Observational epidemiological studies can mitigate genetic confounding with the genetic relatedness matrix

Roshni A. Patel,*[1,2,3] Joshua G. Schraiber,[1] Matt Pennell,*[1,4] Michael D. Edge*[1]

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089,

[2]Department of Data Science, University of Oregon, Eugene, OR 97403

[3]Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403

[4]Department of Computational Biology, Cornell University, Ithaca, NY 14853

*Correspondence: roshnip@uoregon.edu, edgem@usc.edu, mpennell@cornell.edu

Abstract

Observational studies are commonly used in psychology and epidemiology to identify risk factors correlated with health outcomes. However, these studies are vulnerable to confounding when shared genetic variation influences both the putative risk factor and outcome. Researchers have historically controlled for this type of genetic confounding using polygenic scores, but these scores are often noisy and biased estimators of a trait's genetic component. Here, we develop a method that leverages solutions to a similar problem in the field of phylogenetics. Motivated by inference of causal effects in phylogenetics, we show that the genetic relationship matrix (GRM) can be used to control genetic confounding when testing for non-genetic risk factors. In simulations, we find that our method outperforms existing approaches, particularly in the sample sizes characteristic of datasets in psychology and epidemiology. We also demonstrate that while existing methods are susceptible to poor GWAS portability, our method is inherently robust to such concerns. Finally, we apply our method to the UK Biobank to re-analyze social risk factors for health outcomes in previously understudied cohorts.

Introduction

Empirical research across the social, behavioral, and health sciences often relies on observational studies to test for relationships between putative exposures and outcomes. However, these observational studies can be subject to confounding when the exposure and outcome are both complex traits. Complex traits are influenced by tens to hundreds of thousands of causal variants, and these variants also tend to be pleiotropic, influencing multiple traits simultaneously (Bulik-Sullivan *et al.*, 2015; Sinnott-Armstrong *et al.*, 2021; Yengo *et al.*, 2022). This raises the possibility that both exposure and outcome are influenced by shared genetic variation.

Shared genetic variation poses a problem for causal inference in observational epidemiological studies. In many instances, researchers are interested in understanding whether intervening on a particular exposure (e.g. BMI) could modify someone's risk for a particular outcome (e.g. high cholesterol). In other words, to understand the effect of environmental interventions, it is essential to understand whether an observed correlation between exposure and outcome manifests independently of genetics: correlations that arise solely due to shared genetic variation will not translate to effective interventions. For example, previous work has found that associations between developmental risk factors and later-in-life psychopathology are attenuated after accounting for shared genetics and other factors (D'Onofrio *et al.*, 2014). Across applications in psychology and public health, there is a growing awareness of the importance of controlling for shared genetic variation that gives rise to the putative exposure and outcome of interest (D'Onofrio *et al.*, 2020; Jaffee and Price, 2012). Within the social sciences, recognition of this problem has motivated a surge of interest in behavioral genetics, a field that seeks to incorporate genetic information in order to better understand the effects of behavioral interventions (Harden and Koellinger, 2020).

Historically, efforts to control genetic confounding have relied on including covariates that capture the genetic variation underlying the exposure. One approach conditions on a small number of genetic variants that have a large effect on the exposure trait (Choi *et al.*, 2019; Manousaki *et al.*, 2021; Smith and Ebrahim, 2003; Smith and Hemani, 2014). However, this approach assumes that the genetic architecture of the exposure trait is dominated by common, large-effect variants, which is not always true for complex traits (Sinnott-Armstrong *et al.*, 2021). Another line of work has sought to control for genetic confounding by including the polygenic score for the exposure as a covariate when regressing outcome on exposure (Belsky *et al.*, 2016; Conley, 2016; Cuevas *et al.*, 2021; D'Onofrio *et al.*, 2020; Lee *et al.*, 2021; Schmitz and Conley, 2017; Zheng *et al.*, 2023). Polygenic scores (PGS) estimate the genetic component of a trait by combining information across hundreds to thousands of

genetic associations ascertained from genome-wide association studies (GWAS). In theory, one could control genetic confounding by perfectly estimating the genetic value for the exposure and including it as a covariate when regressing outcome on exposure. In practice, PGS are poor approximations of the true genetic value: PGS estimated from data do not capture all causal variants, and the effects of these causal variants are estimated with noise. In response, a considerable body of work has sought to better control genetic confounding by improving PGS estimation and carefully modeling measurement error (DiPrete *et al.*, 2018; Pingault *et al.*, 2021, 2022; Uddin *et al.*, 2022). However, less progress has been made in improving PGS "portability": PGS are known to be biased and less predictive when estimated from and applied to groups with different genetic ancestries or environments (Martin *et al.*, 2019; Mostafavi *et al.*, 2020; Wang *et al.*, 2024). PGS-corrected estimates of the exposure effect are likely to be unreliable when applied to populations that are underrepresented in the cohorts that PGS were estimated from.

Recently, Zhao *et al.* (2024) introduced an alternative approach to control genetic confounding named PENGUIN. PENGUIN-corrected estimates of the exposure effect are a tremendous improvement over previous methods that rely on polygenic scores. Whereas earlier PGS-based methods sought to control genetic confounding by noisily estimating individual-level genetic values, PENGUIN uses LD score regression (LDSC) (Bulik-Sullivan *et al.*, 2015) to directly estimate the genetic correlation between exposure and outcome from GWAS summary statistics. This enables PENGUIN to estimate the exposure effect by subtracting off the correlation between exposure and outcome that is mediated through genetics.

However, at least two problems remain. First, both PENGUIN and PGS-based methods presuppose that data on exposure and outcome has been collected in sample sizes that are large enough for GWAS (that is, tens or hundreds of thousands of individuals). This requirement limits the utility of existing datasets where valuable phenotypic information may have been collected in smaller sample sizes. Moreover, biobank-scale measurements may be fundamentally untenable for many variables analyzed in observational epidemiological studies: for example, psychologists commonly study traits that are measured by combining the responses to multi-question surveys. Second, even when traits have been measured in large cohorts, methods that rely on GWAS summary statistics are susceptible to poor portability across cohorts with differences in genetic ancestries or environmental backgrounds (Mostafavi *et al.*, 2020; Patel *et al.*, 2025; Wang *et al.*, 2024). Despite growing recognition of this problem, progress in expanding the diversity of biobank participants remains slow (Mills and Rahal, 2019). Thus, there remains a need to develop a method that controls genetic confounding when studying traits that are hard to measure, or populations underrepresented in existing GWAS.

