allele frequency contrasts using four different subsamples of the UKBB as GWAS panels, each with varying degrees of overlap with the frequency contrasts. For each contrast/GWAS panel pair we run a polygenic score association test for 17 different phenotypes in the UK Biobank.

3.4.1 Datasets

Test Panels

Human Genome Diversity Project and Thousand Genomes Project

I downloaded high-coverage whole genome sequencing data[108] from 4,094 diverse individuals from the harmonized Human Genome Diversity Project (HGDP) and the Thousand Genome Project (1kGP). From these data, we constructed 15 different sets of allele frequency contrasts (r).

First, we grouped individuals by the provided `project meta project pop` code that corresponds to continental-level ancestry groups and selected only groups with at least 200 samples, yielding 5 groups: East Asia (eas), Non-Finish Europe (nfe), Africa (afr), South Asia (sas), and America (amr). We then computed the allele frequency differences between all pairs of groups, resulting in 10 different contrasts (see Figure 3.21A and Table 3.1 for a map of samples and sample sizes, respectively).

Next, we selected five different test axes where stratification has been of concern in previous studies of polygenic adaption. We note that all of the following examples studied associations with polygenic scores for height, but here we test every contrast for association with polygenic scores for all 17 traits (including height). Additionally, while the motivation behind choosing these axes stem from previous inquiries, we do not use the exact same test panel datasets so our study should not be considered an attempt to replicate any previous study but rather an inquiry into stratification along similar ancestry gradients.

First we consider the test vectors of latitude in Europe and longitude in Eurasia. As discussed in Chapter 1, Berg et al. (2019)[36] and Sohail et al. (2019)[72] found reduced signal for polygenic selection along both latitude in Europe and longitude in Eurasia after stricter control for population stratification. Here we aim to re-test these axes, along with longitude in Europe and latitude in Eurasia. To this end, we first constructed a Eurasian panel by selecting individuals with "fin", "nfe", "eas", "sas" `project meta project pop` codes, and used latitude and longitude as test vectors to generate two sets of allele frequency contrasts (see Figure 3.21B and table 3.1). Next, we constructed a European panel by selecting only individuals with "nfe" as their `project meta project pop` code, and similarly used latitude and longitude as test vectors to generate another pair of frequency contrasts (see Figure 3.21C and table 3.1).

Finally, we consider the test axis from Chen et al. (2020)[73]. As discussed in Chapter 1, this study found evidence of polygenic adaptation for decreased height in Sardinian individuals compared to mainland Europeans using evolutionary diverged GWAS and test panels. Here we aim to test for polygenic score divergence along the same ancestry gradient and generate r by computing the allele frequency difference between individuals with the `project meta project pop` code "nfe" and `project meta project subpop` code "sdi", corresponding to individuals from Sardinia. (see Figure 3.21D and table 3.1).

Single Density Scores

Single density score (SDS), introduced by Field et al. (2016)[34], is a method to infer recent changes in allele frequency. SDS is motivated by the fact that recent selection distorts underlying genealogies, making terminal branches shorter and resulting in longer genomic distances between singleton mutations. This distance is then used as a site-level summary statistic whose value reflects the amount of allele frequency change in the past 2,000 - 3,000 years. SDS values can be combined with GWAS summary statistics to test for recent selection on a trait (tSDS) by testing for a significant correlation between effect size estimates and SDS

values. One interpretation of a significant correlation is recent selection acting on the trait in question to increase the frequency of trait-increasing alleles. In the original SDS paper, Field et al. (2016) used tSDS to conclude that height, alongside 15 other phenotypes, has been under selection in the ancestors of modern British individuals. However, Berg et al. (2019)[36] and Sohail et al. (2019)[72] concluded that the height signal was overestimated, likely due to population stratification in the original GWAS effect size estimates, suggesting that, similar to tests of polygenic score difference, tSDS analysis is susceptible to stratification bias. Here, we use the original SDS values, estimated in the UK10K panel[109], as our vector of allele frequency contrasts to re-evaluate SDS in light of our results. Additionally, we use SDS values estimated by Luo et al. (2023)[110] in the Han Chinese populations as an additional set of allele frequency contrasts.

Ancient DNA contrasts

Le et al. (2022)[6] use ancient DNA from 1,297 individuals from Holocene Europe to identify loci with rapid frequency change across three major epochs of European history: the Neolithic, Bronze Age, and Historical period. The authors start by splitting the total sample into five groups corresponding to different time transects. They then model the ancestry composition of each of the three target epochs as a mixture of the five groups. The expected allele frequency for each epoch is therefore a weighted average of ancestral allele frequencies. The authors detect loci under selection post-admixture by comparing the expected allele frequency predicted by the demographic model to the observed allele frequency in the actual samples. Alleles under positive selection should be observed at higher allele frequencies than predicted by the neutral model. Similar to SDS, the authors compare the correlation between selection statistics and GWAS effect size estimates (from Biobank Japan) to detect selection in polygenic traits, finding a total of 39 traits under selection across the three epochs. Here we use the difference between observed and expected allele frequencies for the three epochs from Le et al. (2022) as three separate r vectors.

British Isles country of birth

In order to construct a set of allele frequency contrasts that capture more subtle, sub-continental ancestry gradients we used individuals from the UK Biobank who were not members of any of the four GWAS panels (see below), and their country of birth to construct 10 sets of allele frequency contrasts. We first used the data field 21000 to select individuals who self-identified as `White` and then used data field 1647 to get individuals who were born in the British Isles and who recorded their country of birth as one of the following five countries: England, Scotland, Northern Ireland, Republic of Ireland, or Wales (see table 3.2 for sample sizes). Finally, we calculated all pairwise allele frequency differences between the five countries, resulting in 10 sets of r vectors representing within the British Isles country of birth differences.

GWAS Panels

Our goal was to use the UK Biobank[43] to construct four different GWAS panels with different levels of genetic diversity while maintaining sufficient sample sizes to conduct well-powered GWASs, allowing us to compare the results of the above polygenic score association tests across a range of realistic scenarios. To this end, we used a sampling scheme developed by Steiner et al. (unpublished) to create four panels of 100,000 individuals each.

To construct each panel, we first calculated the median PC 1 and PC 2 coordinates of the entire biobank and then calculated the Euclidean distance between all individuals and the median. Next we chose four sampling distances, 5ϵ, 10ϵ, 50ϵ, and 200ϵ, with $\epsilon = 1e^{-4}$, and randomly sampled 100 thousand individuals with distances smaller than each of four thresholds. Figure 3.7A shows the PC1 vs PC2 biplot for each of the four panels with the colored dots representing the sampled individuals.

3.4.2 Overlap between panels

In Chapter 2, we demonstrated that bias in polygenic score association tests is produced when the vector of expected confounders aligns with the test vector of interest in a space defined by $\mathbf{F}_{GX}$, the cross-panel kinship matrix (see Equation 2.12). Thus, for any set of L allele frequency contrasts measured in the test panel, r, the susceptibility to confounding will depend on the overlap in structure between r and the GWAS panel $\mathbf{G}$. Here we outline an empirical test for overlap in population structure and apply it to all 30 sets of allele frequency contrasts.

First, suppose that r has been mean-centered and standardized such that it has $Var(r) = 1$ and the columns of $\mathbf{G}$ have been standardized such that $Var(G_{.\ell}) = 1$. To test for overlap in structure between the allele frequency contrasts and the GWAS panel, we first define the statistic,

$$H = \frac{\mathbb{E}_M \left[Cov(r, G_{i.})^2 \right]}{\mathbb{E}_M \left[Var(G_{i.}) \right]} \tag{3.8}$$

where the expectation is taken over individuals in the GWAS panel. H is therefore the average proportion of genetic variance in the GWAS panel that lies along the test axis.

Next, we will propose an estimator of H, but first, we define the LD matrix, R, and its eigen decomposition as

$$\mathbf{R} = \frac{1}{M-1}\mathbf{G}^\top \mathbf{G} = \mathbf{V}^\top \mathbf{\Lambda}^2 \mathbf{V} \tag{3.9}$$

where $\mathbf{V}^\top$ is the $L \times M$ matrix containing the SNP loadings and the diagonal of $\mathbf{\Lambda}^2$ contains the eigenvalues. We denote the rows of $\mathbf{V}$ as v_k which record the loading of the L SNPs onto the k^{th} principal component. Each v_k has $Var(v_k) = \frac{1}{L}$.

Then, we propose that,

$$\hat{H} = \frac{1}{M(L-1)^2} r^\top \mathbf{G}^\top \mathbf{G} r \tag{3.10}$$

$$= \frac{M-1}{M(L-1)^2} r^\top \mathbf{R} r. \tag{3.11}$$

is an estimator of H, and note that if we define our population structure estimator as $\hat{F}_{Gr} = \frac{1}{\sqrt{L-1}}\mathbf{G}r$, then $H = \frac{M-1}{M(L-1)}\hat{F}_{Gr}^\top \hat{F}_{Gr}$, making $\hat{H}$ easy to calculate using our existing methods.

Under the null hypothesis of no overlap in structure between panels ,we can write the expectation of $\hat{H}$

$$\mathbb{E}[\hat{H}|\text{no overlap}] = \frac{M-1}{M(L-1)^2}\mathbb{E}\left[r^\top \mathbf{R} r\right] \tag{3.12}$$

$$= \frac{M-1}{M(L-1)^2}\mathbb{E}\left[r^\top \mathbf{V}^T \mathbf{\Lambda}^2 \mathbf{V} r\right] \tag{3.13}$$

$$= \frac{M-1}{M(L-1)^2}\sum_K \lambda_k^2 L \mathbb{E}\left[(r_\ell v_{k,\ell})^2\right] \tag{3.14}$$

$$= \frac{M-1}{M(L-1)^2}\sum_K \lambda_k^2 \tag{3.15}$$

$$= \frac{(M-1)L}{M(L-1)^2} \tag{3.16}$$

$$\approx \frac{1}{L} \tag{3.17}$$

where λ_k is the k^{th} eigenvalue of the LD matrix. Because r and v_k are independent under the null hypothesis, $\mathbb{E}\left[(r_\ell v_{k,\ell})^2\right] = \frac{1}{L}$ and the sum of the eigenvalues of the LD matrix is L.

Next, we can estimate the variance of $\hat{H}$ via a jackknife approach across $B = 581$ approximately independent LD blocks,

$$\hat{\sigma}_H^2 = \frac{B-1}{B}\sum_{i=1}^{B}\left(\hat{H}^{(-i)} - \frac{1}{B}\sum_{j=1}^{B}\hat{H}^{(-j)}\right)^2 \tag{3.18}$$

and conduct a one-tailed hypothesis test for $\hat{H} > \frac{1}{L}$.

After applying the above procedure to all 30 sets of contrasts outlined in Section 3.4.1 and applying multiple testing correction for 120 tests (30 contrasts $\times$ 4 GWAS panels), we found that 100/120 contrasts GWAS panel pairs had significant overlap. Figure 3.7B shows $\hat{H}$ for all 30 sets of allele frequency contrasts in all four GWAS panels. The top row of Figure 3.7B has the results of the continental-level contrasts in the HGDP1kGP. Unsurprisingly, increasing the sampling distance in the GWAS panel increases the average variance explained by these contrasts, with the European-African ancestry contrasts explaining the largest average proportion of variance (0.0009 in the 200ϵ panel) of any contrast. In contrast to the continental contrasts, the average proportion of variance explained for the within the British Isles country of birth contrasts (third row in Figure 3.7B) only increases very slightly (or not at all) with increasing sampling distance as variation along these axes is well captured by the 5ϵ dataset that contains individuals with primarily white British ancestry.

The value of $\hat{H}$ for the aDNA Le et al. (2022) contrasts and the UK10K Field et al. (2016) SDS contrasts, all of which are meant to capture allele frequency shifts within Europe, does increase with sampling distance but at a smaller magnitude than the HGDP1kGP. The European Neolithic and Historical period aDNA contrasts, in particular, explain less variance than expected for the 5ϵ, 10ϵ, and 50ϵ panels. We are unsure what is causing these observations. It is worth noting that, unlike the HGDP1kGP and British Isles contrasts, the Le et al. contrasts are not allele frequency differences directly; instead, they are the difference between observed and expected allele frequencies, where the allele frequencies themselves were inferred via a maximum likelihood approach from read counts in ancient genomes. It is unclear what, if anything, about this procedure would cause a smaller proportion of average variance explained than we expect under the null. Further investigation is needed to determine possible constraints on what types of contrasts can be used for our test. Finally, the SDS values computed using individuals with Han Chinese ancestry have no overlap in the

5ϵ and 10ϵ panels, but they explain a significant proportion of variance in 50ϵ and 200ϵ panels. Increasing the sampling distance increases the proportion of East Asian ancestry included in the GWAS panel and therefore the ability to capture the Han SDS axis of variation.

Our results indicate that for the 20 GWAS/contrast pairs where we were unable to reject the null hypothesis of independent structure, the resulting polygenic score association tests should be well protected from stratification bias. For tests where the null is rejected, it is important to note that choosing a GWAS panel with a smaller $\hat{H}$ does not *necessarily* lead to less bias in $\hat{q}$ as the strength of potential confounders is unknown. It's possible, though perhaps unlikely, that panels where $\hat{H}$ is small could have more confounding along the shared axis than panels where $\hat{H}$ is larger. So, while a smaller value of $\hat{H}$ should be associated with less bias for a given amount of confounding, we generally know very little about the pattern or strength of confounding and should avoid strong assumptions about the relationship between the value of $\hat{H}$ and the bias of a given polygenic score association test.

3.4.3 Variance in $\hat{F}_{Gr}$ explained by sample PCs

To protect a polygenic association test using set allele frequency contrasts, r, from stratification bias, variation along $\tilde{F}_{Gr}$ must be controlled for in the GWAS panel. In Section 3.2, we discussed two approaches, estimating $\tilde{F}_{Gr}$ directly and including the estimate as a covariate in the GWAS or including some number of sample PCs as covariates, and the relationship between the two. However, even when $\tilde{F}_{Gr}$ is well estimated by $\hat{F}_{Gr}$, including it as the sole control for population structure is sub-optimal as confounding along other axes in the dataset may confound the ascertainment process, leading to a less powerful polygenic score. In contrast, the inclusion of multiple PCs controls for stratification along multiple axes of variation, but is not guaranteed to capture $\tilde{F}_{Gr}$. One benefit of $\hat{F}_{Gr}$ is that it is possible to estimate the error directly using a block jackknife approach, something that is

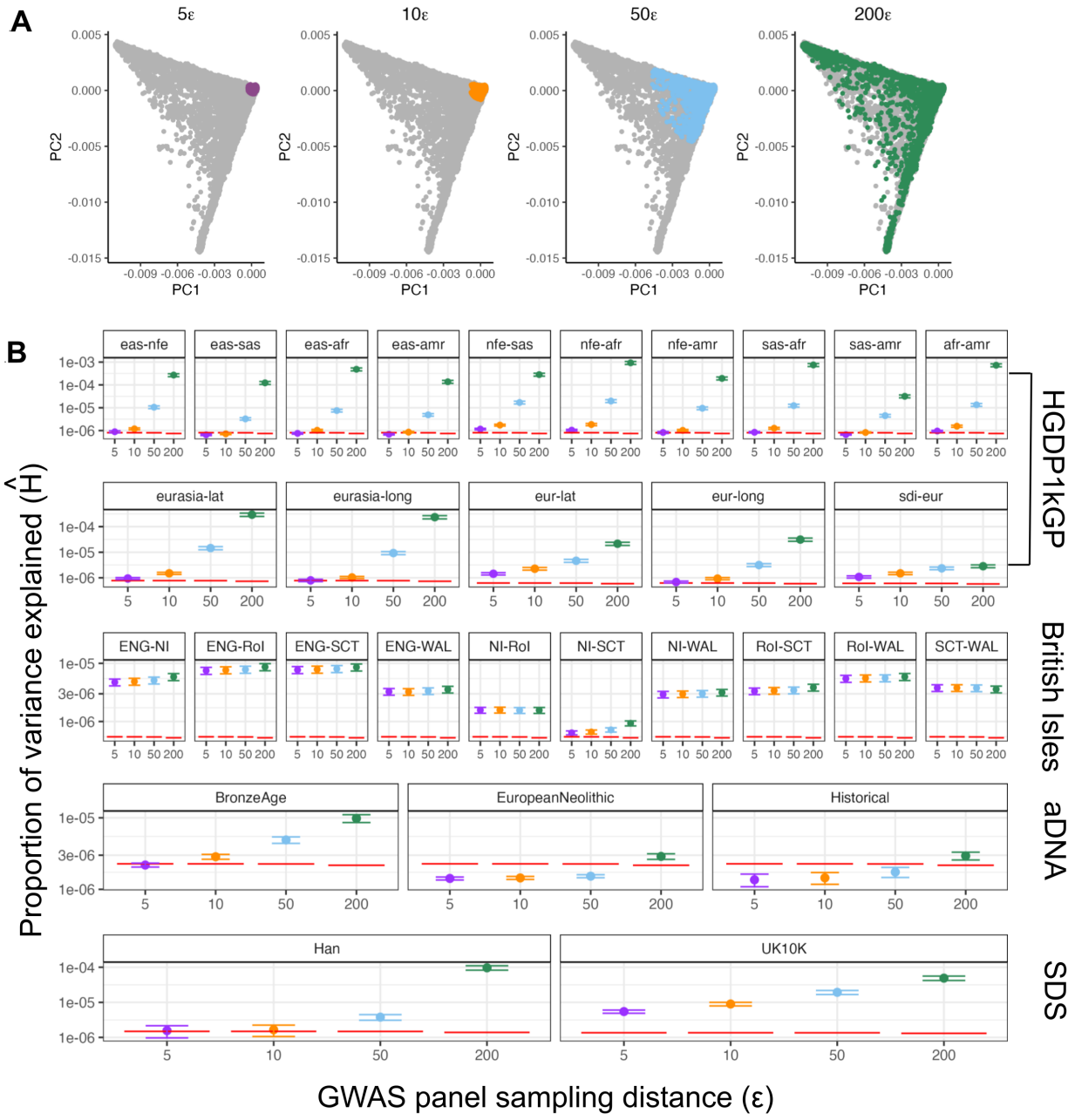

Figure 3.7: **GWAS panel sampling scheme and proportion of variance explained.** (A) We constructed four GWAS panels of 100K individuals each from the UKBB following the sampling scheme outlined by Steiner et al. (unpublished). Setting $\epsilon = 1e^{-4}$, we sampled individuals within 5, 10, 50, and 200 ϵ from the median position on PC 1 and PC 2 using Euclidean distance computed on PCs 1 and 2 as our distance metric (B) We computed $\hat{H}$, the proportion of variance explained by r in the given GWAS panel for all 30 sets of test panel contrasts. The red line is the expectation of $\mathbb{E}[\hat{H}] = \frac{1}{L}$ under the null.

difficult to do with sample PCs (though see Haag et al. (2024)[111]). Here, we outline a procedure for using $\hat{F}_{Gr}$ and the estimate of its error to determine if a set of external allele frequency contrasts is protected by J sample PCs, thereby combining the benefits of both approaches.

We start by writing our estimator $\hat{F}_{Gr}$ as the sum of the true axis of structure $\tilde{F}_{Gr}$ and some error,

$$\hat{F}_{Gr} = \sqrt{\gamma_{F_{Gr}}}\tilde{F}_{Gr} + \sqrt{(1-\gamma_{F_{Gr}})}e_{F_{Gr}} \tag{3.19}$$

where $Var(\tilde{F}_{Gr}) = 1$ and $e_{F_{Gr}}$ is a vector of random Normals with $Var\left(e_{F_{Gr}}\right) = 1$. The accuracy of the estimator $\hat{F}_{Gr}$ is therefore described by $0 < \gamma_{F_{Gr}} < 1$, with values closer to 1 corresponding to a more accurate estimate. Given our standardization and scaling, we have

$$Var\left(\hat{F}_{Gr}\right) = \gamma_{F_{Gr}} + 1 - \gamma_{F_{Gr}} \tag{3.20}$$

$$= 1, \tag{3.21}$$

and the fraction of variance in $\tilde{F}_{Gr}$ that is explained by $\hat{F}_{Gr}$ is

$$\frac{Var(\sqrt{\gamma_{F_{Gr}}}\tilde{F}_{Gr})}{Var(\tilde{F}_{Gr})} = \gamma_{F_{Gr}}. \tag{3.22}$$

As outlined in Section 3.7.2, we can estimate the error $\hat{e}_{F_{Gr}}$ via a block jackknife approach and compute $\hat{\gamma}_{F_{Gr}} = 1 - \hat{e}_{F_{Gr}}$ to determine what fraction of variance in our estimator represents a true signal of population structure. In Figure 3.8, we plot the error in $\hat{F}_{Gr}$ for all contrasts in all GWAS panels against $\hat{H}$, the average proportion of variance in the GWAS panel explained by the contrast. In general, the larger the proportion of variance a contrast explains, the more accurate the estimate of $\tilde{F}_{Gr}$ is.

Next, we turn to PCs. We assume the population PCs are scaled such that $Var(U_i) = 1$

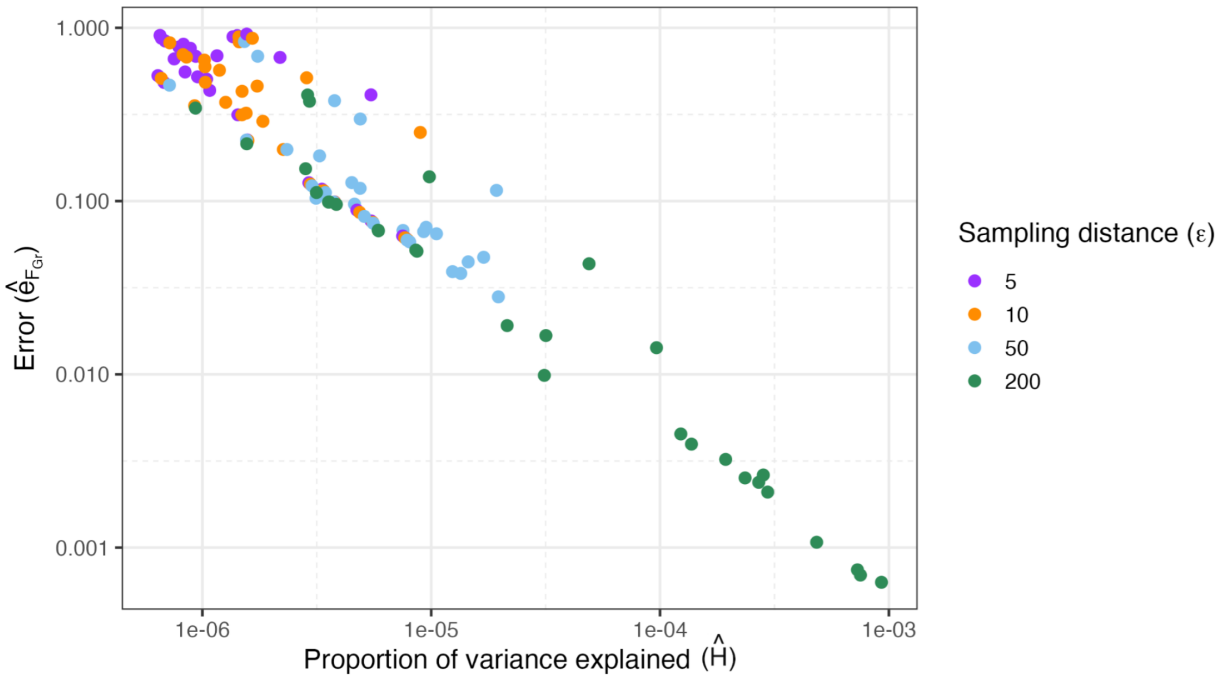

Figure 3.8: **The larger the average proportion of variance a contrast explains, the more accurate the estimates of $\hat{F}_{Gr}$**
For all sets of 30 allele frequency contrasts, we plot the average proportion of variance explained (see 3.7) for each contrast/GWAS panel pairs against our block jackknife estimate of the error in $\hat{F}_{Gr}$ (see 3.7.2).

for all i. $\tilde{F}_{Gr}$ can be interpreted as a linear combination of the population PCs

$$\tilde{F}_{Gr} = \sum_{i=1}^{\min(M,L)} \sqrt{\alpha_i} U_i \tag{3.23}$$

where $\sqrt{\alpha_i} = \text{Cov}\left(U_i, \tilde{F}_{Gr}\right)$, given our standardization. Our standardization also implies that the proportion of variance in $\tilde{F}_{Gr}$ that is explained by U_i is equal to α_i, and because $\tilde{F}_{Gr}$ must lie within the span of the PCs and the U_i are orthogonal to one another, the total variance explained is simply $\sum_i^{\min(M,L)} \alpha_i = 1$.

Generally, we might think that $\tilde{F}_{Gr}$ corresponds to a sufficiently important axis of pop-

ulation structure in the GWAS panel that only some top J PCs are needed, i.e., that

$$\tilde{F}_{Gr} = \sum_{i=1}^{J} \sqrt{\alpha_i} U_i \tag{3.24}$$

for $J \ll \min(M, L)$. This would imply that $\alpha_i = 0$ for all $i > J$, so that $\sum_i^J \alpha_i = 1$, for some $J \ll \min(M, L)$.

Similar to our direct estimator, sample PCs need to be estimated from data. To this end, we write

$$\hat{U}_i = \sqrt{\gamma_i} U_i + \sqrt{(1-\gamma_i)} e_i \tag{3.25}$$

where again e_i is a vector of random Normals with $Var(e_i) = 1$ and $0 < \gamma_i < 1$, so that

$$Var\left(\hat{U}_i\right) = \gamma_i + 1 - \gamma_i \tag{3.26}$$

$$= 1 \tag{3.27}$$

and the fraction of variance in U_i that is explained by $\hat{U}_i$ is

$$\frac{Var(\sqrt{\gamma_i} U_i)}{Var(U_i)} = \gamma_i. \tag{3.28}$$

Now, suppose we fit a model in which we try to explain $\hat{F}_{Gr}$ as a function of the top J sample PCs,

$$\hat{F}_{Gr} = \sum_{i}^{J} \beta_i \hat{U}_i + \epsilon. \tag{3.29}$$

For each i, we have

$$\hat{\beta}_i = Cov\left(\hat{F}_{Gr}, \hat{U}_i\right) \tag{3.30}$$

$$= \sqrt{\gamma_{F_{Gr}}\gamma_i} Cov\left(\tilde{F}_{Gr}, U_i\right) \tag{3.31}$$

$$= \sqrt{\gamma_{F_{Gr}}\gamma_i\alpha_i}, \tag{3.32}$$

and given our standardization, the variance in $\hat{F}_{Gr}$ explained by $\hat{U}_i$ is $\hat{\beta}_i^2 = \gamma_{F_{Gr}}\gamma_i\alpha_i$. Notably, the step from 3.30 to 3.31 requires the assumption that e_i and $e_{F_{Gr}}$ are uncorrelated. The $\hat{U}_i$ are all orthogonal to one another, so the total variance in $\hat{F}_{Gr}$ explained by the top J PCs is

$$R_J^2 = \sum_i \hat{\beta}_i^2 = \gamma_{F_{Gr}} \sum_i^J \gamma_i\alpha_i. \tag{3.33}$$

If the top J PCs are all well estimated (i.e., $\gamma_i \approx 1$ for $i \leq J$) and are sufficient to explain $\tilde{F}_{Gr}$ (i.e., $\sum_i^J \alpha_i = 1$), then

$$R_J^2 = \gamma_{F_{Gr}} \sum_i^J \alpha_i = \gamma_{F_{Gr}} \tag{3.34}$$

i.e., the fraction of variance in $\tilde{F}_{Gr}$ that can be explained by $\hat{F}_{Gr}$. Because the maximum variance in $\hat{F}_{Gr}$ that can be explained by J sample PCs is $\gamma_{F_{Gr}}$, we can combine R_J^2 and our estimate of the signal for each contrast $\hat{\gamma}_{F_{Gr}}$ to estimate the fraction of $\tilde{F}_{Gr}$ captured by the J PCs.

In practice, we first compute $\hat{F}_{Gr}$ and $\hat{\gamma}_{F_{Gr}}$ for all sets of contrasts in each GWAS panel using SNPs on odd chromosomes, and we compute 40 sample PCs in each GWAS panel using SNPs on even chromosomes. Using SNPs on opposite chromosomes allows us to compute R_J^2 without double dipping and ensures that the residuals e_i and $e_{F_{Gr}}$ are uncorrelated. We note that using only half the number of SNPs to estimate both $\hat{F}_{Gr}$ and sample PCs will

increase the error in each. For sample PCs, in particular, reducing $\sqrt{\frac{M}{L}}$ may render some population PCs undetectable or poorly estimated. When the full set of SNPs is used to compute the PCs that get included in the GWAS, it is possible that a larger fraction of $\tilde{F}_{Gr}$ will be captured, making our estimate a lower bound. In Figure 3.9A we show our results for the HGDP1kGP latitude in Europe contrast in all four GWAS panels. The dashed red line is our estimate of the signal in $\hat{F}_{Gr}$. When the sampling distance is 5ϵ, 40 PCs capture only 56% of the signal, but increasing the sampling distance increases both the signal and the faction of signal that is captured by the top PCs. At 200ϵ, 98% of the signal is captured by the top 40 PCs and a polygenic score association test in which 40 sample PCs are included in the GWAS would be well protected from stratification bias.

To better summarize our data, we compute $R^2_J/\hat{\gamma}_{F_{Gr}}$: the fraction of signal explained by J PCs. Figure 3.9B plots this ratio as a function of PCs for the same latitude in Europe contrast. This visualization again suggests that this contrast is well protected by 40 PCs in the 200ϵ panel, whereas all three other panels are susceptible to potential confounding along $\tilde{F}_{Gr}$, even with 40 PCs included. We compute $R^2_J/\hat{\gamma}_{F_{Gr}}$ for all contrasts where we rejected the null hypothesis of independent panels and plot the results in Figure 3.9C. In the HGDP1kGP panel, all continental contrasts are well protected in the 200ϵ panel but are susceptible in all other panels with ratios much less than one. The same is true for all latitude and longitude contrasts in this dataset. However, the Sardinia vs. mainland Europe contrast is not completely captured by 40 PCs in any panel. Unlike the HGDP1kGP contrasts, $R^2_J/\hat{\gamma}_{F_{Gr}}$ is similar across panels for most of the country of birth within the British Isles contrasts, reflecting the fact that the ancestry gradient represented by these subtle contrasts exists mainly within the smallest sampling distance tested. In the aDNA panel, all three contrasts are well captured by the 200ϵ sampling distance. For the Historical period contrasts, we observe a ratio slightly greater than one for the 200ϵ panel ($R^2_{40}/\hat{\gamma}_{F_{Gr}} = 1.02$). This is the only contrast/GWAS panel pair we observe this phenomenon for, and we are unsure what

is causing this violation of our model. For the SDS contrasts, both the Han and UK10K inferences are well protected by the top 40 PCs in the 200ϵ panel. It is perhaps surprising that the UK10K contrasts, designed to capture recent selection in individuals of British ancestry, are not as well protected in the narrow sampling distances. Counterintuitively, using a more diverse panel with 40 PCs should protect these contrasts better than the more homogeneous panels.

3.4.4 Polygenic score association test results from 17 phenotypes

In this Section, we use our 30 sets of allele frequency contrasts and four GWAS panels to conduct polygenic score association tests for 17 different phenotypes in the UK Biobank (see Table 3.3 for a complete list). For each phenotype/GWAS panel combination, we run two GWASs: one including the top 40 sample PCs and one with no correction for population structure. For all GWASs we included age, sex, and genotype measurement batch as covariates and used the simple linear regression model in *fastGWA*[27]. We then selected the minimum p-value per independent LD block (as defined by Berisa and Pickrell (2016)[112]) with $p < 1e^{-4}$ to compute $\hat{q}$. We assess the significance of $\hat{q}$ by estimating the standard error via block jackknife and computing a p-value for the null hypothesis $q = 0$ (see Methods 3.7.2 and Berg et al. (2019)[36]).

First, we compare uncorrected $\hat{q}^2$ to $\hat{q}^2$ computed from a GWASs that included 40 sample PCs across all sets of allele frequency contrasts (see Figure 3.10). When no PCs are included, the larger the average variance explained by the contrasts in the GWAS panel ($\hat{H}$), the larger the inflation of $\hat{q}^2$. This result is consistent with our theory that increased overlap between panels can lead to more bias in polygenic scores if the confounders align with the shared axis of structure. When 40 PCs are included in the GWAS model, inflation is reduced, and the relationship between $\hat{H}$ and $\hat{q}^2$ is weaker. The 5ϵ panel has the largest slope (0.67) at 40 PCs, indicating that many of these tests are not fully protected from stratification, an observation

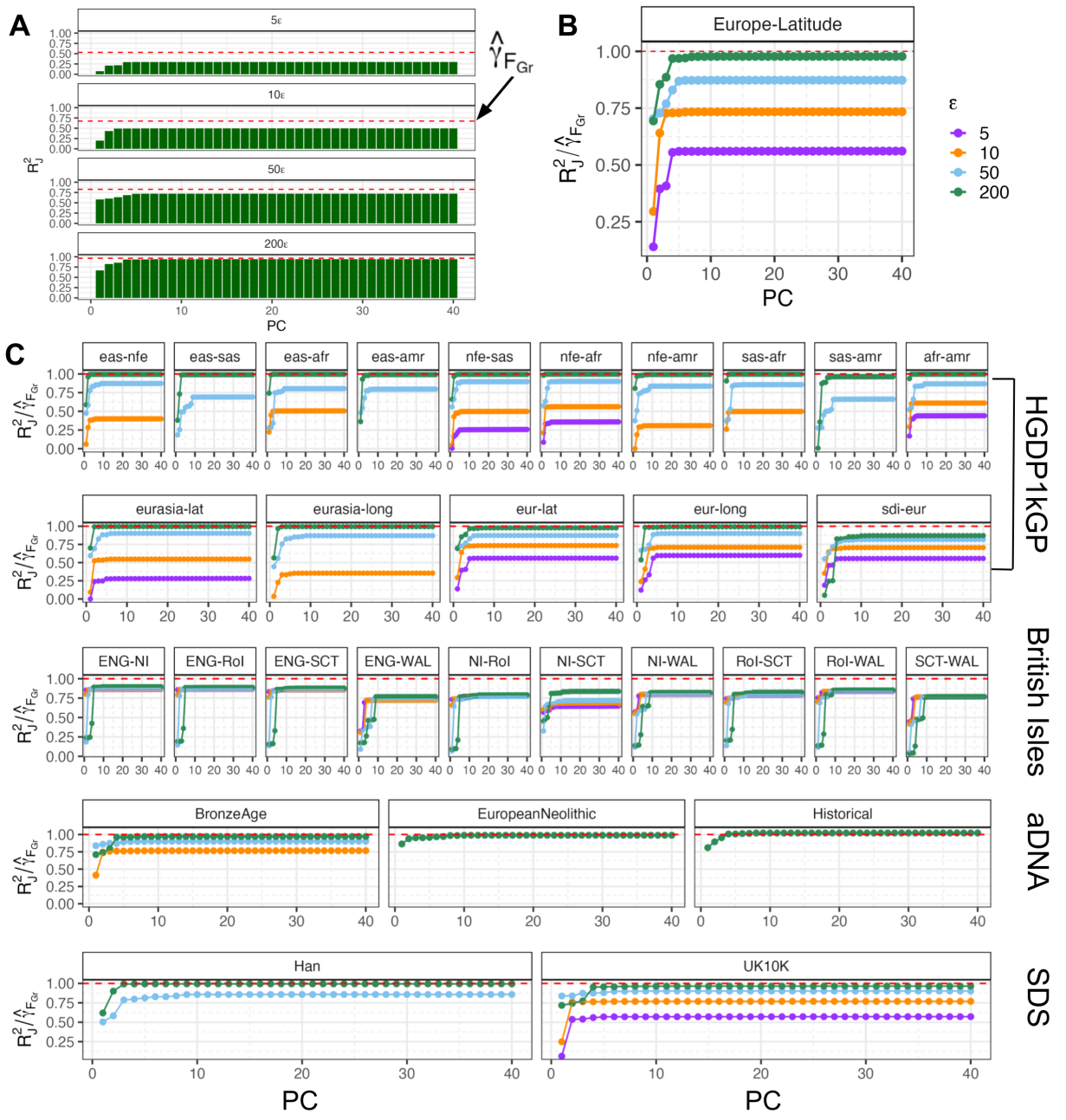

Figure 3.9: **The ratio of variance explained by sample PCs to signal in $\hat{F}_{Gr}$ indicates how well captured $\tilde{F}_{Gr}$ is by J PCs.**

(A) For the HGDP1kGP latitude in Europe contrast, we computed $\hat{F}_{Gr}$ for all four GWAS panel sampling distances. We then compute the cumulative variance explained by each of the top 40 PCs (R^2_J). For this contrast, increasing the sampling distance increases R^2_J. The maximum variance explained is $\hat{\gamma}_{F_{Gr}}$, the signal in our estimate of $\hat{F}_{Gr}$. (B) To better summarize our results, we compute $R^2_J/\hat{\gamma}_{F_{Gr}}$ and plot this ratio for each GWAS panel as a function of cumulative PCs. (C) $R^2_J/\hat{\gamma}_{F_{Gr}}$ for all sets of non-independent contrast/GWAS panel pairs.

that is consistent with our previous results that, across the whole set of contrasts, $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ is generally smallest for the 5ϵ panel.

Next, we analyze the results of individual association tests, grouped by dataset. As discussed in the previous Section, the ratio $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ is an estimate of the fraction of $\tilde{F}_{Gr}$ captured by the top 40 PCs. Therefore, we plot the -log10(p-value) for each association test as a function of this ratio for tests when the panels are not independent. For independent contrasts/GWAS pairs, we still plot the -log10(p-value) and simply group contrasts together on the x-axis. Association tests with a significant p-value and either in independent panels or with a ratio close to one are signals that we are most confident represent real polygenic score divergence, whereas significant tests with $R^2_{40}/\hat{\gamma}_{F_{Gr}} << 1$ cannot be ruled out as false positive results. We note that we don't use a strict cutoff on $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ to distinguish true positives from false positives but instead use $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ as a measure of confidence. Additionally, we draw two different Bonferroni cutoffs: a more liberal one treating each contrast in a given GWAS panel as a separate hypothesis and correcting for the 17 different phenotypes, and a more conservative that corrects for all contrast/GWAS/phenotype combinations in a given dataset. Throughout each dataset we draw attention to examples of signals where we have high confidence and others where we suspect potential confounding, as well as highlighting our results for previously studied cases of polygenic adaptation.