To develop a flexible method to control genetic confounding in observational epidemiological studies, we leverage the similarities to a well-studied problem in phylogenetics. When studying ecological and evolutionary relationships between traits, phylogeneticists often use interspecific data to test for an association between two traits. Crucially, researchers must account for confounding that results from species' shared ancestry via their phylogenetic relationships (Felsenstein, 1985). A vast array of methods have been developed to mitigate phylogenetic confounding, but the most common approach is to use a linear mixed model with random effects structured by the phylogenetic covariance among lineages (Lynch, 1991; Martins and Hansen, 1997; Uyeda *et al.*, 2018; Westoby *et al.*, 2023).

Recent work by Schraiber *et al.* (2024) demonstrated that the phylogenetic covariance matrix is directly analogous to the genetic relatedness matrix (GRM) in statistical genetics. Drawing on this analogy, we sought to control for genetic confounding in observational epidemiological studies using the GRM. The GRM is typically constructed as the variance-covariance matrix of genotypes, and it encodes information on genetic relationships among individuals (Speed and Balding, 2015). Thus, the GRM can be used to partition the naive correlation between exposure and outcome into a component mediated through genetics and a component independent of genetics (Lee *et al.*, 2012; Yang *et al.*, 2010). Here, we demonstrate the conceptual similarity between GRM-corrected estimates and estimates derived from earlier GWAS-based methods. We show that GRM-corrected estimates largely resolve the two problems associated with GWAS-corrected estimates: they can be applied to small datasets and computed within-sample, mitigating portability concerns. We conclude by using our method to re-analyze social risk factors for health outcomes in 6,104 Indian and 8,483 Black British individuals in the UK Biobank.

Results

Statistical model

We begin by introducing a generative statistical model of the exposure, X, and the outcome, Y (Figure 1A):

$$X = G\beta_X + e_X$$

$$Y = G\beta_Y + e_Y + be_X$$

where G is the centered genotype matrix; β_X and β_Y are additive SNP effect sizes; and e_X and e_Y are non-genetic effects on the traits. We assume e_X and e_Y are independently and identically distributed across individuals with a mean of 0, and we further assume that non-genetic effects e_X and e_Y are independent of G. We assume β_X and β_Y have a mean of 0, and make no other assumptions, such that β_X and β_Y could be independent or correlated.

Under this generative model, the outcome is correlated with the exposure through both genetic and non-genetic mechanisms, and the coefficient b measures the effect of the non-genetic component of the exposure. In other words, b is the exact quantity we wish to estimate in observational epidemiological studies: it quantifies the extent to which intervening on the exposure will affect the outcome variable. In contrast, the naive estimate of the exposure effect—obtained by simply regressing the outcome on the exposure—will be confounded by the genetic correlation between the traits.

Though, to our knowledge, this model has not been explicitly described in prior literature, we show that it is the basis for existing methods to control genetic confounding (Supplementary Note 1). Given that e_X is unobserved and unmeasurable, estimating b requires indirect estimation. PENGUIN and PGS-based methods approach this problem by modeling $\mathrm{Cov}(G\beta_X, G\beta_Y)$ and $G\beta_X$, respectively. Here, we propose estimating the interventional effect of the exposure by using the GRM to partition the covariance between exposure and outcome into a genetic component and a non-genetic component.

If we assume that both the effects of genetic variants and the non-genetic effects on traits are normally distributed, the exposure and outcome can be modeled with a bivariate normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_G \otimes R + \Sigma_E \otimes I \right) \tag{1}$$

where R is the GRM, and Σ_G and Σ_E represent the genetic and non-genetic variance-covariance matrices between exposure and outcome. Given a set of genotypes from which to construct the GRM, and phenotypes for the exposure and outcome, Σ_G and Σ_E can be estimated with bivariate variance component analysis. In fact, the model in Equation 1 may be familiar as bivariate GREML, which is commonly used to estimate heritability and genetic correlation using the entries of Σ_G (Yang *et al.*, 2010; Lee *et al.*, 2012). In contrast, we propose using the entries of Σ_E to estimate the interventional effect of an exposure in the context of observational epidemiological studies, such that $\hat{b} = \frac{\sigma_{e_{XY}}}{\sigma^2_{e_X}}$, where $\Sigma_E = \begin{pmatrix} \sigma^2_{e_X} & \sigma_{e_{XY}} \\ \sigma_{e_{XY}} & \sigma^2_{e_Y} \end{pmatrix}$.

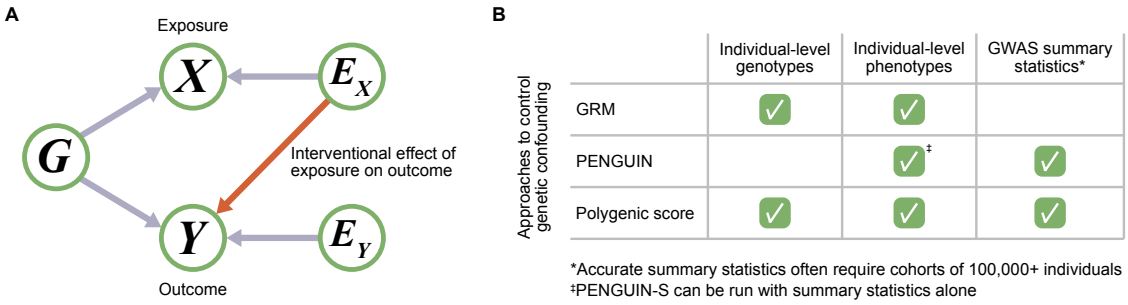

Figure 1: **Overview. A)** Causal graph for observational epidemiological studies where both exposure and outcome are genetic traits. The orange arrow illustrates the quantity of interest: the interventional effect of the exposure on the outcome. **B)** Data requirements for GRM-, PENGUIN-, and PGS-corrected estimates of exposure effect.

Assessing the utility of GRM-corrected estimates

Though PGS-, PENGUIN-, and GRM-corrected estimates share the same underlying model of trait variation, these approaches impose different data requirements (Figure 1B). PGS-corrected estimates require GWAS summary statistics to construct a PGS and individual-level genotypes and phenotypes to regress outcome on exposure. In contrast, PENGUIN-corrected estimates do not require individual-level genotypes. Instead, PENGUIN uses GWAS summary statistics to estimate the genetic correlation between traits, and individual-level phenotypes to estimate the overall phenotypic correlation between exposure and outcome. Finally, GRM-corrected estimates require no GWAS data at all, and instead partition the covariance between exposure and outcome strictly using individual-level genotypes and phenotypes.

Thus, we sought to understand the utility of different methods for controlling genetic confounding. We simulated genotypes for individuals under a realistic model of human demography (Gutenkunst *et al.*, 2009; Adrion *et al.*, 2020). Conditional on these genotypes, we simulated phenotypes for exposure and outcome traits under an additive genetic architecture with frequency-dependent effect sizes (see Methods).

In simulations, we find that the perfect PGS – that is, the true genetic value – can successfully control genetic confounding (Figure 2A). However, as the correlation between the estimated PGS and the true genetic value decreases, PGS-corrected estimates of the exposure effect become increasingly susceptible to genetic confounding. Even when a PGS is constructed with 100% of the true causal variants, noisy estimation of variant effect sizes will result in PGS-corrected estimates being biased towards the genetic correlation between exposure and outcome. Given that current GWAS are not expected to identify 100% of the causal variants for a trait, this represents a relatively optimistic

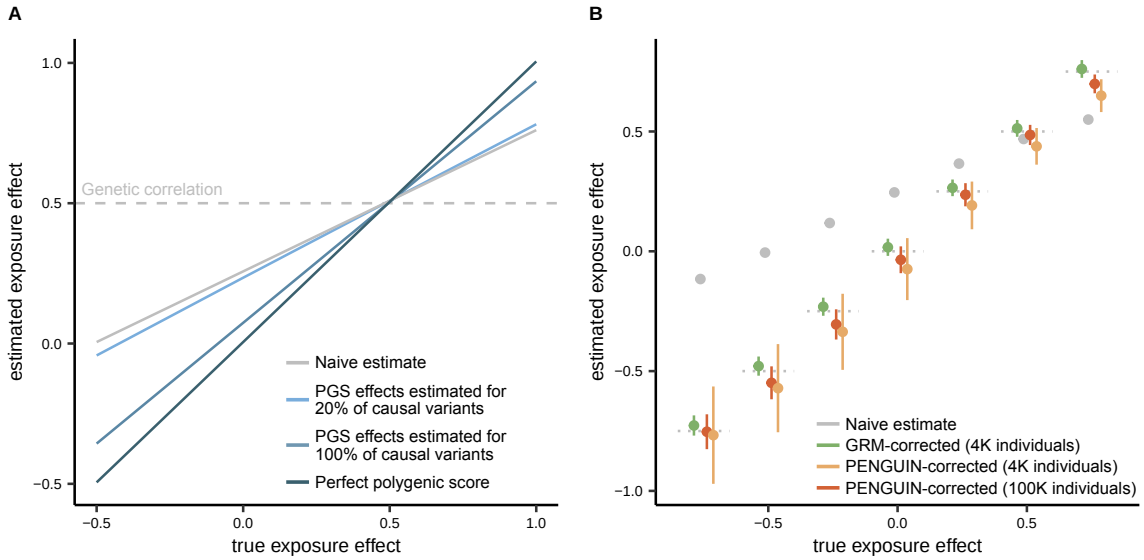

Figure 2: **Utility of GRM-corrected estimates. A)** PGS-corrected estimates of exposure effect, for i) PGS constructed from 20% of causal variants and effects estimated with noise; ii) PGS constructed from 100% of causal variants and effects estimated with noise; and iii) the perfect PGS. **B)** GRM- and PENGUIN-corrected estimates of exposure effect in simulated data, compared to naive estimates obtained by directly regressing outcome on exposure. Error bars correspond to mean $\pm$ 1 standard deviation.

scenario. When only 20% of causal variants are captured, the PGS-corrected estimate of exposure effect barely performs better than the naive marginal estimate. For the remainder of this manuscript, we therefore prioritize comparing GRM-corrected estimates with PENGUIN-corrected estimates.

We next compared the performance of PENGUIN- and GRM-corrected estimates in realistic scenarios (Figure 2B). Some observational epidemiological studies are conducted in biobanks with information on hundreds of thousands of individuals, while others rely on datasets with a few thousand individuals (Conley, 2016; D'Onofrio *et al.*, 2020). We find that GRM-corrected estimates derived from sample sizes of just 4,000 individuals perform as well as PENGUIN-corrected estimates derived from biobank-scale data. In smaller sample sizes, we find that GRM-corrected estimates substantially outperform PENGUIN-corrected estimates, reflecting the reliance of PENGUIN on GWAS summary statistics. At a sample size of 4,000 individuals, GWAS summary statistics are too noisy for accurate estimation of heritability and genetic correlation via LD score regression. This highlights the particular utility of GRM-corrected estimates in studies with sample sizes that are too small for reliable GWAS. On the other hand, the utility of PENGUIN-corrected estimates becomes apparent as sample sizes approach hundreds of thousands of individuals. In this setting, constructing a GRM becomes computationally infeasible, but GWAS summary statistics become less noisy and more accurate.

Interrogating assumptions about trait heritability

To identify the potential limitations of GRM- and PENGUIN-corrected estimates, we next examined their robustness to assumptions about trait heritability. Both approaches estimate the interventional effect of the exposure by partitioning the genetic correlation between traits from the non-genetic correlation between them, which requires making assumptions about how trait heritability is distributed across individuals and variants. Thus, we sought to understand how violations of these assumptions affect our ability to recover the true interventional estimate of exposure effect.

We first evaluated the effect of violating assumptions about random mating. Population structure and cross-trait assortative mating can both induce spurious correlations between individuals' genotypes and are known to bias traditional heritability and genetic correlation estimators (Border *et al.*, 2022, 2024). To characterize the effect of population structure, we simulated two subpopulations with an F_{ST} of 0.005, a value comparable to those observed between populations within continental Europe. We find that when the mean exposure phenotype is shared between subpopulations, the confidence intervals for GRM- and PENGUIN-corrected estimates generally recover the true exposure effect, though PENGUIN-corrected estimates have much wider confidence intervals (Figure S1A). However, if the mean exposure effect differs between subpopulations (e.g., due to environmental differences), we find that PENGUIN-corrected estimates exhibit greater bias than GRM-corrected estimates (Figure S1B). In simulations of cross-trait assortative mating, we find that confidence intervals for GRM-corrected estimates generally recover the true exposure effect, but estimates can be prone to bias as the strength of cross-trait assortative mating increases (Figure S1C). We were unable to evaluate PENGUIN-corrected estimates under assortative mating given the large sample sizes required for PENGUIN and the computational complexity of assortative mating simulations (Border *et al.*, 2024).