Human Genome Diversity Project and Thousand Genomes Project

We ran a total of 2,040 polygenic score association tests using the Human Genome Diversity Project and Thousand Genomes Project combined dataset. Figure 3.11A summarizes our results for all contrasts, phenotypes, and GWAS panels, when including 40 sample PCs in the GWAS model. 17 association tests were significant (Table 3.4) at our more liberal threshold, with 6 tests occurring in independent panels that should be protected from stratification. Of the remaining 11 tests, only 3 tests have $R^2_{40}/\hat{\gamma}_{F_{Gr}} > 0.8$.

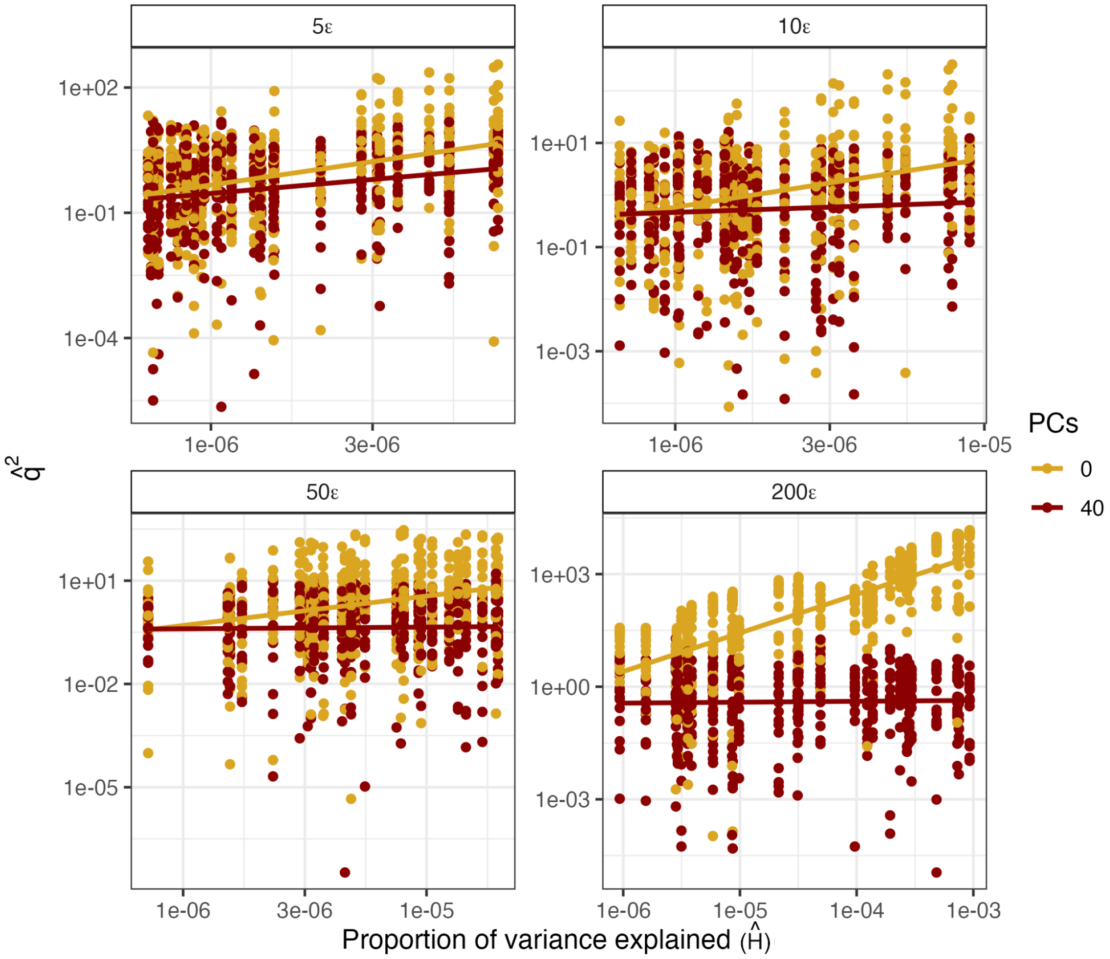

Figure 3.10: **Confounding increases with the proportion of variance explained.** Each panel plots both the uncorrected estimate of $\hat{q}^2$ and $\hat{q}^2$ corrected for 40 sample PCs for all 30 sets of contrasts as a function of the average proportion of variance explained by each contrast in the GWAS panel with that sampling distance.

The most well-protected significant test ($R^2_{40}/\hat{\gamma}_{F_{Gr}} = 0.99$) is the difference in polygenic scores for glucose levels between non-Finnish Europeans and American continental-level ancestry groups (Figure 3.11B). In Figure 3.11C we also highlight a putative false positive result where there is a positive association between mean corpuscular volume polygenic scores and longitude in Europe when the effect sizes were estimated in the 5ϵ GWAS panel. The value of $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ was 0.60, indicating that this test was not well protected from stratification bias by the included PCs. The uncorrected $\hat{q}$ was significantly less than zero, similar to the corrected test, suggesting that uncorrected stratification might be slightly downwardly biasing this association. When we increase the sampling distance, the better protected the test is ($R^2_{40}/\hat{\gamma}_{F_{Gr}} = 0.9$ and 0.99 for 50ϵ and 200ϵ, respectively) and there is no longer a signal of divergence.

The strongest association, and the only one to pass the global threshold is the previously discovered height polygenic score divergence between Sardinian and mainland non-Finnish European individuals (see Figure 3.12). We see a lower polygenic score in Sardinians in all of our analyses with a significant divergence in three of four of our GWAS panels (50ϵ is not significant after multiple testing correction). In Figure 3.12, we show that the $\hat{q}$ values are similar across panels despite the direction of the difference, using the uncorrected effect sizes, switching from 50ϵ to 200ϵ. The values of $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ range from 0.56 to 0.89 (see Figure 3.12B) so although none of the tests are fully protected from confounding, the fact that the direction of confounding switches while $\hat{q}$ (including 40 PCs) remains similar across panels lends additional support. Follow-up analyses could verify the signal by including additional PCs to increase $R^2_J/\hat{\gamma}_{F_{Gr}}$ or by testing additional GWAS panels for a higher $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ for this contrast.

The positive association between polygenic scores for height and latitude in Europe is perhaps the most well-studied case of polygenic adaptation in humans. In 2019, both Berg et al.[36] and Sohial et al.[72] demonstrated that the original signal was overestimated, likely due

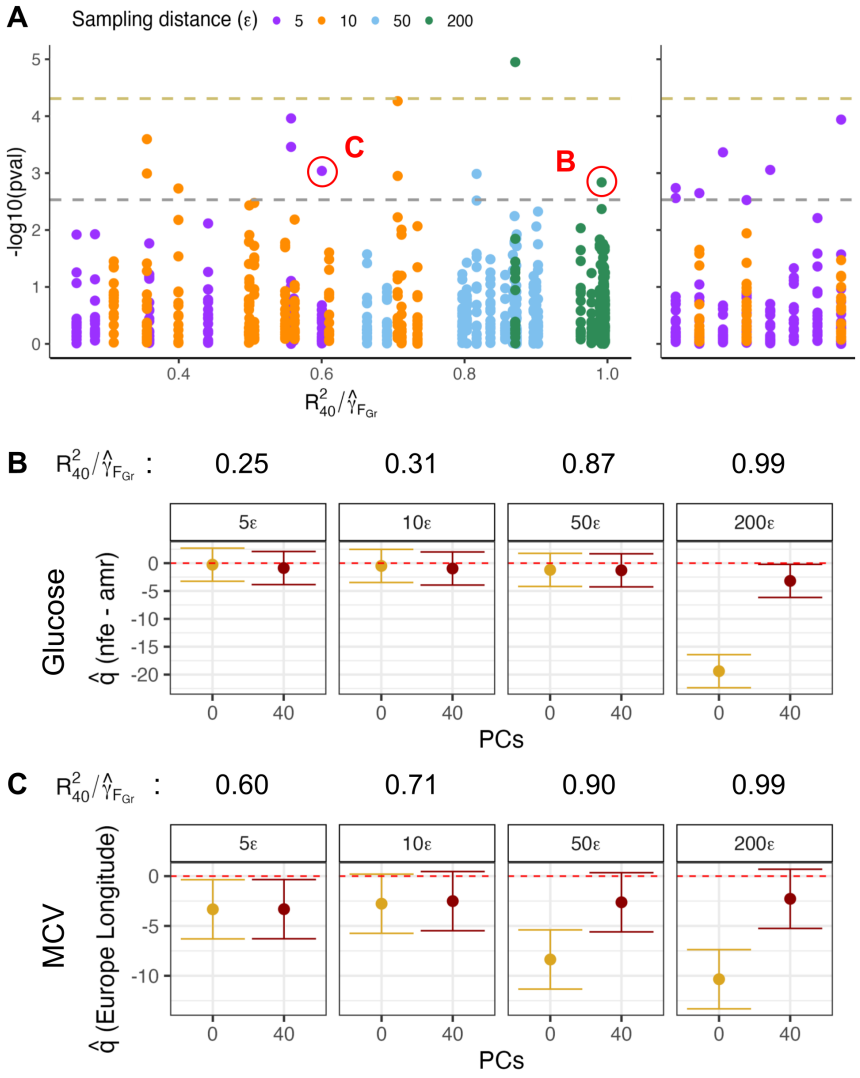

Figure 3.11: **Polygenic association test results for 17 phenotypes in the HGDP and 1kGP combined dataset.**

(A) Manhattan plot for the results of association tests for 15 contrasts with effects sizes for 17 phenotypes estimated in four different GWAS panels. All GWASs included 40 sample PCs. For the left panel, the x-axis is $R^2_{40}/\hat{\gamma}_{FGr}$ with values closer to 1 indicating better protection from potential confounders. The right panel contains the independent contrasts GWAS pairs. The gray dotted line represents our more liberal significance threshold that treats each contrast/GWAS panel pair as a separate hypothesis. The gold dotted line represents our more conservative global threshold that corrects for all contrast/GWAS panel/phenotype groupings in the dataset. (B) Glucose polygenic scores differences between non-Finnish Europeans and American individuals across all four GWAS panels. There is strong statistical support for lower glucose levels in non-Finnish Europeans. (C) The covariance between mean corpuscular volume polygenic score differences and longitude in European samples. There is a significant association in the 5ϵ panel where the test is not well protected from stratification bias.

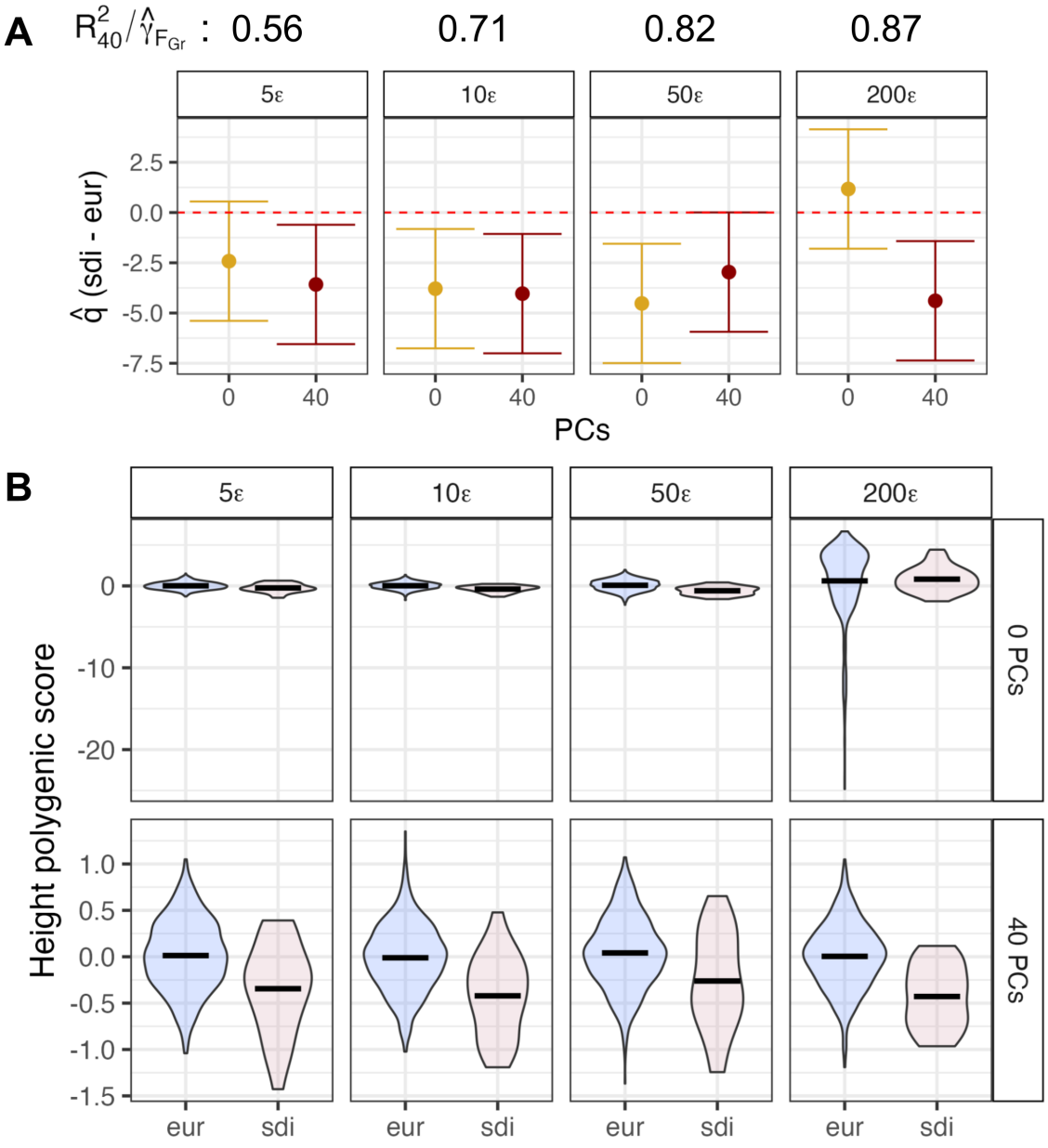

Figure 3.12: **Evidence for decreased height polygenic scores in Sardinian individuals compared to mainland Europeans.**

(A) We computed the covariance between allele differences between Sardinia and mainland Europe and GWAS effect sizes for height estimated in our four different GWAS panels. When 40 PCs are included in the GWAS, $q \approx -4$ across all panels indicating that trait-associated alleles are at lower frequency in Sardinia. However, none of the tests are fully protected from potential confounders with the largest $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ equal to 0.87. (B) Each row shows the distribution of height polygenic scores in each group without any control for structure (top row) and with 40 PCs included (bottom row)

to uncorrected stratification bias in the GIANT height effect sizes estimates. Both studies find the association to be either very weak or completely absent, depending on the exact methodology. Here, we test for association of height polygenic scores with latitude in non-Finish European individuals in all four GWAS panels. We find that $\hat{q}$ is positive in all four panels, but the signal is non-significant after multiple testing correction (Figure 3.13). In the 200ϵ panel, where the test is best protected ($R^2_{40}/\hat{\gamma}_{F_{Gr}} = 0.98$), we find that $\hat{q} = 1.9$ (adjusted $p = 0.057$). Our results support the conclusions from Berg et al. and Sohail et al. in that we observe the same direction of signal, but no statistical support in this test panel dataset.

Originally reported in Berg et al. (2017) [69], Berg et al. (2019) also investigated a signal of decreasing height polygenic scores from east to west along the Eurasian continent. While the original signal (detected using effect sizes estimated in GIANT) was strongly overestimated, when using effect sizes from the UKBB the authors find a weakly significant signal in the same direction. Here we also find a negative relationship between height polygenic scores and longitude using 40 PCs as population structure correction, but the relationship is not significant in any of the GWAS panel (Gigure 3.14), including in the 200ϵ where stratification is well controlled for ($R^2_{40}/\hat{\gamma}_{F_{Gr}} = 0.99$).

British Isles country of birth

All 10 country of birth contrasts have significant overlap with all four GWAS panels with the narrowest sampling distances capturing nearly as much variance as the most diverse panels (see Figure 3.7). Additionally, $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ for each contrast is similar across sampling distances with most values clustering around 0.8-0.85. Of the 680 polygenic score association tests we conducted, we find no globally significant results and only 6 that pass our more liberal threshold (see table 3.5 and Figure 3.15A). For all six tests, $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ ranges from 0.77 to 0.86 and none of the tests replicated in another GWAS panel. In Figure 3.15B and 3.15C, we plot two of the significant tests, polygenic score divergence for triglycerides and

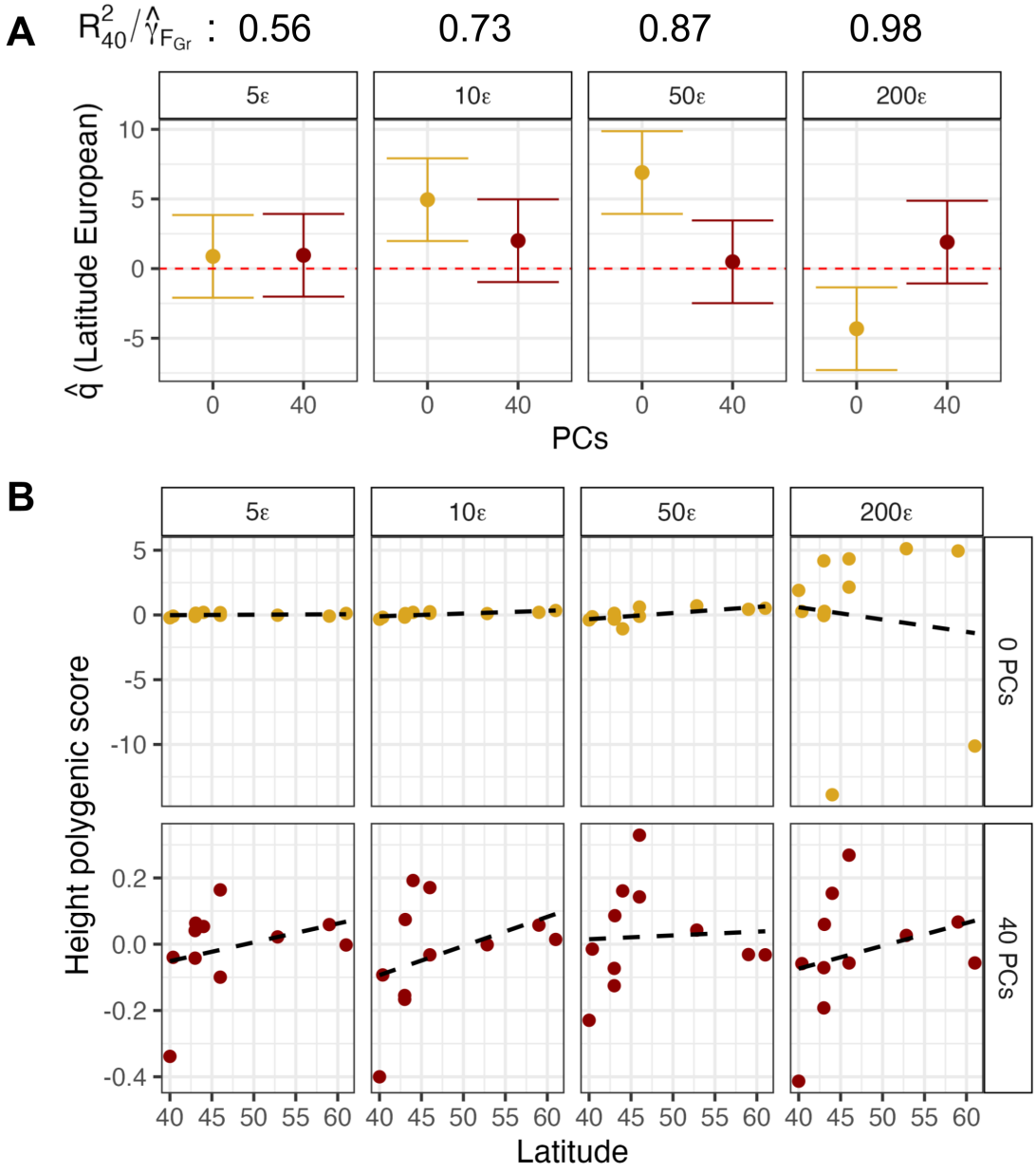

Figure 3.13: **No significant relationship between height polygenic scores and latitude in Non-Finnish European samples.**

(A) Each column shows the covariance between latitude of samples from Europe and polygenic scores for height calculated from effect sizes estimated in each of the four GWAS panels with different sampling distances. The 200ϵ panel is well protected from stratification and indicates a positive, but non-significant relationship. (B) Each row shows the relationship between latitude and the average height polygenic score per sub-population for uncorrected polygenic scores (top row) and those that include 40 PCs (bottom row).

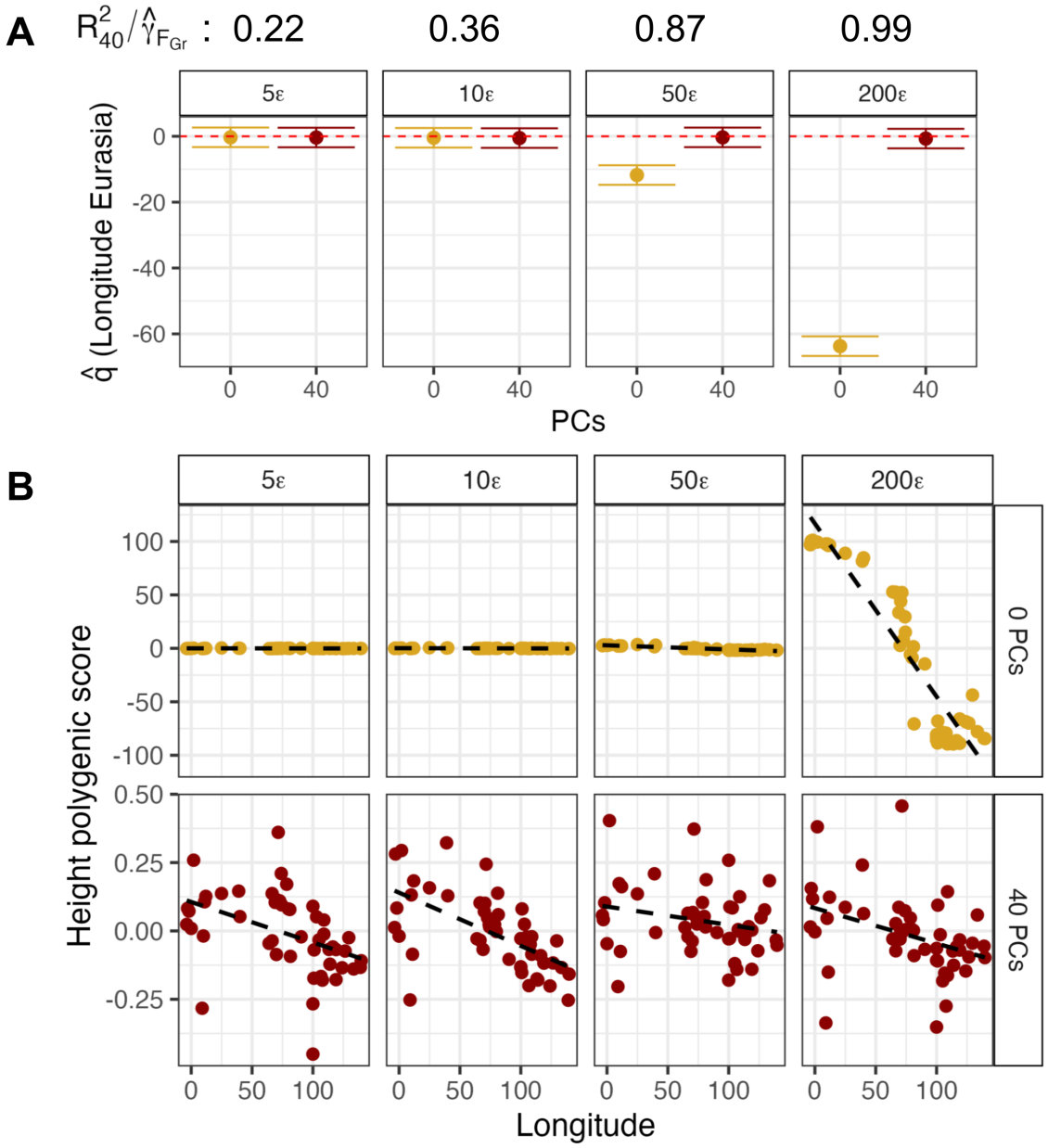

Figure 3.14: **No significant relationship between height polygenic scores and longitude in Eurasian samples.**

(A) We compute $\hat{q}$ using longitude of Eurasian samples and polygenic scores for height. Each column displays our results for both the uncorrected effect sizes (PCs = 0) and the effect sizes corrected with 40 sample PCs across all 4 sampling distances in the GWAS panel. The larger the sampling distance, the better protected the test is by 40 PCs (larger $R^2_{40}/\hat{\gamma}_{F_{Gr}}$), and the larger the magnitude of confounding along the east-west axis. (B) We plot the average height polygenic score for each subpopulation as a function of longitude with the dotted line as the line of best fit.

mean corpuscular hemoglobin concentration (MCHC) between individuals born in Scotland and those born in Wales. The triglyceride polygenic score difference is significant in the 50ϵ panel (adjusted $p = 0.01$), but not in any other panel. All panels have $R^2_{40}/\hat{\gamma}_{F_{Gr}} \approx 0.77$ but have varying magnitudes and directions of confounding. The uncorrected $\hat{q}$ in the 50ϵ panel also has a significant negative value and we hypothesize that residual confounding is downwardly biasing the 40 PC $\hat{q}$ in this panel. We observe a similar pattern with the 5ϵ panel for MCHC polygenic score divergence with a significantly positive uncorrected $\hat{q}$ and $\hat{q}$ when 40 PCs are included.

We note that all of our significant results are for three phenotypes (mean corpuscular volume, MCHC, and triglycerides) that relate to blood. While this is a potentially interesting result that might warrant further follow-up, we observe very limited statistical support for any of our observations. None of the significant results were fully protected from stratification bias, and none of them replicated across panels where the patterns of confounding varied. Therefore, with our current analyses, we cannot rule out that these results are false positives driven by residual stratification.

Ancient DNA contrasts

Next, we analyze the polygenic association test results for the three sets of contrasts from Le et al. (2022). These contrasts are designed to capture selection post admixture for three different epochs (European Neolithic, Bronze Age, and Historical). The original authors find evidence for selection acting on 39 different traits across all three epochs using effect sizes estimated in Biobank Japan. Of the 17 phenotypes we tested, all of which overlapped with the 220 phenotypes they tested, we find no globally significant tests and only 3 tests that pass our more liberal threshold. First, we see a marginally significant signal for selection on trait increasing-alleles for mean corpuscular volume and selection on trait-decreasing alleles for systolic blood pressure in the Historical period. Le et al. find no signal for mean corpuscular

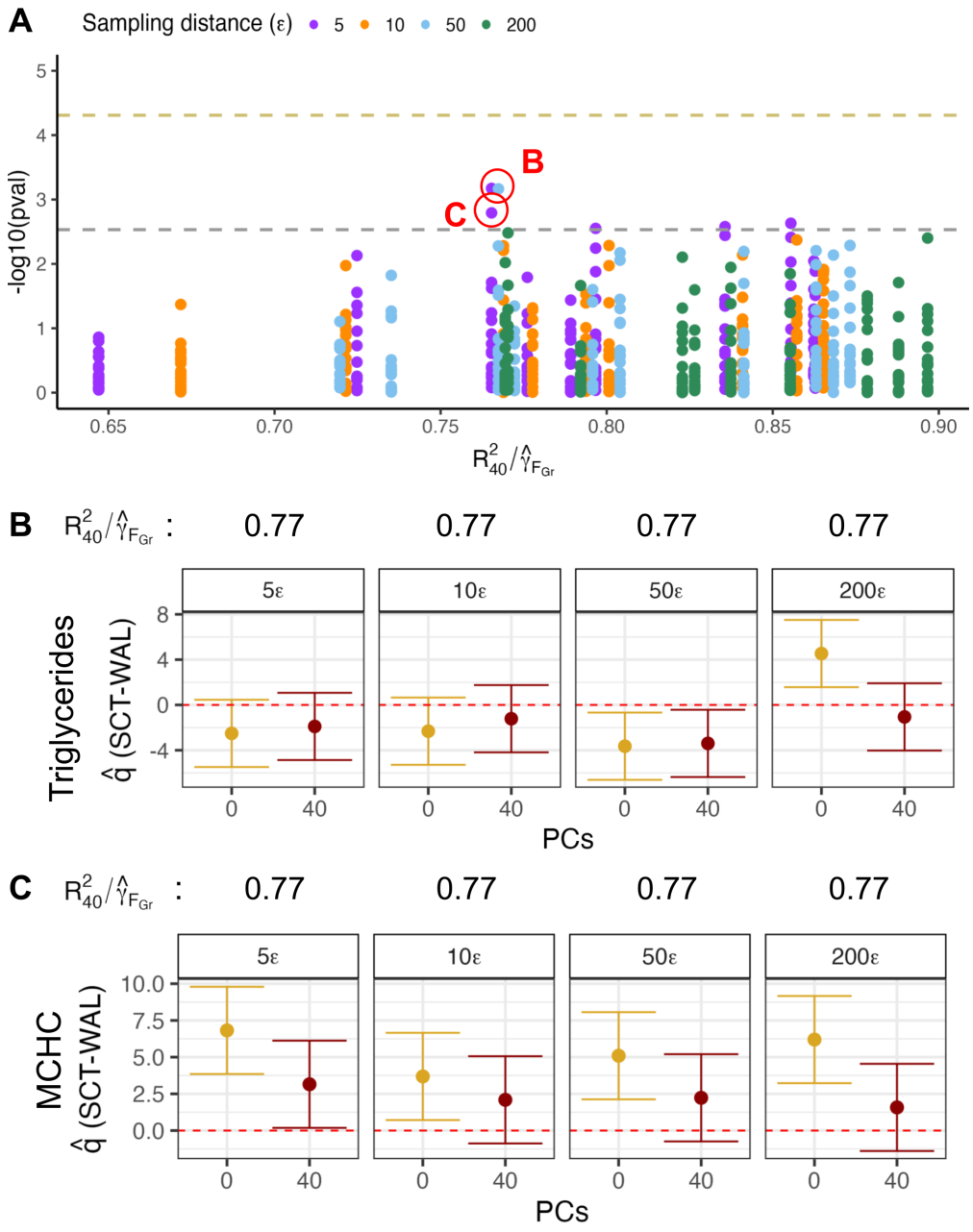

Figure 3.15: **Limited evidence for polygenic score divergence between countries of birth in the British Isles.**

(A) Manhattan plot for the results of association tests for 10 contrasts with effect sizes for 17 phenotypes estimated in four different GWAS panels plotted against $R^2_{40}/\hat{\gamma}_{F_{Gr}}$. All GWASs included 40 sample PCs. (B) Polygenic score for triclycerides differences between individuals born in Scotland and those born in Wales in all four panels. (C) Polygenic score differences for mean corpuscular hemoglobin concentration for the same Scotland-Wales contrast.

volume and but they do find evidence for selection on trait-*increasing* alleles for systolic blood pressure for the same epoch. Additionally, we find evidence for selection on trait-decreasing alleles for systolic blood pressure in the European Neolithic epoch whereas the original authors do not see a signal.

There are a lot of differences between our study design and Le et al. (2022), including differences in GWAS sample size, ascertainment process, p-value threshold, and ancestry of the GWAS panel. Notably, the Biobank Japan GWAS has $\sim 80,000$ more ($n = 179,000$) samples and we may simply be underpowered to detect their observed selection signals. The three selection signals we do detect occur in independent panels that should be immune to stratification. However, none of the signals replicate across panels and are not globally significant. Overall, more work is needed to directly replicate the study design of Le et al. in the context of our approach to stratification.

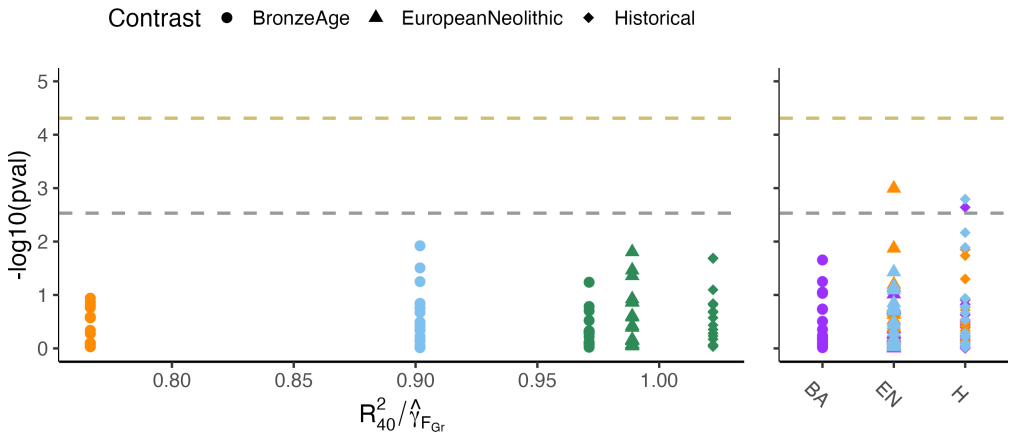

Figure 3.16: **No replication of selection signals in ancient DNA contrasts.** Manhattan plot for the results of association tests for 3 contrasts with effects sizes for 17 phenotypes estimated in four different GWAS panels plotted against $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ in the left panel and by the contrast type in the right panel (independent contrasts/GWAS pairs). Only 3 tests pass our liberal threshold, none of which were discovered (either in the same direction or at all) in the original publication.

Singleton Density Scores

Finally, we ran polygenic score association tests for all phenotypes using two sets of singleton density scores, one estimated in a sample of individuals from the UK[34] and one estimated in a sample of individuals from the Han Chinese ancestry group[110]. We find six significant associations at our lower threshold (see table 3.7 and Figure 3.17A), all of which came from the UK10K contrasts. The strongest signal we find, and the only one to pass the global threshold, is for height (Figure 3.17B) which is significantly positive in all but the 50ϵ GWAS panel. As discussed above, multiple studies, including the original Field et al. (2016) publication, found a highly significant correlation between SDS values and GWAS effect sizes for height-associated alleles. Berg et al. (2019) and Sohail et al. (2019). found either a greatly attenuated signal or no signal at all depending on the GWAS panel used. However, Howe et al. (2022) recovered a significant signal using effect sizes estimated in a sibling GWAS, the gold standard for estimating direct effects immune from stratification bias. In our analyses, we also observe the positive signal using effect sizes estimated in a well protected GWAS panel where $R^2_{40}/\hat{\gamma}_{F_{Gr}} = 0.97$. We also observe a significant positive relationship between SDS values and body weight effect sizes in the same well-protected GWAS panel.

Outside of height and body weight, the other two signals we find are for increased basophil percentage and decreased total protein levels, both in the 50ϵ with $R^2_{40}/\hat{\gamma}_{F_{Gr}} = 0.9$. In Figure 3.17 we plot $\hat{q}$ for basophil percentage. By comparing $\hat{q}$ from using no PCs across panels, it appears that confounding increases in the same direction as the putative signal as sampling distance increases from 5ϵ to 50ϵ, whereas at 200ϵ, where we have the best protection against stratification, the confounding changes direction, and the signal is no longer significant. Therefore we hypothesize that the signal for 50ϵ is being driven by confounding that is not large enough to cause a false positive in 5ϵ or 10ϵ and is corrected for in 200ϵ. We observe a similar phenomenon for total protein levels.

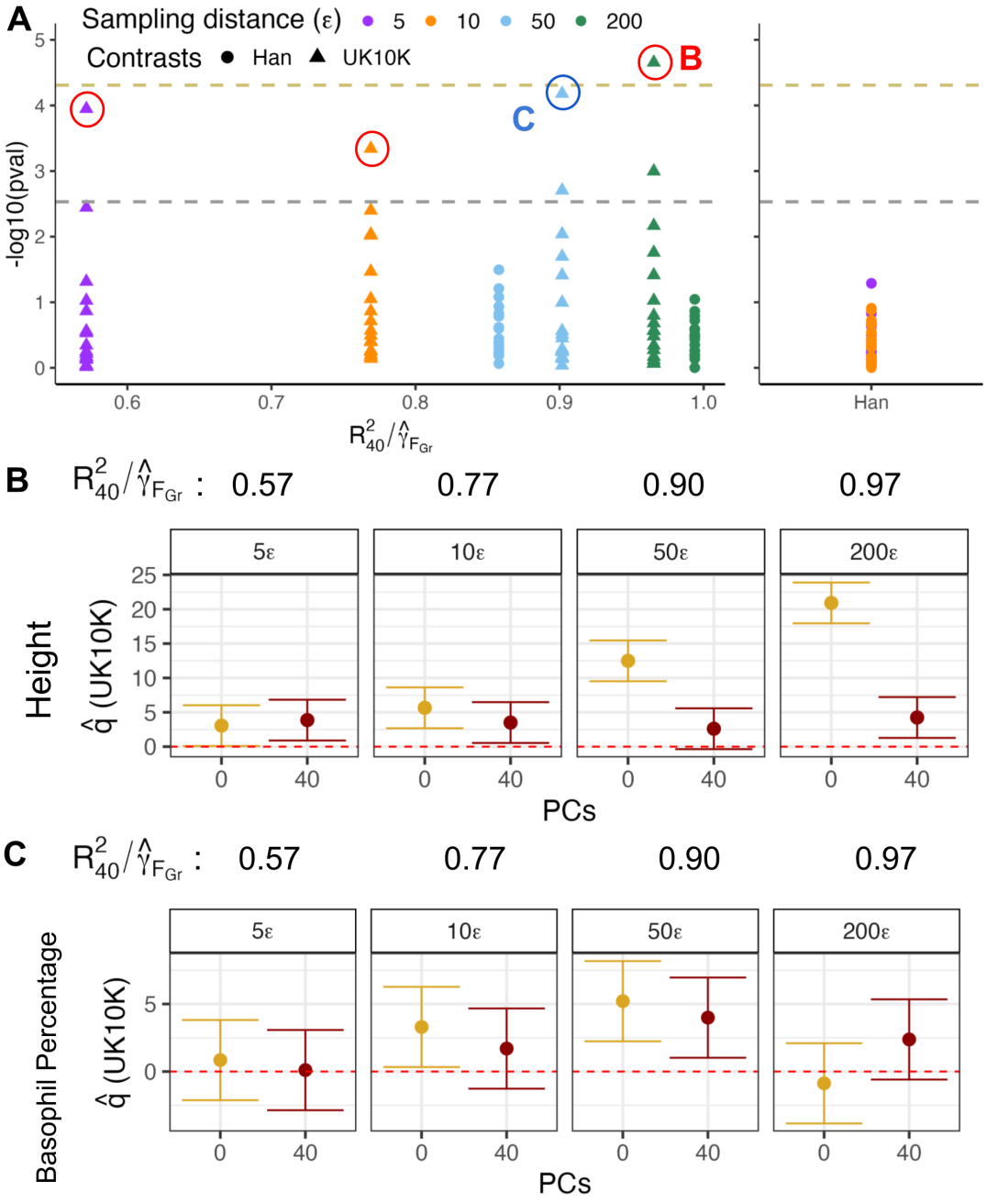

Figure 3.17: **Evidence for recent selection on height using singleton density scores.** (A) Manhattan plot for the results of association tests for 2 contrasts with effects sizes for 17 phenotypes estimated in four different GWAS panels plotted against $R^2_{40}/\hat{\gamma}_{F_{Gr}}$. The Han contrasts were determined to be independent from the 5ϵ and 10ϵ panels. (B) We find strong evidence for a positive association between SDS values and trait-increasing alleles for height with a significant association in the best protected 200ϵ panels where $R^2_{40}/\hat{\gamma}_{F_{Gr}} = 0.97$. (C) The association between trait-increasing alleles for basophil percentage and SDS values in the 50ϵ panel is a putative false positive.