Lastly, we considered assumptions about how heritability is distributed across variants. We find that if the exposure and outcome are simulated with a heritability model that violates method assumptions, both GRM- and LDSC-based estimates of heritability will be biased, consistent with previous findings (Supp Fig S2A) (Speed *et al.*, 2017). Nevertheless, we find that GRM-corrected estimates of exposure effect are generally more successful at recovering the true effect (Supp Fig S2B).

Portability of methods to control genetic confounding

Finally, we sought to characterize the portability of methods to control genetic confounding. While GRM-corrected estimates can only be obtained from a single dataset with individual-level genotype

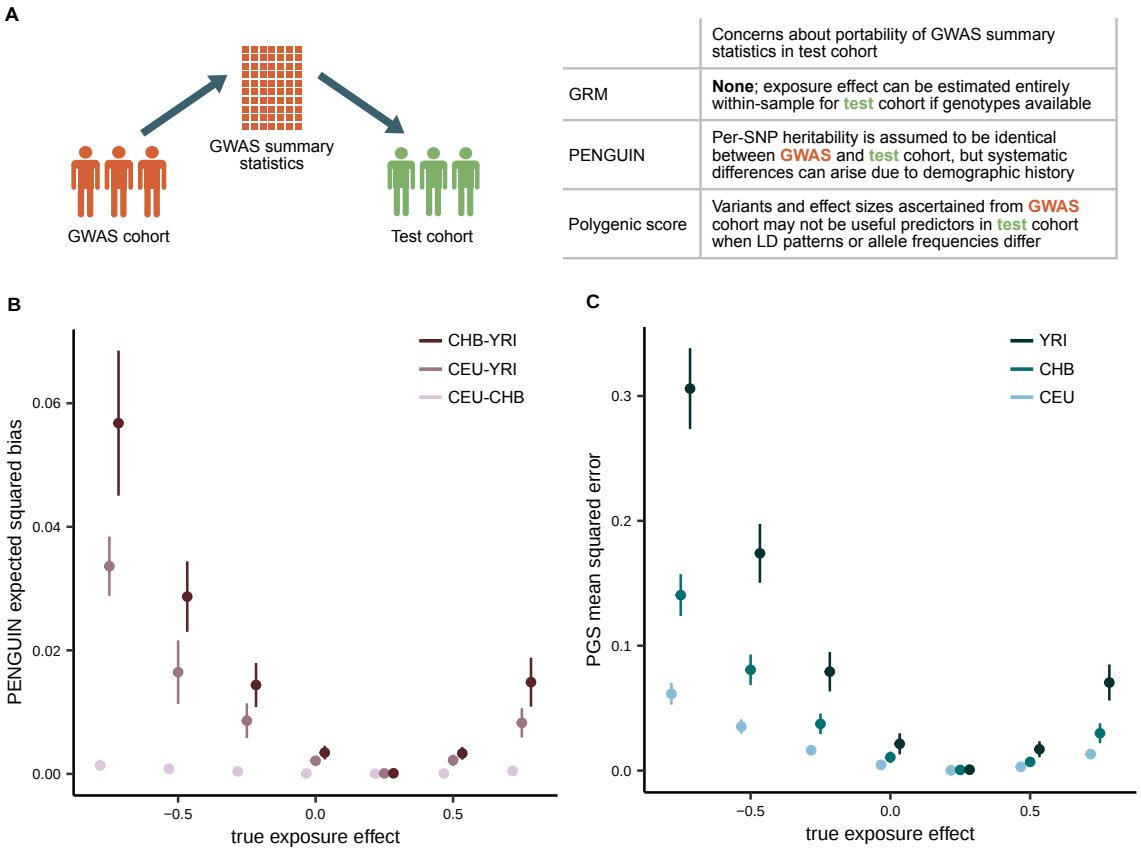

Figure 3: **Portability of methods.** **A)** Overview of portability concerns associated with GRM-, PENGUIN-, and PGS-corrected estimates of exposure effect. **B)** Expected squared bias in PENGUIN-corrected estimates of exposure effect when GWAS summary statistics and individual-level data are obtained from two different populations. **C)** Mean squared error in PGS-corrected estimates of exposure effect when GWAS summary statistics are obtained from simulated CEU individuals. Error bars correspond to mean ± 1 standard deviation.

and phenotype data, both PGS- and PENGUIN-corrected estimates can be obtained by combining information across multiple cohorts. In the case of PENGUIN-corrected estimates, this feature presents a considerable methodological advantage: by leveraging external GWAS summary statistics, researchers can use PENGUIN to estimate exposure effect even in the absence of individual-level genotype data (Figure 1B). At the same time, combining information across cohorts makes both PENGUIN- and PGS-corrected estimates susceptible to concerns about poor portability across cohorts.

To characterize the portability of PGS- and PENGUIN-corrected estimates, we first distinguish between the types of portability concerns relevant to both methods (Figure 3A). The portability of PGS has been written about extensively (Berg *et al.*, 2019; Martin *et al.*, 2019; Mostafavi *et al.*, 2020; Sohail *et al.*, 2019; Wang *et al.*, 2024), and is defined as a loss of accuracy when GWAS are conducted in one cohort and the resulting PGS is applied to a second cohort. This loss of accuracy has many

causes, including differences in allele frequency of causal variants; differences in LD between causal variants and tag variants identified in GWAS; differences in the role of the environment; and gene-by-gene or gene-by-environment interactions (Mostafavi *et al.*, 2020; Patel *et al.*, 2022, 2025; Wang *et al.*, 2020; Yair and Coop, 2022).