3.5 Conclusion

In polygenic score association tests the goal is to test for a relationship between an ancestry gradient in the prediction panel and the additive component of genetic variation associated with a trait of interest. Therefore it is a concerning observation that different sets of GWAS effect sizes for the same trait produce polygenic scores whose relationships vary widely in both magnitude and direction with major ancestry gradients within the prediction panel[53,113,36,72]. How much of this variation is driven by uncorrected effect sizes vs other factors is still an open question and one that cannot be answered until stratification can be ruled out.

In this Chapter we built on results from Chapter 2 to develop and test approaches to control for stratification bias in practice, where population structure covariates must be estimated from genetic data. We first focus on PCA, where, in order to sufficiently protect against stratification, the population PCs must capture $\tilde{F}_{Gr}$ (or the confounders), and the sample PCs must be accurate estimates of the population PCs. In our simulations, the GWAS panel sample size was orders of magnitude smaller than typical biobank scale GWASs. Consequently, the error lower bound was large for all but the first few PCs and the performance of the PCs as population structure estimates suffered. If we scaled up our simulations, we would expect the performance of PCs to improve. Our simulation results highlight the importance of investigating estimation error in PCs, as it is crucial to understand how many PCs are detectable and how many are accurately estimated, especially when working with real data. For example, one explanation for the residual stratification in the GIANT height GWAS effect size estimates is that, in the meta-analysis study design, sample PCs are estimated in each individual smaller cohort, and important axes of structure in the combined dataset may have been poorly estimated. In general, polygenic scores built from effect sizes estimated in a meta-analysis are more correlated with ancestry gradients within the prediction panel, suggesting potential stratification bias[113]. Another area of possible

concern is eQTL studies where sample sizes for individual cohorts are typically on the order of hundreds rather than hundreds of thousands. Top PCs are usually included as stratification control, and it is worth investigating how well estimated these PCs are for different datasets. Our results here highlight the importance of developing tools to more robustly estimate the error in population structure estimates.

One benefit of our direct approach is that it is straightforward to estimate error directly using a block jackknife approach. In both our simulation studies and empirical analysis the larger the variance explained by $\tilde{F}_{Gr}$ in the GWAS panel, the better the estimate of $\hat{F}_{Gr}$. In our empirical analysis, the only contrasts where we had less than 10% error were for continental-level contrasts in the most diverse GWAS panel. This is the precise situation for which we expect PCs to also be successful in capturing the relevant axis of structure. However, we can utilize even noisy estimates of $\hat{F}_{Gr}$ to estimate how well protected the test is by a given number of sample PCs by calculating $R^2_J/\hat{\gamma}_{F_{Gr}}$.

So, how well protected were our association tests by 40 sample PCs? First, we noticed that the first 5-10 PCs explain a large proportion of the variance in $\tilde{F}_{Gr}$ with only marginal increases including higher PCs. Some patterns are easily explainable, for example in the 200ϵ panel the first PC explains nearly all the variance for the non-Finish European/African ancestry allele frequency differences. In contrast, the British Isles country of birth contrasts require a few more PCs to reach a point of diminishing returns. Even with these contrasts, which capture subtle structure within the biobank, it appears that including 10 PCs is not very different than including 40 PCs in terms of protecting those specific contrasts. This was surprising to me, as I expected higher PCs to contribute more for subtle axes of structure. It is possible that a higher PC we did not calculate will significantly increase the percent of variance explained (e.g PC 41 explains the remaining 20% of variance) but given what we observe it seems more likely that the remaining variance is spread out along a lot of PCs and may even be undetectable. The observation that the country of birth contrasts, which

should have complete overlap with the GWAS panel, are not fully protected in any panel suggests that perhaps, for these contrasts, other estimators of population structure might be more useful. Notably, our framework is not limited to common variant PCs and could be applied to any method that represents patterns of population structure in terms of a factor model. For example, previous work has suggested that PCs of rare variants or local ancestry assignments might successfully capture more subtle structure, and our framework could be used to assess the extent to which this is true in any given case. More generally, as new population structure estimates are developed we hope that our method will allow researchers to test multiple estimators to find the optimal combination that best protects their contrast of interest.

We also note that for all contrasts, besides the country of birth contrasts, $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ was highest for the most diverse GWAS panel. Conventional wisdom suggests more homogeneous GWAS reduce stratification. While homogeneous panels may reduce the scale of confounding, which we see evidence for in Figure 3.10, they also seem to make it more difficult to completely protect contrasts from stratification bias with common PCs. In this case we have two opposing forces. Broader sampling tends to increase $R^2_J/\hat{\gamma}_{F_{Gr}}$ while also increasing confounding, potentially making the fraction of unexplained confounding larger, despite a larger $R^2_{40}/\hat{\gamma}_{F_{Gr}}$. It is difficult to quantify this phenomenon without knowledge of specific confounders. Here we have taken the uncorrected $\hat{q}$ as a proxy for confounding and in future work we could develop a more sophisticated approach that uses effect sizes from non-associated SNPs (frequency matched to associated SNPs) as a measure of confounding along r in the test panel.

However, even using our naive proxy, computing both the uncorrected $\hat{q}$, the $\hat{q}$ including 40 PCs, and $R^2_{40}/\hat{\gamma}_{F_{Gr}}$ we were able to use effect sizes estimated in multiple panels to generate additional lines of support (or lack thereof) for observed signals. The strongest signal in any of our datasets was for height polygenic score divergence between Sardinian and mainland

Europeans. Despite the fact that this contrast was not completely protected in any GWAS panel ($R^2_{40}/\hat{\gamma}_{F_{Gr}}$ ranged from 0.56 to 0.87), the value of $\hat{q}$ including 40 PCs was consistent, even as the values of the uncorrected $\hat{q}$ varied in direction across panels. This observation, along with independent evidence from Chen et al. (2020)[73], suggests that this signal is likely driven by true divergence, even though stratification cannot be ruled out in our analysis. This example demonstrates the utility of replicating polygenic score association tests using effect sizes estimated in different GWAS panels where the relationship between ancestry and environment varies, an idea I discuss in more depth in Chapter 4.

There is no one-size-fits-all solution to controlling for stratification bias in polygenic score association tests in practice. In this Chapter, guided by solid theoretical results, we have developed new tools to tackle this challenge. At first, I had hoped that including our direct estimator $\hat{F}_{Gr}$ in conjunction with PCA and/or LMMs in the GWAS model would provide a guarantee that the test of interest was completely protected from residual bias. Upon further investigation, and testing our approach in empirical data, we realized that for many of our contrasts of interest, $\hat{F}_{Gr}$, was noisy and did not add much additional protection on top of PCA. After a deeper exploration of the relationship between PCs, $\hat{F}_{Gr}$, and $\hat{H}$, we developed the $R^2_J/\hat{\gamma}_{F_{Gr}}$ statistic as a way to quantify the amount of variance in the shared axis of structure captured by population structure estimators. Our results demonstrate the need for continued development of new population structure estimates and further emphasis on quantification of uncertainty and bias in both population structure estimates and polygenic scores.

3.6 Author contributions

Jennifer Blanc: Conceptualization, methodology, formal analysis, investigation, writing - original draft, writing - review and editing, visualization, funding acquisition, software, data curation

Jeremy Berg: Conceptualization, methodology, formal analysis, investigation, writing - original draft, writing - review and editing, supervision, project administration, funding acquisition, resources

3.7 Materials and Methods

3.7.1 Simulations

Simulating genotypes

We used *msprime*[114] to simulate genotypes under different models with 100 replicates per model. The first model, shown in Figure 3.1, 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. We then sample 5,000 individuals per population and create two configurations of GWAS and test panels ($N, M = 10,000$) based on the diagrams in Figure 3.1A and Figure 3.1C. For every model replicate we simulate a large number of independent sites and down-sample to $L = 10,000$ SNPs with MAF $>$ 0.01 in both GWAS and test panels. We use these genotype simulations for Figure 3.1 and Figure 3.18. When the populations in the GWAS and test panel are non-sister (i.e., Figure 3.1A) the average within panel F_{ST}[115] was 0.01, whereas in the configuration in Figure 3.1C the average F_{ST} was 0.005.

For Figure 3.2 we use the same model setup but adjust the split times to 12/0, 12/4, and 12/10 generations in the past for population models A, B, and C, respectively. The average F_{ST} for the overlapping structure scenario is approximately 0.0006. To reduce computational burden, we scale down the sample size to 1,000 individuals per panel (500 per population). We simulate large number of independent SNPs and down-sample to L sites (MAF $>$ 0.01 in both panels) which we vary from 500 to 100,000.

For Figure 3.3 we use a model, modified from Zaidi and Mathieson 2020[47], that is a 6×6

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

Simulating phenotypes

To study the effect of environmental stratification on association tests, we first simulated non-genetic phenotypes for an individual i in the GWAS panel as $y_i \sim N(0,1)$. In our discrete 4 population models we then generate a phenotypic difference between populations by adding Δ_{AB} to y_i for individuals in population B. For Figure 3.1 we vary Δ_{AB} from 0 to 0.1 standard deviations. In order to compare across models and values of $\frac{M}{L}$ in Figure 3.2 we compute Δ_{AB} as $\frac{5000}{0.05 \times L}$.

In our grid simulations we generated three different phenotypic gradients where the largest phenotypic shift is always equal to Δ. To generate a latitudinal gradient (Figure 3.3A), 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 3.3B, 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 3.3C we shifted the phenotype of individuals in deme (1,4) by Δ. For all grid simulations in Figure 3.3 we set $\Delta = 0.2$. In order to compare across values of L in Figure 3.4 we compute Δ as $\frac{60}{0.015}$

To study the effect of controlling for stratification in cases where there is a true signal of association between polygenic scores and the test vector (Figure 3.18), we used our 4 population demographic model and followed the protocol outlined in Zaidi and Mathieson (2020)[47] to simulate a neutral trait with $h^2 = 0.3$. We first randomly select 300 variants to be causal and sample their effect sizes from $\beta_\ell \sim N(0, \sigma_i^2[p_\ell(1-p_\ell)]^\alpha)$, 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 300 causal sites in the test panel and flip the sign of the effect sizes in the GWAS panel such that $p_D - p_C > 0$ and $\beta_\ell > 0$ with probability θ. θ therefore controls the strength of the association with $\theta = 0.5$ representing no expected association and $\theta = 1$ representing the most extreme case where trait increasing alleles are always at a higher frequency in population D. We use θ ranging from $0.5 - 0.62$. We then draw the environmental component of the phenotype $e_{i,k} \sim N(0, 1-h^2)$ and generate an environmental confounder by adding $\Delta_{AB} \in \{-0.1, 0, 0.1\}$ to $e_{i,k}$ for individuals in population B.

Computing covariates

For each polygenic score association test we computed $\hat{F}_{Gr}$. We first construct T as either population ID, latitude or the single deme of interest, depending on the simulation. Given this test vector, we compute $r = \mathbf{X}^\top T$ using the plink2[116] function `--glm`. Finally we compute $\hat{F}_{Gr}$ (see Equation 3.2) using `--sscore` in plink2, taking care to standardize by the variance in the GWAS panel genotypes. Additionally we used plink2 `--pca` or `--pca approx` to compute sample PCs from the GWAS panel genotype matrix.

GWAS

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

1. $y = \beta_\ell G_\ell + \epsilon$
2. $y = \beta_\ell G_\ell + \omega \hat{F}_{Gr} + \epsilon$

3. $y = \beta_\ell G_\ell + \omega_1 \hat{U}_1 + ... + \omega_j \hat{U}_j + \epsilon$.

Additionally, we conducted a fourth GWAS, $y = \beta_\ell G_\ell + \omega \tilde{F}_{Gr} + \epsilon$, for the discrete 4 population model where $\tilde{F}_{Gr}$ is known. All GWASs were done using the plink2[116] function `--glm`.

We then ascertain S SNPs based on minimum p-value for inclusion in the polygenic score. For Figure 3.1 and Figure 3.3 we set $S = 300$. In order to compare across values of $\frac{M}{L}$ in Figure 3.2 and Figure 3.4, we set $S = 0.05 \times L$ and $S = 0.015 \times L$, respectively. For Figure 3.18 we use estimated effect sizes at the 300 causal sites rather than ascertaining based on p-value.

Polygenic score association test

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

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

Estimating the error in our population structure estimators for the grid model

Direct estimator

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

$$\hat{F}_{Gr,ij} = \left(\hat{F}_{Gr,ij} - \overline{\hat{F}_{Gr,j}}\right) + \left(\overline{\hat{F}_{Gr,j}} - \tilde{F}_{Gr,j}\right) + \tilde{F}_{Gr,j} \tag{3.35}$$

where $\overline{\hat{F}_{Gr,j}} = \frac{1}{m_j} \sum_i^{m_j} \hat{F}_{Gr,ij}$ is the empirical average of $\hat{F}_{Gr,ij}$ within deme j (m_j is the number of individuals in deme j), and $\tilde{F}_{Gr,j}$ is the entry of the true population structure axis

$\tilde{F}_{Gr}$, for all individuals in deme j. Individuals within demes are exchangeable in our model, so the deviations $\left(\hat{F}_{Gr,ij} - \overline{\hat{F}_{Gr,j}}\right)$ and $\left(\overline{\hat{F}_{Gr,j}} - \tilde{F}_{Gr,j}\right)$ both represent sources of error in our estimator. The fraction of variance in $\hat{F}_{Gr}$ that is attributable to error is therefore

$$\text{error} = \frac{\mathbb{E}_j\left[Var_i\left(\hat{F}_{Gr,ij} - \overline{\hat{F}_{Gr,j}}\right)\right] + Var_j\left(\overline{\hat{F}_{Gr,j}} - \tilde{F}_{Gr,j}\right)}{Var\left(\hat{F}_{Gr}\right)}. \tag{3.36}$$

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

$$\frac{1}{H}\sum_{h}^{H}\frac{1}{J}\sum_{j}^{J}\frac{1}{m_j - 1}\sum_{i}^{m_j}\left(\hat{F}_{Gr,ijh} - \overline{\hat{F}_{Gr,jh}}\right)^2, \tag{3.37}$$

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

$$\overline{\hat{F}_{Gr,jh}} = \frac{1}{m_j}\sum_{i}^{m_j}\hat{F}_{Gr,ijh} \tag{3.38}$$

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

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

$$\frac{1}{J}\sum_{j}^{J}\frac{1}{H-1}\sum_{h}^{H}\left(\overline{\hat{F}_{Gr,jh}} - \frac{1}{H}\sum_{\ell}^{H}\overline{\hat{F}_{Gr,j\ell}}\right)^2. \tag{3.39}$$

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

The denominator, in turn, can be estimated straightforwardly as

$$\frac{1}{M-1}\sum_{i}^{M}\left(\hat{F}_{Gr,i}-\frac{1}{M}\sum_{\ell}^{M}\hat{F}_{Gr,\ell}\right)^2 \tag{3.40}$$

where we now use ℓ to index individuals within the mean calculation. Our estimate of the error is then given by summing 3.37 and 3.39 and dividing by 3.40.

Principal components

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

$$\text{error}_k > \mathbb{E}_j\left[Var_i\left(\hat{U}_{ijk}-\overline{\hat{U}_{jk}}\right)\right], \tag{3.41}$$

which we estimate as

$$\frac{1}{H}\sum_{h}^{H}\frac{1}{J}\sum_{j}^{J}\frac{1}{m_j-1}\sum_{i}^{m_j}\left(\hat{U}_{ijkh}-\frac{1}{m_j}\sum_{\ell}^{m_j}\hat{U}_{\ell jkh}\right)^2. \tag{3.42}$$

3.7.2 Empirical Analyses

Computing allele frequency vectors

Human Genome Diversity Project and Thousand Genomes project

We downloaded VCF files from the combined Human Genome Diversity Project and Thousand Genomes Project dataset[108] and then converted them to plink2 format after

converting the coordinates to the hg19 genome build. As described in Section 3.4.1, we constructed 15 sets of allele frequency contrasts. For each set of contrasts we kept SNPs at greater than 1% MAF in both the test panel and the GWAS panel. For pairwise allele frequency comparisons, we compute the within population allele frequencies and take the difference. For continuous contrasts, we use plink2 `--glm` to regress the test vector on the vector of genotypes.

Ancient DNA contrasts

Le et al. 2022[6] used a maximum likelihood approach to estimate the allele frequency in five different discrete time periods of European history. Following their approach, we use their MLE estimates and calculate three sets of allele frequency contrasts that capture the difference between expected and observed allele frequencies post-admixture in 3 different epochs. We compute the expected allele frequencies via a weighted average,

1. European Neolithic = 0.84 Anatolia Neolithic + 0.16 Mesolithic
2. Bronze Age = 0.52S Steppe + 0.48 European Neolithic
3. Historical = 0.15 European Neolithic + 0.85 Bronze Age

and then subtract the estimated allele frequency to get our contrasts. We restrict to SNPs with greater than 1% MAF for each GWAS panel.

British Isles country of birth

We selected all individuals from the UKBB not belonging to any of our four GWAS panels and whose country of birth was listed as Scotland, England, Northern Ireland, the Republic of Ireland, or Wales (see table 3.2 for sample sizes). We then selected SNPs with greater than 1% MAF in both the test panel and the GWAS panel and computed all pairwise allele frequency differences for a set of 10 contrasts.

Singleton density scores

We downloaded SDS values estimated from Field et al. (2016)[34] and limited to SNPs with SDS values and greater than 1% MAF in the GWAS panel. Additionally, we downloaded SDS values from Luo et al. (2023)[110] and converted the coordinates to the hg19 genome build and took only the SNPs with greater than 1% MAF in the GWAS panel.

Estimating $\hat{F}_{Gr}$ and it's error

For each set of contrasts, we first LD prune the total set of SNPs using `--indep-pairwise 100kb 0.8`. With these L SNPs we calculate $\hat{F}_{Gr} = \frac{1}{\sqrt{(L-1)}} \sum_{\ell=1}^{L} G_{\cdot\ell} r_\ell$ using `--sscore` in plink2.

To get the proportion of overlap between contrasts and panels we compute

$$H = \frac{M-1}{M(L-1)} \hat{F}_{Gr}^{\top} \hat{F}_{Gr} \tag{3.43}$$

where $M = 100,000$ for all GWAS panels and L varies by contrast.

In empirical data $\tilde{F}_{Gr}$ is unknown so we estimated error in $\hat{F}_{Gr}$ using a block jackknife, using $B = 581$ independent LD blocks to get a rough estimate of the error for each individual, and then averaging across all individuals in the GWAS panel to estimate the average error. Specifically, we compute,

$$\hat{e}_{F_{Gr}} = \frac{\frac{1}{M}\sum_i^M \frac{B-1}{B}\sum_{j=1}^{B}\left(\hat{F}_{Gr}^{(-j)} - \frac{1}{B}\sum_i^B \hat{F}_{Gr}^{(-j)}\right)^2}{\frac{1}{M-1}\sum_i^M \left(\hat{F}_{Gr,i} - \frac{1}{M}\sum_\ell^M \hat{F}_{Gr,\ell}\right)^2} \tag{3.44}$$

where the numerator is the average block jackknife estimate of the standard error across individuals, and the denominator is the total variance in our estimator. We then calculate the signal in $\hat{F}_{Gr}$ as $\hat{\gamma}_{F_{Gr}} = 1 - \hat{e}_{F_{Gr}}$. When computing $\hat{\gamma}_{F_{Gr}}$ to calculate the ratio $R_J^2/\hat{\gamma}_{F_{Gr}}$, we only use SNPs on odd chromosomes.

PCA

For each GWAS panel we use a set of 147,604 SNPs selected by the UKBB[43] to do PCA using `--pca approx` in plink2. We extract 40 sample PCs which we later include as covariates in our GWAS analyses. To compute $R^2_J/\hat{\gamma}_{F_{Gr}}$, we repeat PCA only using SNPs on even chromosomes and then calculate the variance in $\hat{F}_{Gr}$ (computed using odd chromosome SNPs) explained by $J = 1$ to $J = 40$ PCs.

GWAS

We conducted two GWAS in each panel for each of the 17 phenotypes listed in table 3.3 using *fastGWA*[27] `--fastGWA-lr`. In the first GWAS, we included age at recruitment (21022), genetic sex (22001), and genotype measurement batch (22000) as covariates with no additional control for population structure. In the second GWAS, we included the same covariates and 40 sample PCs.

Polygenic score association tests

Next, we ascertained SNPs for inclusion in the polygenic score association test by selecting the minimum p-value SNP per approximately independent LD block (1703 blocks as defined by Berisa and Pickrell (2016)[112]) and included the SNP if the GWAS p-value was less than 10^{-4}. We computed a raw polygenic score association test statistic as $\hat{q}_{raw} = \hat{\beta}^\top r$. Finally, we assessed the significance of $\hat{q}_{raw}$ using a block jackknife approach to estimate the standard error of $\hat{q}_{raw}$. Specifically, we estimated the variance of $\hat{q}_{raw}$ attributable to neutral processes as

$$\hat{\sigma}^2 = \frac{B-1}{B}\sum_{i=1}^{B}\left(\hat{q}_{raw}^{(-i)} - \frac{1}{B}\sum_{\ell=1}^{B}\hat{q}_{raw}^{(-\ell)}\right)^2, \tag{3.45}$$

where B is the total number of blocks, and $\hat{q}_{raw}^{(-i)}$ is the estimate of $\hat{q}_{raw}$ obtained when leaving block i out. Motivated by the observation in Berg et al (2019)[36] Appendix 1 that even weak linkage between SNPs can lead to evolutionary covariance causing a miscalibrated null distribution, we combined groups of three neighboring blocks into larger blocks so that in the end we have $B = 581$ non-overlapping block for calculating $\hat{\sigma}^2$. Similar to other subsections in this method section, ℓ also indexes blocks, but we use a different letter to avoid confusion between the two sums. Now, $\hat{q}_{raw} \sim N(0, \hat{\sigma}^2)$, which we used to compute a two tailed p-value for each polygenic score association test. In figures, we plot a standardized version

$$\hat{q}_{std} = \frac{\hat{q}raw}{\hat{\sigma}} \tag{3.46}$$

which has a variance of 1, and we include Bonferroni corrected error bars of $\pm\Phi^{-1}\left(1 - \frac{0.025}{17}\right) = \pm 2.97$, where Φ^{-1} is the inverse of the standard Normal CDF.

3.8 Extended results and supplemental figures

3.8.1 Downward bias with true signal

Expected bias

The direct estimator approach (i.e computing $\hat{F}_{Gr}$ and including it as a covariate in the GWAS) 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 that are used to compute the correction is large relative to the number included in the polygenic score.

To see why, we can rewrite the regression Equation 3.3 from the main text in terms of a sum of the contribution from our focal site and all other sites as

$$y = G_\ell \beta_\ell + \left(\frac{L-1}{L} \hat{F}_{Gr,-\ell} + \frac{1}{L} \frac{G_\ell r_\ell}{G_\ell^\top G_\ell / M} \right) \omega + e \tag{3.47}$$

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 r_ℓ. If r_ℓ is positive, $\hat{\beta}_\ell$ has a slight negative bias, whereas if r_ℓ is negative, the bias will be positive. Similar effects are noted elsewhere in the statistical genetics literature, for example in correcting for PCs of gene expression data[117], and in the use of linear mixed models in GWAS[103], where it has been termed "proximal contamination". Assuming that the variance of $\frac{r_\ell}{\sqrt{G_\ell^\top G_\ell / M}}$ 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.

$$\mathbb{E}\left[\hat{q} \mid q\right] \approx q \left(1 - \frac{S}{L}\right). \tag{3.48}$$

Notably, controlling for sample PCs of the GWAS panel genotype matrix will induce a similar effect if the sample PCs capture $\tilde{F}_{Gr}$. 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[118], 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[119,120]. Further concern about downward biases in applications could likely be ameliorated via the "leave one chromosome out" scheme (we implement this approach in 3.4) commonly implemented in the context of linear mixed models[103,23] 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.

Toy model simulations

Next, we wanted to confirm that including $\hat{F}_{Gr}$ or $\hat{U}_1$ does not regress out true signals of polygenic score divergence, consistent with our theoretical argument above. To do this, we modified our simulations of the confounded topology from Figure 3.1 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 a positive test statistic and is conceptually equivalent to adding a selection event on the internal branch in the population phylogeny.

When there was no environmental stratification (Figure 3.18, middle panel), genetic stratification eventually created an upward bias in the uncorrected $\hat{q}$, as the strength of the divergence signal increased. Similarly, when we added environmental stratification in the same direction as the genetic stratification (Figure 3.18, right panel) the upward bias increased in magnitude. Finally, when we added environmental stratification in the opposite direction (Figure 3.18, left panel), environmental and genetic stratification opposed one

another and the observed bias depended on the strength of each. We then observe that in all cases including $\tilde{F}_{Gr}$, $\hat{F}_{Gr}$, or $\hat{U}_1$ (here $\hat{F}_{Gr}$, and $\hat{U}_1$ are well estimated) as a covariate in the GWAS eliminates bias while still capturing the true association signal. This is expected in all situations for $\tilde{F}_{Gr}$ and when $S << L$ for $\hat{F}_{Gr}$ and $\hat{U}_1$.

3.8.2 Supplemental tables and figures

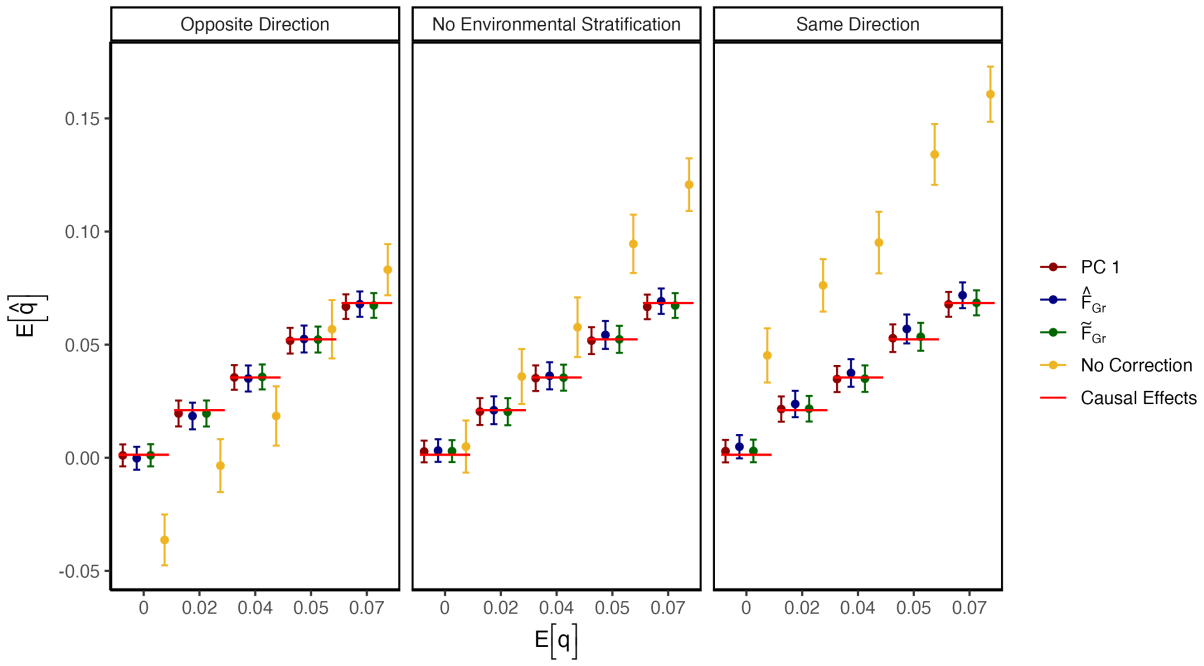

Figure 3.18: **Including $\tilde{F}_{Gr}$, $\hat{F}_{Gr}$, or PC 1 as a covariate in the GWAS model maintains power to detect true association signal.**

GWAS and test panels were simulated in the overlapping structure configuration (see Figure 3.1A). Heritable phenotypes ($h^2 = 0.3$) were simulated with a true difference in polygenic scores by flipping the sign of a proportion of causal effects to align with allele frequency contrasts, $p_{D,\ell} - p_{C,\ell}$, in the test panel. When stratification is in the same direction as the true difference, $\hat{q}$ is upwardly biased, as it is when there is no environmental stratification, once genetic stratification is strong enough. When stratification is in the opposite direction, environmental and genetic stratification are opposed and the direction of bias depends on the strength of each. As expected, $\tilde{F}_{Gr}$ perfectly captures true association regardless of the direction of stratification. Estimators of $\tilde{F}_{Gr}$ (i.e $\hat{F}_{Gr}$ and PC 1) also capture true association, consistent with out theoretical arguments that downward bias is minimal when $S << L$.

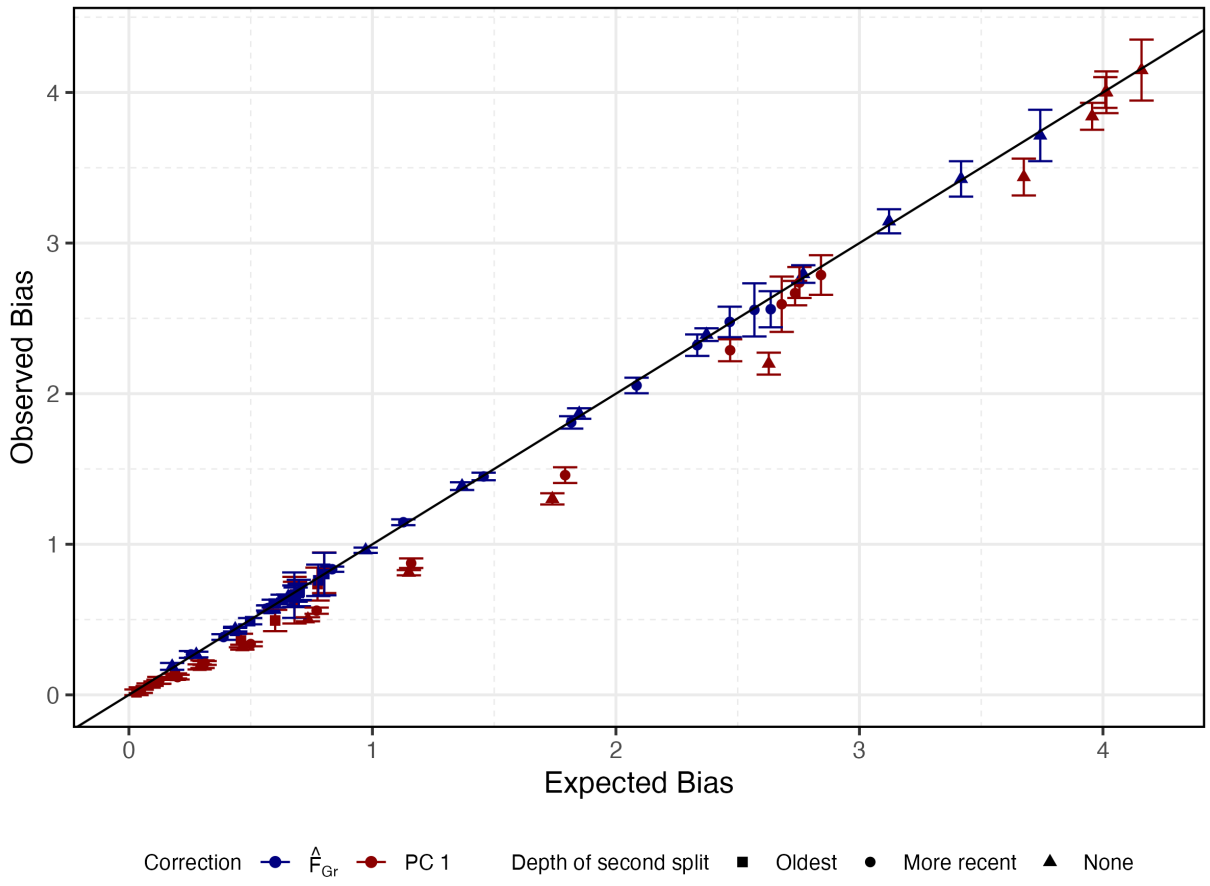

Figure 3.19: **Error in estimates of $\tilde{F}_{Gr}$ predicts bias in $\hat{q}$ across population models.** For all simulations in Figure 3.2 we compute the expected bias as $\mathbb{E}[\text{Error}] \times \mathbb{E}[\hat{q}_{nc}]$ where $\hat{q}_{nc}$ is the observed bias using effect sizes that were estimated with no correction. We then compare this expected bias to the observed bias when using that estimator as a covariate in the GWAS. The error in both $\hat{F}_{Gr}$ and sample PC 1 is highly predictive of the observed bias, though we observe that sample PC 1 exhibits a slight increase in bias reduction compared to the expected.

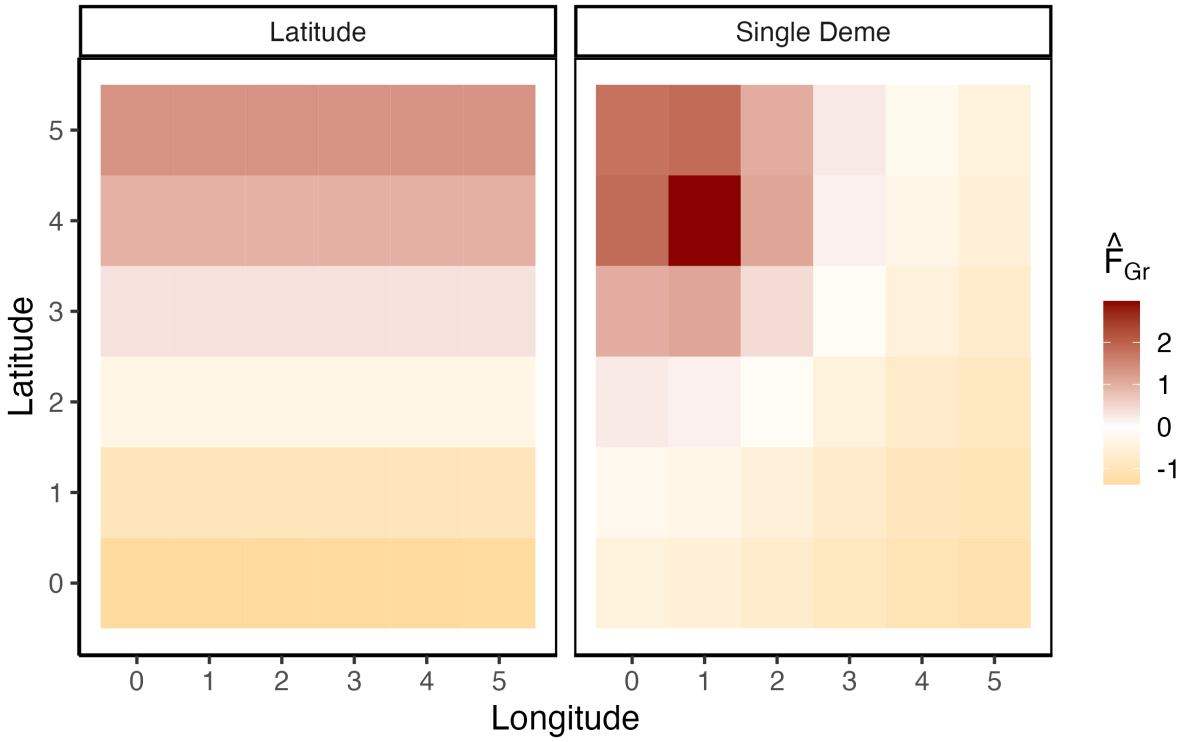

Figure 3.20: $\hat{F}_{Gr}$ **as observed in the GWAS panel.**
For both of the test vectors used in the grid simulations we plotted the average $\hat{F}_{Gr}$ per deme across 100 replicates. For the latitudinal test vector, $\hat{F}_{Gr}$ simply recapitulates latitude, which is unsurprising given the symmetric migration model we use. For the single deme test vector, $\hat{F}_{Gr}$ largely reflects the distance to the focal test deme.