In contrast, the PENGUIN estimator faces a different kind of portability challenge. In the context of the PENGUIN estimator, "poor portability" means that PENGUIN-corrected estimates are inaccurate when GWAS summary statistics and individual-level phenotypes are obtained from two different cohorts. We show analytically that the PENGUIN estimator is no longer guaranteed to recover the true exposure effect when per-SNP heritability differs between the two cohorts due to factors such as demographic history or selection shaping allele frequencies, or cohort-specific gene-by-gene and gene-by-environment interactions modifying effect sizes (Supplementary Note 2). Specifically, we find that the expected value of the PENGUIN estimator is:

$$\mathbb{E}[\hat{b}_{\mathrm{PENG}}] \approx \frac{b\sigma_e^2 + \sum_{k=1}^{M} 2\beta_{X,k}\beta_{Y,k}\Big(p_{k,\mathrm{ind}}(1-p_{k,\mathrm{ind}}) - p_{k,\mathrm{GWAS}}(1-p_{k,\mathrm{GWAS}})\Big)}{\sigma_e^2 + \sum_{k=1}^{M} 2\beta_{X,k}^2\Big(p_{k,\mathrm{ind}}(1-p_{k,\mathrm{ind}}) - p_{k,\mathrm{GWAS}}(1-p_{k,\mathrm{GWAS}})\Big)} \tag{2}$$

To clarify the magnitude of error associated with these two types of portability concerns, we simulated populations under a previously published demographic model inferred from 1000 Genomes CEU (Utah residents with Northern and Western European ancestry), CHB (Han Chinese in Beijing, China), and YRI (Yoruba in Ibadan, Nigeria). Importantly, the portability concerns we modeled are purely due to differences in allele frequency and linkage disequilibrium arising from demographic history. We do not consider gene-by-environment interactions, population structure, or population-specific exposure effects, all of which could further reduce the portability of methods to control genetic confounding.

To characterize PENGUIN portability, we computed the expected squared bias of the estimator as $(b - \mathbb{E}[\hat{b}_{\mathrm{PENG}}])^2$. As before, we find that the expected squared bias increases as the true exposure effect becomes more dissimilar to the genetic correlation (Figure 3B). We also find that the error is most pronounced for comparisons between CEU and YRI, or CHB and YRI. This reflects systematically different per-SNP heritabilities in CEU and CHB relative to YRI, and can be explained by the different amounts of genetic drift in their respective demographic histories.

We next characterized PGS portability by comparing the mean squared error associated with out-of-sample GWAS and in-sample GWAS. Briefly, we performed GWAS on the exposure trait in simulated CEU individuals; clumped trait-associated variants to construct a PGS; and applied that PGS to control genetic confounding in CHB and YRI. Similar to PENGUIN, we find that PGS-

corrected estimates exhibit poor portability between groups (Figure 3C). However, we reiterate that unlike PENGUIN, PGS-corrected estimates can be biased even with in-sample GWAS (Figure 2A).

Empirical data analyses

Having demonstrated the utility of the GRM in controlling genetic confounding, as well as its robustness to assumptions about trait heritability, we turned our attention to empirical data. Specifically, we sought to understand the effect of loneliness on health outcomes in the UK Biobank. Chronic loneliness has an incidence of 5-10% in modern societies (Beutel *et al.*, 2017; Hakulinen *et al.*, 2018), and is recognized as a significant public health concern (Office of the Surgeon General, 2023). Using observational epidemiological studies, researchers have found that loneliness is associated with numerous health and social outcomes, including increased risk for cardiovascular disease (Hakulinen *et al.*, 2018; Liu *et al.*, 2024; Paul *et al.*, 2021; Rødevand *et al.*, 2021; Xia and Li, 2018), increased inflammation (Matthews *et al.*, 2024; Van Bogart *et al.*, 2022), and poor mental health (Abdellaoui *et al.*, 2019; Dennis *et al.*, 2021; Lee *et al.*, 2021; Rødevand *et al.*, 2021). At the same time, loneliness seems to be genetically correlated with some of these same health outcomes (Abdellaoui *et al.*, 2019; Moshtael *et al.*, 2024), demonstrating the need to control for potential genetic confounding when testing for loneliness as a risk factor.

We estimated the effect of loneliness on three cardiovascular traits, four mental health traits, and two inflammation markers in the UK Biobank. Given that risk factors can manifest differently across social and environmental contexts (Nianogo *et al.*, 2022; Proto and Quintana-Domeque, 2021; Williams *et al.*, 2012), we separately analyzed the effect of loneliness in three subsets of the UK Biobank: individuals who identified as White and British (hereafter "White British"), individuals who identified as Indian, and individuals who identified as Black or Black British (hereafter "Black British"). For each subset, we computed both GRM- and PENGUIN-corrected estimates of loneliness. To compute PENGUIN-corrected estimates, we used GWAS summary statistics generated from White British individuals, the only subset large enough to perform GWAS.

Across all three groups, we generally find that loneliness either has no effect or is associated with poorer health outcomes, corroborating existing literature (Figure 4). We find that GRM- and PENGUIN-corrected estimates are most concordant for White British individuals ($r^2 = 0.96$, compared with 0.85 for Black British and 0.92 for Indian). This likely reflects the fact that PENGUIN-corrected estimates face portability challenges when individual-level phenotypes and GWAS summary statistics are obtained from two different populations.

This has implications for interpreting the effect of loneliness on health outcomes. For example,

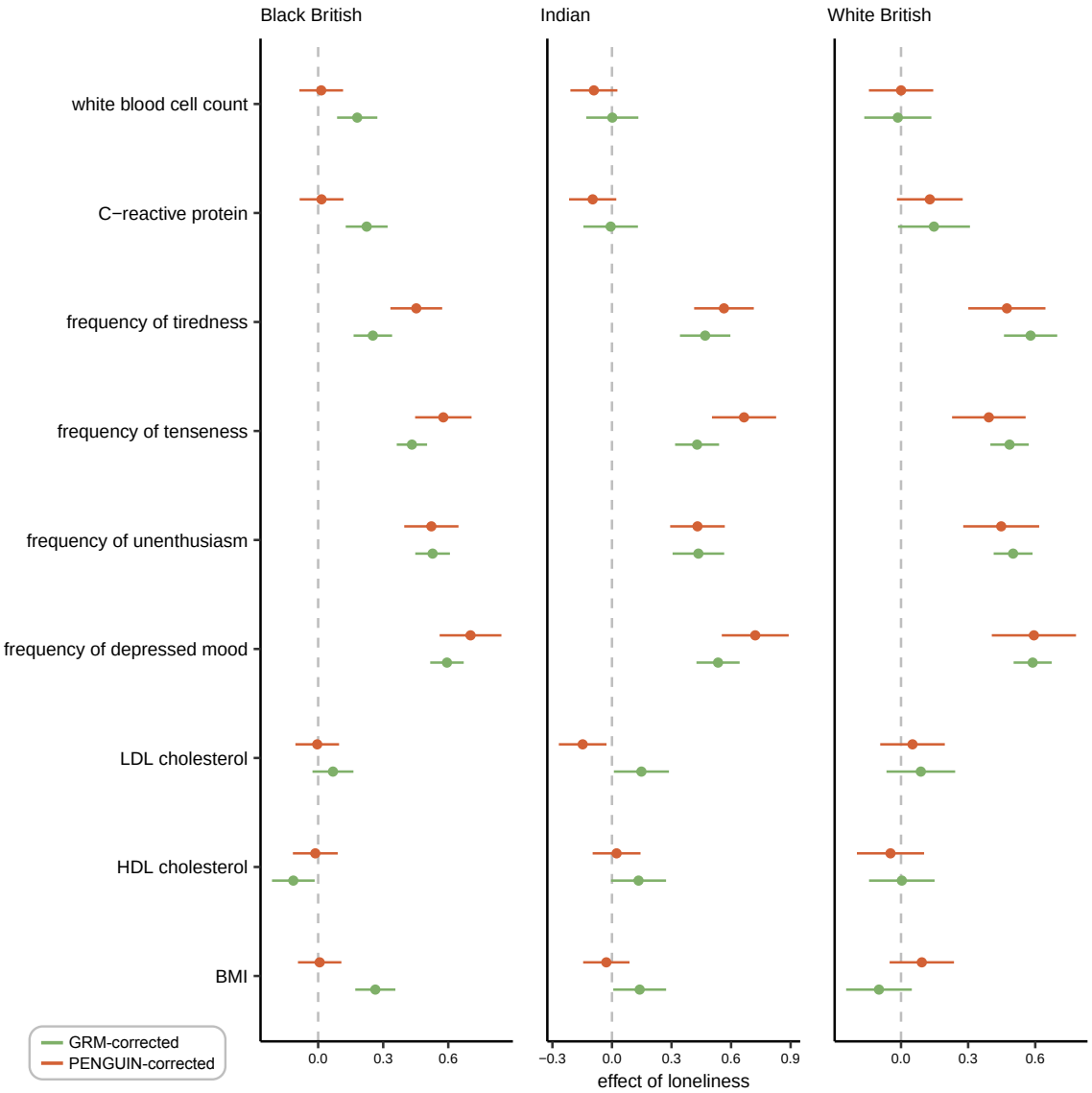

Figure 4: **Effect of loneliness on health outcomes in UK Biobank.** GRM- and PENGUIN-corrected estimates of the effect of loneliness on inflammatory markers, mental health traits, and cardiovascular risk factors in the UK Biobank Black British, Indian, and White British. PENGUIN-corrected estimates use individual-level data from the specified group and GWAS summary statistics from the UK Biobank White British. Error bars correspond to mean ± 1 standard deviation.

PENGUIN trained on a White British GWAS finds no effect of loneliness on inflammation markers in Black British individuals. However, GRM-corrected estimates indicate that loneliness is associated with both increased white blood cell count and increased C-reactive protein. This is particularly notable because loneliness does not appear to have any effect on inflammation in the Indian or White British subsets – thus, the association between loneliness and inflammation is only evident with GRM-corrected estimates, and only in the Black British subset. In some rare instances, we also find that PENGUIN-corrected estimates and GRM-corrected estimates have opposite signs, though we caution that these differences are not significant.

Discussion

We presented an approach that uses the GRM to control genetic confounding in observational epidemiological studies. This approach was inspired by analogous problems in phylogenetics, where researchers account for shared evolutionary history when testing for trait associations across species. Our results highlight the deep connections between statistical genetics and phylogenetics, and echoing earlier work (Bulik-Sullivan, 2015), they also reiterate the conceptual similarity between LDSC- and GRM-based estimates of heritability. Here, we show that the estimator obtained by the LDSC-based method PENGUIN is equivalent to the estimator obtained when the GRM is used to control genetic confounding. Thus, although our motivation stemmed from recent work on the similarity between the GRM and phylogenetic variance-covariance matrix (Schraiber *et al.*, 2024), one could easily arrive at the same method by using the well-known correspondence between LDSC- and GRM-based approaches. Although our approach to the problem via phylogenetics might be seen as roundabout—the key thing is to correct for the genetic correlation, and this correlation is estimated by bivariate GREML—the possibility of our approach does not appear in previous work (to our knowledge), even the pathbreaking work leading to PENGUIN.

In this work, we focused on analyzing a particular causal model (Fig. 1A), and on estimating the interventional effect of the exposure. Under this same causal model, it would be trivial to use our approach to obtain different measures of exposure effect, such as the combined estimate of the genetic and interventional effect of the exposure. Under alternative causal models—for example, if the effect of genetics on the outcome is entirely mediated through the exposure—the suitability of GRM- and PENGUIN-corrected estimates of exposure effect remains to be characterized.

The GRM is central to our approach and merits further discussion. There are different ways to construct GRMs, corresponding to different assumptions about how genetic similarity between

individuals translates into phenotypic similarity. Our analyses used the canonical GRM, which is generally constructed as the variance-covariance matrix of centered, standardized genotypes, and corresponds to an assumption that the magnitude of variant effect size is inversely proportional to the frequency of the variant. In simulations, we find that the canonical GRM is generally sufficient to control genetic confounding, even when the data are generated under different models. Nevertheless, other GRMs could easily be used in the context of observational epidemiological studies, and their performance is an object of future study.

Our results provide several lessons for empirical studies. First, our results showcase the value of using the GRM to control genetic confounding. Both the Black British and Indian subsets are too small for in-sample GWAS, which means that methods like PENGUIN require out-of-sample GWAS that risk portability concerns. Second, we find differences in the effect of loneliness on several health and social outcomes across the Black British, Indian, and White British individuals in the UK Biobank. This highlights the importance of cohort-specific analyses in epidemiological settings where the goal is to design effective interventions. The differences we observe could represent either gene-by-environment interactions or environment-by-environment interactions, given that both genetic and non-genetic factors might vary across and within groups. Leveraging the GRM to characterize such interactions is an important area for future work.