HGDP1kGP Test Panel		
Group	Code	Sample Size
East Asia	eas	825
Africa	afr	1003
South Asia	sas	790
Non-Finish Europe	nfe	689
America	amr	552
Eurasia	eurasia	2224
Europe	eur	510
Sardinia	sdi	27

Table 3.1: **Sample sizes for different groups in the combined HGDP 1kGP dataset.**

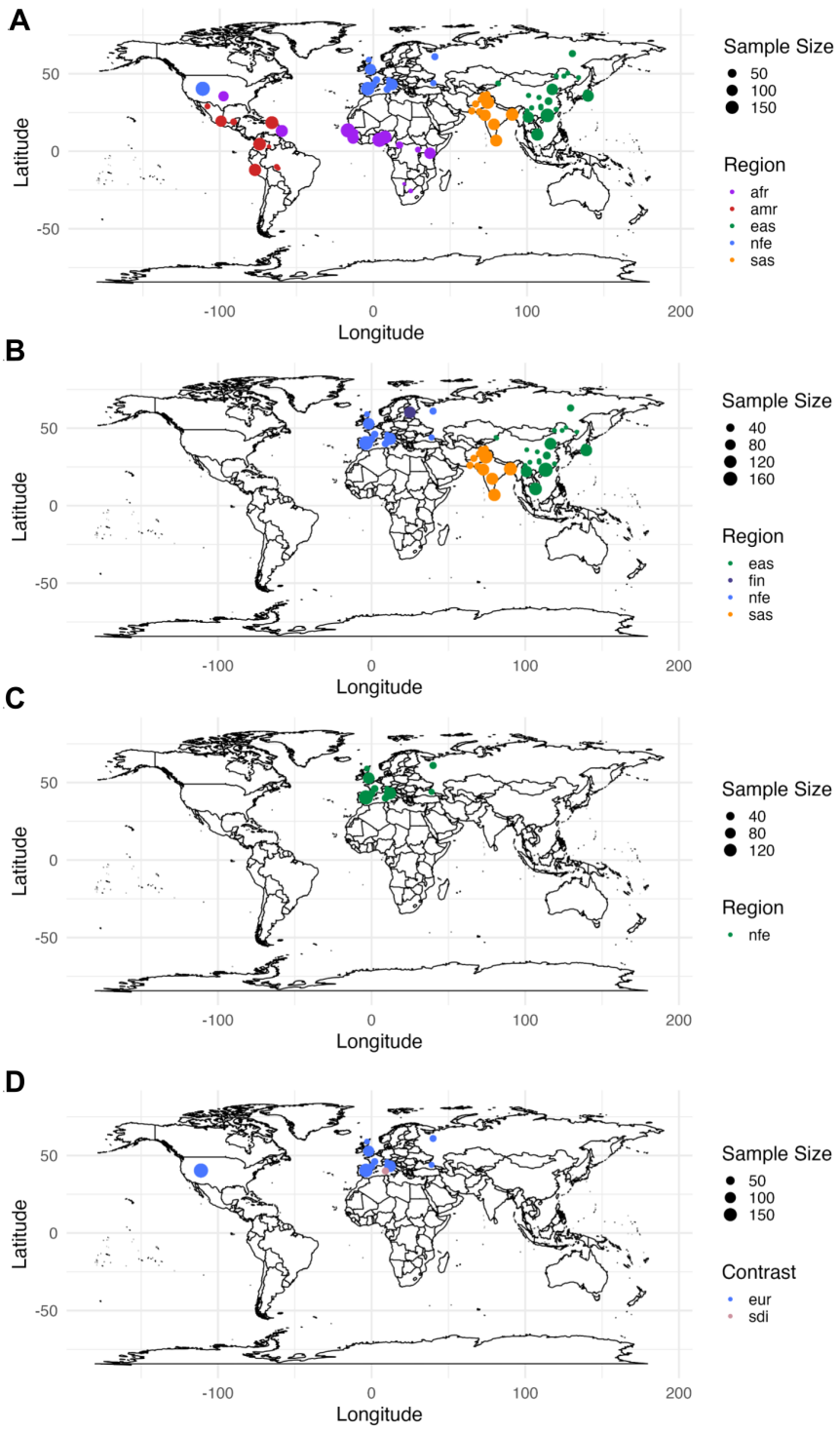

Figure 3.21: **HGDP and 1kGP test panels**

(A) Panel used to compute pairwise continental allele frequency differences. (B) Eurasian panel. (C) Non-Finnish European panel. (C) Sardinian panel that includes all Sardinian and non-Finnish Europeans.

Within the British Isles Country of Birth Test Panel		
Country	Code	Sample Size
England	ENG	142278
Northern Ireland	NI	1121
Republic of Ireland	RoI	1803
Scotland	SCT	14595
Wales	WAL	8128

Table 3.2: **Sample sizes for different countries of birth within the British Isles.**

Phenotype	Code
Standing Height	50
Alkaline Phosphate	30610
Aspartate Aminotransferase	30650
Basophill Percentage	30220
Body Weight	21002
Cholesterol	30690
Eosinophill Percentage	30210
Glucose	30740
HbA1c	30750
HDL	30760
LDL	30780
Mean Corpuscular Haemoglobin Concentration (MCHC)	30060
Mean Corpuscular Volume (MCV)	30040
Platelet Count	30080
Systolic Blood Pressure (SBP)	4080
Total Protein	30860
Triglycerides	30870

Table 3.3: **UKBB phenotypes used for polygenic score association tests.**

Contrasts	$\hat{q}$	**p-value**	**Bonf. p-value**	**GWAS**	**Phenotype**	$R^2_{40}/\hat{\gamma}_{F_{Gr}}$
eurasia-long	3.327	8.789e-04	1.494e-02	5	LDL	NA
eur-long	-3.315	9.150e-04	1.555e-02	5	MCV	0.600
sdi-eur	-3.869	1.093e-04	1.859e-03	5	MCV	0.557
sdi-eur	-3.578	3.463e-04	5.887e-03	5	Height	0.557
eas-nfe	3.52	4.315e-04	7.336e-03	5	LDL	NA
eas-afr	3.117	1.827e-03	3.107e-02	5	Cholesterol	NA
eas-afr	2.995	2.746e-03	4.668e-02	5	LDL	NA
eas-amr	3.055	2.251e-03	3.827e-02	5	LDL	NA
sas-amr	-3.857	1.148e-04	1.952e-03	5	MCV	NA
eurasia-long	3.286	1.017e-03	1.728e-02	10	Cholesterol	0.355
eurasia-long	3.659	2.533e-04	4.306e-03	10	LDL	0.355
sdi-eur	-3.258	1.122e-03	1.908e-02	10	BodyWeight	0.706
sdi-eur	-4.036	5.439e-05	9.246e-04	10	Height	0.706
eas-nfe	3.111	1.862e-03	3.165e-02	10	LDL	0.399
sdi-eur	-3.281	1.033e-03	1.756e-02	50	MCV	0.817
sdi-eur	-4.392	1.121e-05	1.905e-04	200	Height	0.871
nfe-amr	-3.184	1.453e-03	2.470e-02	200	Glucose	0.992

Table 3.4: **Significant polygenic score association tests in the HGDP1kGP panel.**

Contrasts	$\hat{q}$	**p-value**	**Bonf. p-value**	**GWAS**	**Phenotype**	$R^2_{40}/\hat{\gamma}_{F_{Gr}}$
ENG-SCT	-3.04	2.340e-03	0.0398	5	MCV	0.855
NI-WAL	3.00	2.807e-03	0.0477	5	MCV	0.797
RoI-WAL	3.01	2.641e-03	0.0449	5	MCHC	0.836
SCT-WAL	3.40	6.713e-04	0.0114	5	MCV	0.765
SCT-WAL	3.15	1.613e-03	0.0274	5	MCHC	0.765
SCT-WAL	-3.40	6.826e-04	0.0116	50	Triglycerides	0.767

Table 3.5: **Significant polygenic score association tests using British Isles country of birth contrasts.**

Contrasts	$\hat{q}$	**p-value**	**Bonf. p-value**	**GWAS**	**Phenotype**	$R^2_{40}/\hat{\gamma}_{F_{Gr}}$
Historical	3.05	2.229e-03	0.0388	5	MCV	NA
European Neolithic	-3.29	1.009e-03	0.0172	10	SBP	NA
Historical	-3.15	1.256e-03	0.02732	50	SBP	NA

Table 3.6: **Significant polygenic score association tests in the ancient DNA contrasts.**

Contrasts	$\hat{q}$	**p-value**	**Bonf. p-value**	**GWAS**	**Phenotype**	$R^2_{40}/\hat{\gamma}_{F_{Gr}}$
UK10K	3.862	1.126e-04	1.914e-03	5	StandingHeight	0.574
UK10K	3.506	4.547e-04	7.731e-03	10	StandingHeight	0.773
UK10K	-3.095	1.968e-03	3.346e-02	50	TotalProtein	0.905
UK10K	3.99	6.608e-05	1.123e-03	50	BasophillPercentage	0.905
UK10K	3.288	1.009e-03	1.715e-02	200	BodyWeight	0.968
UK10K	4.242	2.212e-05	3.761e-04	200	StandingHeight	0.968

Table 3.7: **Significant polygenic score association tests using SDS constrasts.**

CHAPTER 4

CONCLUSION

The relationship between ancestry, genetics, and the environment is complicated. The hope of polygenic scores is that they are able to separate out the direct genetic effects on phenotypes of interest, allowing researchers to study the effect of genetics in the context of different environments and across individuals of different ancestral backgrounds. Throughout the past 15 years the field has repeatedly observed intriguing patterns in polygenic scores, including an inferred decrease in polygenic scores for BMI over the past 10,000 years in west Eurasia[35], a positive relationship between polygenic scores for educational attainment and enrollment in more advanced math classes among students of European ancestry in U.S high schools[121], and differences in the distribution in polygenic scores for type 2 diabetes subtypes across ancestry groups[122]. The conclusions drawn from these studies, and others like them, have significant consequences for the fields of evolutionary biology, social science, and public health. Therefore, it is vitally important that we, as a field, have confidence in the methodology behind these analyses.

Interpreting patterns in the distribution of polygenic scores is difficult on multiple levels, especially if confounding cannot be ruled out. In this thesis, I analyze the procedure of testing for differences in polygenic scores in the presence of confounding. In Chapter 2, we first characterize patterns of stratification bias in the distribution of polygenic scores as a function of the expected genetic similarity between GWAS and test panels. For any given polygenic score association test axis, the amount of bias in the association test statistic depends on the strength of stratification along exactly one axis of population structure in the GWAS panel ($\tilde{F}_{Gr}$). We show that including this singular vector as a covariate in the GWAS model results in an unbiased test. Additionally, we analyze the standard approach of including top principal components of the GWAS panel in the context of our model, and show that including J population PCs as covariates in the GWAS model succeeds when they

capture either $\tilde{F}_{Gr}$ or the expected confounders.

The results in Chapter 2 illustrate the importance of developing comprehensive and precise models. Despite extensive work on correcting for population stratification in GWAS study designs, relatively little work has been done on modeling the GWAS and polygenic score procedures jointly. Our work clarifies what constitutes overlapping structure between panels and why evolutionary diverged panels decrease the potential for confounding in the distribution of polygenic scores. However, there are several elements of our model that differ from reality, and it is worth highlighting what these are, and what their effects are. For example, our model ignores linkage among sites and assumes that we use marginal effects, rather than jointly estimated effects, to construct our polygenic scores. Firstly, linkage among sites does not change the fundamental point that controlling for $\tilde{F}_{Gr}$ is sufficient to render the effect size estimates uncorrelated with the test panel genotype contrasts under the null. However, in practice, we would still prefer to estimate effects jointly, for at least two reasons. The first is simply because doing so increases the accuracy of the polygenic scores, which will increase our power. The second is because, in the presence of residual stratification, polygenic scores constructed with jointly estimated effects should be less biased than those constructed using marginal effects. This is because, when effect sizes are estimated marginally, each site experiences the entirety of the stratification effect, and therefore gets a "full dose" of it. The stratification effect is then being added into the polygenic score multiple times across SNPs. This is why we find the bias in the polygenic score association test statistic to be proportional to the number of loci included in the polygenic score. In contrast, if effects were estimated jointly, the stratification effect will be spread out more evenly across sites, and so we would expect the effect on the polygenic score to be less extreme, but not eliminated.

Additionally, 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 it could be an issue for more complex demographic histories experienced

by real human populations, as mutations of different ages experience the demographic history differently. In practice, one solution to this problem would be to bin variants by allele frequency or estimated age[123] and compute different corrections for different frequency/age bins, similar to the suggestion by Zaidi and Mathieson (2020)[47] to use principal components computed on rare variants to correct for stratification along axes of recent genetic divergence.

We also wish to emphasize that our results are relevant for a broader set of analyses than those explicitly covered by our model. For example, with a slight shift in perspective, our model is applicable to studies that use GWAS summary statistics together with coalescent methods to test for signals of directional polygenic selection[34,63,64,124]. The key to this is to recognize such methods use patterns of haplotype variation to estimate genotype contrasts between the sampled present-day individuals and a set of unobserved ancestors, and then ask whether these estimated genotype contrasts correlate with effect size estimates for a trait of interest. Thus, within such an analysis, there also exists an $\tilde{F}_{Gr}$ that describes the extent to which individuals in the GWAS panel are more closely related to the present-day sample or the hypothetical ancestors. The SDS values we use in Chapter 3 are an example of this type of analysis. For both the coalescent approaches, as well as methods relying on direct comparison of polygenic scores, the evolutionary hypothesis being tested and the degree of susceptibility to bias follow directly from the set of genotype contrasts used in the test. Some prior work has suggested that certain coalescent methods of testing for polygenic selection are more robust to stratification bias than others[64,124], but our results show that, in general, two different methods that test the same evolutionary hypothesis using the same set of estimated effect sizes should have the same susceptibility to stratification bias. If there *are* differences in robustness to stratification bias among methods, then it is likely coming from either changing the evolutionary hypothesis being tested or from overall differences in the statistical power of the methods.

Our model in Chapter 2 also raises interesting questions about polygenic score prediction

overall. Specifically, our expression for the bias in the distribution of polygenic scores (see Equation 2.9) indicates that it is proportional to the cross-panel GRM ($\mathbf{F}_{GX}$). Canonical correlation analysis (CCA), similar to PCA, finds the linear combination of two separate vectors that explains the maximum variance. In the genomics literature, CCA has been used to integrate multiple types of -omics data[125,126]. For polygenic score analyses, if the individual level GWAS panel data is available, developing a CCA based method that identifies top axes of structure between datasets is a compelling approach to dealing with stratification along multiple shared axes of variation. Specifically, one could imagine regressing out enough canonical variables from both panels, such that $\hat{H}$ fails to reject the no overlapping structure null hypothesis. However, conventional CCA is not effective when the number of SNPs is much larger than the number of samples[127] and is likely still limited in what axes of structure are detectable in a similar to way to PCA (see Section 3.2.1). While there are significant methodological challenges to overcome, I do think efforts to jointly analyze the structure between GWAS and prediction panels may prove fruitful.

In Chapter 3, I turn my attention to controlling for stratification bias in polygenic score association tests in practice. We compare our novel direct approach and the standard PCA across a range of simulations and empirical examples. We then develop a multi-step procedure that combines elements of both approaches to estimate how well protected by J sample PCs a set of allele frequency contrasts is from confounding in a given GWAS panel. After applying our approach to thirty sets of allele frequency contrasts for 17 different phenotypes we find one globally significant signal, a positive relationship between height effect sizes and UK10K SDS values, that is well protected from stratification bias.

Why did we not find more associations that we are confident in? It is important to note some of the limitations of our study design. First, because we split the UK Biobank into four panels of 100,000 individuals each, our sample size is quite a bit smaller than the most commonly used biobanks. Additionally, in order to standardize our pipeline for each GWAS,

we chose a lenient genome-wide significance threshold and took a simple minimum p-value approach to ascertainment. Combined, these choices may mean we are simply under-powered to detect more subtle signals of divergence. Additionally, we did find other significant signals where the test was not fully protected by 40 PCs. These may or may not represent true signals of divergence, but with our current combination of GWAS panels and PCs, we cannot rule out residual stratification. Notably, the divergence in height polygenic scores between Sardinian and mainland Europeans is a globally significant signal that was not fully protected in our analysis but, combined with evidence from Chen et al. (2020), we believe is likely driven by true divergence. Finally, it is also possible that there are few signals of polygenic score divergence for the contrasts and phenotypes that we tested and and that increasing power and/or more completely controlling for stratification would not change the number of associations.

It is worth considering what an improved study design, one with no limitations on time or data access, would look like. First, it would be ideal to have access to multiple diverse biobanks with sample sizes in the multiple hundreds of thousands where the relationship between ancestry and the environment differes. For example, we might expect that All of US[128], the UK Biobank[43], Biobank Japan[44], and FinnGen[45], all of which have different recruitment strategies, healthcare systems, and population structure, might have different confounders along different ancestry gradients. Though it is worth noting that this is an assumption that warrants further investigation. With multiple GWAS panels in hand, I would then compute $\hat{H}$ for each contrast and identify GWAS panels where the contrast is independent of the structure in the GWAS panel. I would then conduct a GWAS using state-of-the-art LMM software (e.g FastGWAS[27], BOLT[23], REGENIE[29]) and polygenic score construction software (e.g ldpred2[119], Vilma[120]) that accounts for LD. Because I know that the contrasts are immune to ancestry stratification in these GWAS panels, I could focus on increasing the power of my polygenic score when choosing parameters and covariates for

these models. Next, I would select GWAS panels that have large $\hat{H}$ values and compute $R^2_J/\hat{\gamma}_{F_{Gr}}$ for a range of J PCs. Depending on the results, I might also try computing $R^2_J/\hat{\gamma}_{F_{Gr}}$ for rare variant PCs and/or ancestry components. For GWAS panels where I could achieve a high value of $R^2_{40}/\hat{\gamma}_{F_{Gr}}$, I would go forward with the same LMM and polygenic score methods, including my selected population structure estimates as fixed effects in the GWAS. I would then compare my results for the same contrast across all selected GWAS panels. I expect that, if there is a true association signal, $\hat{q}$ would be largest in GWAS panels with larger $\hat{H}$ values. These panels will have more overlap in ancestry, which will increase the portability of the polygenic scores an improve the power of our test. Observing a significant association in both independent and overlapping panels where confounding is well controlled would be compelling evidence that the signal is not being driven by residual stratification bias.

It is also worth noting some of the limitations of the above approach. First, it necessitates having access to individual-level data and the same phenotypes across multiple datasets. One future direction for this line of work could be to develop a summary statistic version that utilizes ancestry-matched LD matrices and summary statistics to regress out r from effect sizes in a way that does not remove signal. Additionally, we noticed that for two sets of contrasts that we tested, the Historical and European Neolithic one from Le et al. (2022), $\hat{H}$ was significantly smaller than expected under the null hypothesis. As discussed in Section 3.4.2, these contrasts differ in multiple aspects from our other sets of contrasts. We are unsure which of those factors, if any, is causing them to behave differently. Future work could more systematically explore limitations of our approach in terms of the type of contrasts, number of SNPs, and number of individuals.