Overall, our work demonstrates that the GRM is particularly useful in settings in which large, accurate GWAS are not available for a given trait or population. In settings where biobank-scale data or GWAS summary statistics for the specific trait and population are readily available, existing methods such as PENGUIN are well-suited to the task. For understudied traits or populations, however, observational epidemiological studies will benefit from using GRM-corrected estimates of exposure effect.

Materials and methods

Estimating exposure effect

GRM-corrected estimates

To construct a GRM for a particular set of individuals, we used all biallelic variants with minor allele frequency greater than 0.01. Unless specified otherwise, the GRM was constructed with an algorithm corresponding to the assumption that all variants contribute equally to heritability (GCTA flag --make-grm-alg 0).

To compute GRM-corrected estimates of the interventional exposure effect, we used bivariate GREML (Lee *et al.*, 2012) to decompose the variance of the exposure and outcome into genetic and non-genetic components. We estimated the exposure effect using the estimated variance-covariance matrix for the non-genetic components of the exposure and outcome. Specifically, we estimated b as

$$b = \frac{\hat{\sigma}_{e_{XY}}}{\hat{\sigma}^2_{e_X}},$$

where $\hat{\sigma}_{e_{XY}}$ is the estimated covariance between the non-genetic components of the exposure and outcome, and $\hat{\sigma}^2_{e_X}$ is the estimated variance of the non-genetic component of the exposure.

To obtain standard errors of the GRM-corrected estimate, we used two approaches. First, we computed the standard deviation of the estimate across multiple replicates in simulations. Second, we computed the standard error of the estimate using the delta method:

$$\mathrm{Var}\left(\frac{\hat{\sigma}_{e_{XY}}}{\hat{\sigma}^2_{e_X}}\right) = \left(\frac{\hat{\sigma}_{e_{XY}}}{\hat{\sigma}^2_{e_X}}\right)^2 \cdot \left(\frac{\mathrm{Var}(\hat{\sigma}_{e_{XY}})}{(\hat{\sigma}_{e_{XY}})^2} + \frac{\mathrm{Var}(\hat{\sigma}^2_{e_X})}{(\hat{\sigma}^2_{e_X})^2} - 2\frac{\mathrm{Cov}(\hat{\sigma}_{e_{XY}}, \hat{\sigma}^2_{e_X})}{\hat{\sigma}_{e_{XY}}\hat{\sigma}^2_{e_X}}\right)$$

To obtain the variance and covariance of the estimated components, we lightly modified GCTA (https://github.com/roshnipatel/GCTA) to report the Fisher information matrix, which is already computed by the software during each REML iteration. We verified in simulations that both standard errors were comparable (Supp Fig S3).

PENGUIN-corrected estimates

To compute PENGUIN-corrected estimates in simulated data, we first generated GWAS summary statistics for biallelic variants with minor allele frequency greater than 0.01 using plink2 (Chang *et al.*, 2015). We next generated LD scores for our simulated genotypes using LDSC (Bulik-Sullivan *et al.*, 2015) with a window size of 100 kb. Using our GWAS summary statistics and LD scores, we ran individual-level PENGUIN with default parameters. In simulations of random mating, we did not include additional covariates when running GWAS or PENGUIN. In simulations of population structure or assortative mating, we included one principal component when generating GWAS summary statistics and subsequently computing PENGUIN-corrected estimates of exposure effect.

PGS-corrected estimates

Due to the many technical decisions necessary to construct polygenic scores (Lin *et al.*, 2025), we sought to place an upper bound on the utility of PGS-corrected estimates through semi-analytical means. We simulated trait architecture and genotypes and phenotypes for 100,000 individuals as

described below (see *Random mating*). We then simulated the performance of PGS-corrected estimates in scenarios in which 20% and 100% of causal variants are identified in GWAS and included in the polygenic score. To do so, we started by randomly sampling the aforementioned proportion of causal variants, given that all variants contribute equally to heritability in expectation under the model we assume (see *Random mating*). In particular, for the subset of causal variants identified in GWAS, we simulated effect sizes estimated with noise using a standard approach (Vilhjálmsson *et al.*, 2015). Estimated effect sizes $\hat{\beta}$ were drawn from a normal distribution centered on the true effect size, with a variance that depends on the true effect size, the allele frequency of the variant, f; and the number of individuals in the GWAS, n.

$$\hat{\beta} \sim N\Big(\beta, \frac{1 - 2f(1-f)\beta^2}{2nf(1-f)}\Big)$$

Using the estimated effect sizes $\hat{\beta}$, we then simulated an estimated polygenic score for each individual. Using this estimated polygenic score, we computed the PGS-corrected estimate of the exposure effect for ten replicates. We reported the mean of the estimated exposure effect across all ten replicates.

Simulating genotypes and phenotypes

Random mating

Given that our empirical analyses rely on GWAS summary statistics obtained from the UK Biobank White British cohort, we conducted the majority of simulation analyses in simulated CEU individuals, except when analyzing portability (see *Analyzing portability* below). We simulated genotypes on chromosome 22 for CEU individuals under a published demographic model using the msprime engine in stdpopsim (Adrion *et al.*, 2020; Baumdicker *et al.*, 2022; Gutenkunst *et al.*, 2009). For downstream analyses, we centered but did not standardize genotypes. To compare the utility of GRM- and PENGUIN-corrected estimates of exposure effect across different sample sizes, we randomly partitioned the simulated individuals into a GWAS cohort of 100,000 individuals and a test cohort of 4,000 individuals.

We simulated genetic values for exposure and outcome under an additive genetic architecture. To simulate causal variants, we first filtered variants for biallelic sites with a minor allele frequency greater than 0.01 in the test cohort. Next, we randomly sampled 10,000 variants to be the causal

variants. We simulated effect sizes for the ith variant using a bivariate normal distribution

$$\begin{pmatrix} \beta_{X,i} \\ \beta_{Y,i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{h^2}{2mf_i(1-f_i)} & \frac{\rho h^2}{2mf_i(1-f_i)} \\ \frac{\rho h^2}{2mf_i(1-f_i)} & \frac{h^2}{2mf_i(1-f_i)} \end{pmatrix} \right),$$

where h^2 is the heritability of the exposure trait, ρ is the effect size correlation between exposure and outcome, f_i is the allele frequency of the ith variant in the test cohort, and m is the number of causal variants. This distribution corresponds to the assumption that each causal variant contributes equally to heritability (often known in statistical genetics as "the alpha model" (Speed *et al.*, 2017)). For a subset of results, we also analyzed alternative models with a weaker dependence between effect size and frequency, such that $\mathrm{Var}(\beta_i) \propto (f_i(1-f_i))^\alpha$ for $\alpha \in [0, -0.3, -0.5]$.