Finally, we note that even if we could guarantee that effect sizes were completely immune to stratification, the interpretation of the results of polygenic score association tests is not straightforward[129]. For example, these analyses use effect sizes estimated in one set of genetic and environmental background, and there is no guarantee that the effects will be

the same in other backgrounds. Effect size heterogeneity can cause many difficulties with the interpretation of positive associations between polygenic scores and axes of population structure (as several papers have noted[129,130,131]). Another difficulty with interpretation arises from allelic turnover[76] and differences in tagging across populations, as a given polygenic score will have less power to detect differences between populations that are genetically more distant from the GWAS panel, and this can lead to a biased picture of how selection has actually affected the trait across populations[77]. However, none of these phenomena are expected to generate false signals of directional selection where none exist. This is because the fact that the effect size might vary across populations has no impact on the correlation between the effect size measured in only one of the populations and patterns of allele frequency differentiation among populations. One subtle caveat to this claim is that certain forms of directional interaction effects (e.g., directional dominance) could in principle create correlations between the direction of recent allele frequency change on the lineage leading to the GWAS panel individuals and the average effect as estimated under an additivity assumption, and this *would* violate the null model. However, there is little evidence for substantial interaction variance among common variants in human complex traits, so this is unlikely to be an issue in practice.

Moving beyond the specific issue of associations between polygenic scores and population structure axes, we note that GWAS can also be impacted by other forms of genetic confounding beyond the simple associations between ancestry and genetic background that we consider here, including dynastic effects, assortative mating, and stabilizing selection[39]. This issue is especially pertinent in the context of commonly studied behavioral/social outcome traits (e.gg, educational attainment, household income, intelligence, etc.) where there is evidence of strong assortative mating[132] and phenotype-biased migration[83,49], and some evidence of indirect effects[37].

It is also is worth remembering that signals of polygenic score divergence, even if they are

free from all types of confounding, do not necessarily map in a straightforward way to *phenotypic* divergence. Complex traits are influenced by both genetics and the environment, and we need to take care in our interpretation of polygenic score differences, especially as these types of analyses are ripe for misinterpretation and and misappropriation[133]. Therefore, as a scientist, especially one who studies patterns of genetic differentiation among human populations, it is important to me to clearly communicate what my results mean. Polygenic scores offer exciting opportunities for advancements in human genetics and it is important to continue improving their implementation while also being careful in our interpretation.

REFERENCES

1. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J., Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

2. Purcell, S.M. *et al.*, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009), number: 7256 Publisher: Nature Publishing Group.

3. Khera, A.V. *et al.*, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**, 1219–1224 (2018).

4. Sella, G. & Barton, N.H., Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics* **20**, 461–493 (2019), _eprint: https://doi.org/10.1146/annurev-genom-083115-022316.

5. Cox, S.L., Ruff, C.B., Maier, R.M. & Mathieson, I., Genetic contributions to variation in human stature in prehistoric Europe. *Proceedings of the National Academy of Sciences* **116**, 21484–21492 (2019).

6. Le, M.K. *et al.*, 1,000 ancient genomes uncover 10,000 years of natural selection in Europe (2022), pages: 2022.08.24.505188 Section: New Results.

7. Irving-Pease, E.K. *et al.*, The selection landscape and genetic legacy of ancient eurasians. *bioRxiv* 2022.09.22.509027 (2022).

8. Cox, S.L. *et al.*, Socio-cultural practices affect sexual dimorphism in stature in Early Neolithic Europe (2023), pages: 2023.02.21.529406 Section: New Results.

9. Harden, K.P. & Koellinger, P.D., Using genetics for social science. *Nature Human Behaviour* **4**, 567–576 (2020), bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Economics;Genetics;Politics and international relations;Psychology;Sociology Subject_term_id: economics;genetics;politics-and-international-relations;psychology;sociology.

10. Price, A.L. *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006), number: 8 Publisher: Nature Publishing Group.

11. Novembre, J. *et al.*, Genes mirror geography within europe. *Nature* **456**, 98–101 (2008).

12. Menozzi, P., Piazza, A. & Cavalli-Sforza, L., Synthetic maps of human gene frequencies in europeans. *Science* **201**, 786–792 (1978).

13. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A., *The History and Geography of Human Genes*. Princeton University Press (1994).

14. Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

15. Tian, C. *et al.*, Analysis and application of european genetic substructure using 300 K SNP information. *PLoS Genet.* **4**, e4 (2008).

16. Marigorta, U.M., Rodríguez, J.A., Gibson, G. & Navarro, A., Replicability and prediction: Lessons and challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).

17. Palmer, C. & Pe'er, I., Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.* **13**, e1006916 (2017).

18. Yao, Y. & Ochoa, A., Limitations of principal components in quantitative genetic association models for human studies. *bioRxiv* 2022.03.25.485885 (2023).

19. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N., New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).

20. Zhu, C. & Yu, J., Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* **182**, 875–888 (2009).

21. Thornton, T. & McPeek, M.S., ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* **86**, 172–184 (2010).

22. Kang, H.M. *et al.*, Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354 (2010), bandiera_abtest: a Cg_type: Nature Research Journals Number: 4 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome-wide association studies Subject_term_id: genome-wide-association-studies.

23. Loh, P.R. *et al.*, Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015), number: 3 Publisher: Nature Publishing Group.

24. Hoffman, G.E., Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLOS ONE* **8**, e75707 (2013), publisher: Public Library of Science.

25. Zhang, Y. & Pan, W., Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genet. Epidemiol.* **39**, 149–155 (2015).

26. Schraiber, J.G., Edge, M.D. & Pennell, M., Unifying approaches from statistical genetics and phylogenetics for mapping phenotypes in structured populations. *bioRxiv* (2024).

27. Jiang, L. *et al.*, A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).

28. Zhou, W. *et al.*, Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

29. Mbatchou, J. *et al.*, Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

30. Vilhjálmsson, B.J. & Nordborg, M., The nature of confounding in genome-wide association studies. *Nature Reviews Genetics* **14**, 1–2 (2013).

31. Zhang, Y. & Pan, W., Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements? *Genetic Epidemiology* **39**, 149–155 (2015), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.21879.

32. Bulik-Sullivan, B.K. *et al.*, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295 (2015), number: 3 Publisher: Nature Publishing Group.

33. Bulik-Sullivan, B. *et al.*, An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236–1241 (2015), number: 11 Publisher: Nature Publishing Group.

34. Field, Y. *et al.*, Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016), publisher: American Association for the Advancement of Science.

35. Akbari, A. *et al.*, Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation. *bioRxiv* (2024).

36. Berg, J.J. *et al.*, Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019), publisher: eLife Sciences Publications, Ltd.

37. Young, A.I., Benonisdottir, S., Przeworski, M. & Kong, A., Deconstructing the sources of genotype-phenotype associations in humans (2019).

38. Brumpton, B. *et al.*, Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses (2020).

39. Veller, C. & Coop, G., Interpreting population and family-based genome-wide association studies in the presence of confounding (2023), pages: 2023.02.26.530052 Section: New Results.

40. Howe, L.J. *et al.*, Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat. Genet.* **54**, 581–592 (2022).

41. Lee, J.J. *et al.*, Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

42. Young, A.I. *et al.*, Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* **54**, 897–905 (2022).

43. Bycroft, C. *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018), bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 7726 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome;Genome-wide association studies;Genotype;Haplotypes;Population genetics Subject_term_id: genome;genome-wide-association-studies;genotype;haplotypes;population-genetics.

44. Nagai, A. *et al.*, Overview of the BioBank japan project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

45. Kurki, M.I. *et al.*, Author correction: FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **615**, E19 (2023).

46. Lawson, D.J. *et al.*, Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics* **139**, 23–41 (2020).

47. Zaidi, A.A. & Mathieson, I., Demographic history impacts stratification in polygenic scores. preprint, Genetics (2020).

48. Kerminen, S. *et al.*, Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *The American Journal of Human Genetics* **104**, 1169–1181 (2019).

49. Abdellaoui, A. *et al.*, Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour* **3**, 1332–1342 (2019), number: 12 Publisher: Nature Publishing Group.

50. Haworth, S. *et al.*, Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature Communications* **10**, 333 (2019), number: 1 Publisher: Nature Publishing Group.

51. Trochet, H., Pelletier, J., Tadros, R. & Hussin, J., Comparison of polygenic risk scores for heart disease highlights obstacles to overcome for clinical use. Technical report, bioRxiv (2021), section: New Results Type: article.

52. Racimo, F., Berg, J.J. & Pickrell, J.K., Detecting polygenic adaptation in admixture graphs. *Genetics* **208**, 1565–1584 (2018).

53. Kerminen, S. *et al.*, Geographic variation and bias in the polygenic scores of complex diseases and traits in finland. *Am. J. Hum. Genet.* **104**, 1169–1181 (2019).

54. Latta, R., Differentiation of Allelic Frequencies at Quantitative Trait Loci Affecting Locally Adaptive Traits. *The American Naturalist* **151**, 283–292 (1998), publisher: The University of Chicago Press.

55. Latta, R.G., Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist* **161**, 51–58 (2004), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1469-8137.2003.00920.x.

56. Pritchard, J.K. & Di Rienzo, A., Adaptation – not by sweeps alone. *Nature Reviews Genetics* **11**, 665–667 (2010), bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Population genetics Subject_term_id: population-genetics.

57. Pritchard, J.K., Pickrell, J.K. & Coop, G., The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Current Biology* **20**, R208–R215 (2010).

58. Turchin, M.C. *et al.*, Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* **44**, 1015–1019 (2012), number: 9 Publisher: Nature Publishing Group.

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

60. Racimo, F., Berg, J.J. & Pickrell, J.K., Detecting Polygenic Adaptation in Admixture Graphs. *Genetics* **208**, 1565–1584 (2018).

61. Josephs, E.B., Berg, J.J., Ross-Ibarra, J. & Coop, G., Detecting Adaptive Differentiation in Structured Populations with Genomic Data and Common Gardens. *Genetics* **211**, 989–1004 (2019), publisher: Genetics Section: Investigations.

62. Uricchio, L.H., Kitano, H.C., Gusev, A. & Zaitlen, N.A., An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evolution Letters* **3**, 69–79 (2019), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/evl3.97.

63. Edge, M.D. & Coop, G., Reconstructing the History of Polygenic Scores Using Coalescent Trees. *Genetics* **211**, 235–262 (2019), publisher: Genetics Section: Investigations.

64. Stern, A.J., Speidel, L., Zaitlen, N.A. & Nielsen, R., Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics* **108**, 219–239 (2021).

65. Lango Allen, H. *et al.*, Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010), number: 7317 Publisher: Nature Publishing Group.

66. Wood, A.R. *et al.*, Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186 (2014), number: 11 Publisher: Nature Publishing Group.

67. Robinson, M.R. *et al.*, Population genetic differentiation of height and body mass index across Europe. *Nature Genetics* **47**, 1357–1362 (2015), number: 11 Publisher: Nature Publishing Group.

68. Zoledziewska, M. *et al.*, Height-reducing variants and selection for short stature in Sardinia. *Nature Genetics* **47**, 1352–1356 (2015), number: 11 Publisher: Nature Publishing Group.

69. Berg, J.J., Zhang, X. & Coop, G., Polygenic Adaptation has Impacted Multiple Anthropometric Traits. preprint, Evolutionary Biology (2017).

70. Guo, J. *et al.*, Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat. Commun.* **9**, 1865 (2018).

71. Mathieson, I. *et al.*, Genome-wide patterns of selection in 230 ancient eurasians. *Nature* **528**, 499–503 (2015).

72. Sohail, M. *et al.*, Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019), publisher: eLife Sciences Publications, Ltd.

73. Chen, M. *et al.*, Evidence of Polygenic Adaptation in Sardinia at Height-Associated Loci Ascertained from the Biobank Japan. *The American Journal of Human Genetics* **107**, 60–71 (2020).

74. Martin, A.R. *et al.*, Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics* **100**, 635–649 (2017).

75. Wang, Y. *et al.*, Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications* **11**, 3865 (2020), bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation;Genome-wide association studies;Statistical methods Subject_term_id: genetic-variation;genome-wide-association-studies;statistical-methods.

76. Carlson, M.O., Rice, D.P., Berg, J.J. & Steinrücken, M., Polygenic score accuracy in ancient samples: Quantifying the effects of allelic turnover. *PLOS Genetics* **18**, e1010170 (2022), publisher: Public Library of Science.

77. Yair, S. & Coop, G., Population differentiation of polygenic score predictions under stabilizing selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **377**, 20200416 (2022), publisher: Royal Society.

78. Ding, Y. *et al.*, Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* 1–8 (2023), publisher: Nature Publishing Group.

79. Martin, A.R. *et al.*, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584–591 (2019), number: 4 Publisher: Nature Publishing Group.

80. *Recommendations use reporting race, ethnicity, ancestry genetic reseach Experiences from NHLBI TOPMed program.*

81. *Using Population Descriptors Genetics Genomics Research: New Framework Evolving Field.*

82. McVean, G., A Genealogical Interpretation of Principal Components Analysis. *PLOS Genetics* **5**, e1000686 (2009), publisher: Public Library of Science.

83. Abdellaoui, A., Dolan, C.V., Verweij, K.J.H. & Nivard, M.G., Gene–environment correlations across geographic regions affect genome-wide association studies. *Nature Genetics* 1–10 (2022), publisher: Nature Publishing Group.

84. Mostafavi, H. *et al.*, Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**, e48376 (2020), publisher: eLife Sciences Publications, Ltd.

85. Le Corre, V. & Kremer, A., The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology* **21**, 1548–1566 (2012), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-294X.2012.05479.x.

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

87. Refoyo-Martínez, A. *et al.*, How robust are cross-population signatures of polygenic adaptation in humans? *Peer Community J.* **1**, 1–None (2021).

88. Cavalli-Sforza, L.L., Barrai, I. & Edwards, A.W.F., Analysis of Human Evolution Under Random Genetic Drift. *Cold Spring Harbor Symposia on Quantitative Biology* **29**, 9–20 (1964), publisher: Cold Spring Harbor Laboratory Press.

89. Nicholson, G. *et al.*, Assessing Population Differentiation and Isolation from Single-Nucleotide Polymorphism Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**, 695–715 (2002).

90. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J.K., Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics* **185**, 1411–1423 (2010).

91. Pickrell, J.K. & Pritchard, J.K., Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics* **8**, e1002967 (2012), publisher: Public Library of Science.

92. Bulmer, M.G., The Effect of Selection on Genetic Variability. *The American Naturalist* **105**, 201–211 (1971), publisher: The University of Chicago Press.

93. Kremer, A. & Le Corre, V., Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity* **108**, 375–385 (2012), number: 4 Publisher: Nature Publishing Group.

94. Consortium, T..G.P. & The 1000 Genomes Project Consortium, A global reference for human genetic variation (2015).

95. Patel, R.A. *et al.*, Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am. J. Hum. Genet.* **109**, 1286–1297 (2022).

96. Hou, K. *et al.*, Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).

97. Reisberg, S. *et al.*, Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLoS One* **12**, e0179238 (2017).

98. Sanderson, E., Richardson, T.G., Hemani, G. & Davey Smith, G., The use of negative control outcomes in mendelian randomization to detect potential population stratification. *Int. J. Epidemiol.* **50**, 1350–1361 (2021).

99. Patterson, N., Price, A.L. & Reich, D., Population Structure and Eigenanalysis. *PLOS Genetics* **2**, e190 (2006), publisher: Public Library of Science.

100. Baik, J., Ben Arous, G. & Péché, S., Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697 (2005).

101. Johnstone, I.M. & Paul, D., PCA in high dimensions: An orientation. *Proc. IEEE Inst. Electr. Electron. Eng.* **106**, 1277–1292 (2018).

102. Bloemendal, A. & Chen, C., PCA and stratification in GWAS / a primer on random matrix theory. `https://www.youtube.com/watch?v=B7ub92OLw1g` (2019).

103. Listgarten, J. *et al.*, Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525–526 (2012), number: 6 Publisher: Nature Publishing Group.

104. Reich, D. *et al.*, Reconstructing indian population history. *Nature* **461**, 489–494 (2009).

105. Patterson, N. *et al.*, Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).

106. Mathieson, I. & McVean, G., Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics* **44**, 243–246 (2012), number: 3 Publisher: Nature Publishing Group.

107. Novembre, J. & Stephens, M., Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**, 646–649 (2008).

108. Koenig, Z. *et al.*, A harmonized public resource of deeply sequenced diverse human genomes. *Genome Res.* **34**, 796–809 (2024).

109. The UK10K Consortium *et al.*, The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

110. Luo, H. *et al.*, Recent positive selection signatures reveal phenotypic evolution in the han chinese population. *Sci Bull (Beijing)* **68**, 2391–2404 (2023).

111. Haag, J., Jordan, A.I. & Stamatakis, A., Pandora: A tool to estimate dimensionality reduction stability of genotype data. *bioRxiv* 2024.03.14.584962 (2024).

112. Berisa, T. & Pickrell, J.K., Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

113. Refoyo-Martínez, A. *et al.*, How robust are cross-population signatures of polygenic adaptation in humans? preprint (2020).

114. Kelleher, J., Etheridge, A.M. & McVean, G., Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* **12**, e1004842 (2016), publisher: Public Library of Science.

115. Bhatia, G. *et al.*, Correcting subtle stratification in summary association statistics. *bioRxiv* (2016).

116. Chang, C.C. *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

117. Dahl, A. *et al.*, Adjusting for Principal Components of Molecular Phenotypes Induces Replicating False Positives. *Genetics* **211**, 1179–1189 (2019).

118. Li, M.X., Yeung, J.M.Y., Cherny, S.S. & Sham, P.C., Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics* **131**, 747–756 (2012).

119. Privé, F., Arbel, J. & Vilhjálmsson, B.J., LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).

120. Spence, J.P., Sinnott-Armstrong, N., Assimes, T.L. & Pritchard, J.K., A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics (2022), pages: 2022.04.18.488696 Section: New Results.

121. Harden, K.P. *et al.*, Genetic associations with mathematics tracking and persistence in secondary school. *npj Science of Learning* **5**, 1–8 (2020), number: 1 Publisher: Nature Publishing Group.

122. Smith, K. *et al.*, Multi-ancestry polygenic mechanisms of type 2 diabetes elucidate disease processes and clinical heterogeneity. *Res. Sq.* (2023).

123. Albers, P.K. & McVean, G., Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biology* **18**, e3000586 (2020), publisher: Public Library of Science.

124. Song, W. *et al.*, A selection pressure landscape for 870 human polygenic traits. *Nature Human Behaviour* **5**, 1731–1743 (2021), number: 12 Publisher: Nature Publishing Group.

125. Jendoubi, T. & Strimmer, K., A whitening approach to probabilistic canonical correlation analysis for omics data integration (2019).

126. Lin, D. *et al.*, Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics* **14**, 245 (2013).

127. Parkhomenko, E., Tritchler, D. & Beyene, J., Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.* **8**, Article 1 (2009).

128. All of Us Research Program Genomics Investigators, Genomic data in the all of us research program. *Nature* **627**, 340–346 (2024).

129. Novembre, J. & Barton, N.H., Tread Lightly Interpreting Polygenic Tests of Selection. *Genetics* **208**, 1351–1355 (2018).

130. Rosenberg, N.A., Edge, M.D., Pritchard, J.K. & Feldman, M.W., Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evolution, Medicine, and Public Health* **2019**, 26–34 (2019).

131. Harpak, A. & Przeworski, M., The evolution of group differences in changing environments. *PLOS Biology* **19**, e3001072 (2021), publisher: Public Library of Science.

132. Robinson, M.R. *et al.*, Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 0016 (2017).

133. Carlson, J. & Harris, K., Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation. *PLoS Biol.* **18**, e3000860 (2020).