We simulated trait-specific environmental effects as $N(0, 1-h^2)$. Conditional on genetic values and environmental effects, we simulated phenotypes for exposure and outcome under the generative model we introduce in the Results:

$$X = G\beta_X + e_X$$

$$Y = G\beta_Y + e_Y + be_X$$

Note that this model ensures that both the variance of genetic values and the heritability of the exposure, X, is equal to h^2, while only the variance of genetic values for the outcome is equal to h^2; the heritability will vary with the magnitude of b.

Population structure

To model population structure, we simulated a population split followed by no migration between subpopulations. Specifically, we simulated a split time of 200 generations ago, and a constant population size of 10,000 individuals for each subpopulation. For downstream analysis, we sampled 52,000 individuals from each subpopulation. We simulated both stratified and unstratified phenotypes, and stratified phenotypes were simulated such that the environmental effect of both exposure and outcome was shifted by 0.5 units.

To compute GRM-corrected estimates, we analyzed a joint sample of 2,000 individuals from each subpopulation. To compute PENGUIN-corrected estimates, we used a joint sample of 50,000 individuals from each subpopulation to perform a GWAS. We then used a (non-overlapping) joint sample of 2,000 individuals from each subpopulation to comprise individual-level data for PENGUIN. We computed LD scores in a random subset of 4,000 individuals from one population, and incorpo-

rated principal components as covariates to correct population structure for both the GWAS and PENGUIN itself.

Assortative mating

We simulated genotypes and phenotypes under assortative mating using the software package xftsim (Border *et al.*, 2024). Given the computational complexity of assortative mating simulations, we were unable to simulate haplotypes at a sample size capable of computing PENGUIN-corrected estimates. To evaluate GRM-corrected estimates in simulations of assortative mating, we simulated founder haplotypes of 4,000 CEU individuals on chromosome 22 using stdpopsim (Adrion *et al.*, 2020). We seeded our simulations with these founder haplotypes and simulated 5 generations of cross-trait assortative mating with correlations ranging from -0.6 to +0.6. We simulated causal variants, effect sizes, and phenotypes under the models described previously (see *Random mating* above for more details).

Analyzing portability

Portability of PENGUIN-corrected estimates

To characterize the portability of PENGUIN-corrected estimates, we simulated 10,000 CEU, CHB, and YRI individuals under the demographic model inferred by Gutenkunst *et al.* (2009). We randomly sampled 10,000 causal variants from the pool of biallelic sites that were segregating at a frequency of 1×10^{-4} in all populations. We simulated multivariate effect sizes independent of frequency (i.e. with $\alpha = 0$), given the impossibility of simulating frequency-dependent effects in a multi-population model. We simulated a genetic correlation of 0.25 between exposure and outcome and an exposure heritability of 0.5 in CEU. Using simulated frequencies and effect sizes, we computed the analytical expectation of the PENGUIN estimator as follows (see Supplementary Note 2 for derivation).

$$\mathbb{E}[\hat{b}_{\mathrm{PENG}}] = \frac{b\sigma_e^2 + \sum_{k=1}^{M} 2\beta_{X,k}\beta_{Y,k}\Big(p_{k,\mathrm{ind}}(1-p_{k,\mathrm{ind}}) - p_{k,\mathrm{GWAS}}(1-p_{k,\mathrm{GWAS}})\Big)}{\sigma_e^2 + \sum_{k=1}^{M} 2\beta_{X,k}^2\Big(p_{k,\mathrm{ind}}(1-p_{k,\mathrm{ind}}) - p_{k,\mathrm{GWAS}}(1-p_{k,\mathrm{GWAS}})\Big)}$$

Across each of 10 simulation replicates, we computed the squared bias as $(b - \mathbb{E}[\hat{b}_{\mathrm{PENG}}])^2$ and reported the mean squared bias across simulations.

Portability of PGS-corrected estimates

To characterize the portability of PGS-corrected estimates, we simulated 100,000 CEU and 10,000 CHB and YRI individuals under the demographic model inferred by Gutenkunst *et al.* (2009). We simulated causal variants and their effect sizes as described in *Portability of PENGUIN-corrected estimates*. We generated GWAS summary statistics with plink2 (Chang *et al.*, 2015) in the set of variants with MAF ≥ 0.01 in CEU, and clumped variants with $p \leq 1 \times 10^{-5}$, $r^2 \leq 0.5$, and a window size of 250 kb. We computed PGS by summing over the product of clumped variants and their estimated effect sizes, and reported the mean squared error in each population.

UK Biobank analyses

We analyzed the effect of loneliness on three cardiovascular traits (BMI, HDL cholesterol, and LDL cholesterol); four mental health traits (frequency of depressed mood, unenthusiasm, tiredness, and tenseness); and two immune-related traits (C-reactive protein and white blood cell count). To encode loneliness phenotypes, we used responses to the question "Do you often feel lonely?", where "No" was encoded as 0, and "Yes" was encoded as 1. To increase the interpretability of exposure effect, we inverse normal transformed phenotypes for quantitative traits (white blood cell count, C-reactive protein, LDL cholesterol, HDL cholesterol, and BMI) and removed outliers with a z-score greater than 3. We separately computed GRM- and PENGUIN-corrected estimates in three subsets of the UK Biobank: individuals who self-identified their ethnic background as Black or Black British (hereforth "Black British"); Indian; and both White and British (hereforth "White British").

To compute GRM-corrected estimates, we constructed a GRM in each subset using all genotyped variants with minor allele frequency greater than 0.01. Given that there were 442,719 individuals who identified as White British, we randomly sampled 7,000 individuals to obtain a tractable sample size for computing GRM-corrected estimates. We performed variance components analysis using bivariate GREML and included covariates that captured age, sex, assessment center, and 10 principal components.

We computed PENGUIN-corrected estimates using GWAS summary statistics obtained from the UK Biobank White British (http://www.nealelab.is/uk-biobank/), and individual-level phenotype data from Black British, Indian, and White British individuals. As before, we included covariates that captured age, sex, and 10 principal components.

Acknowledgments and funding sources

We thanks members of the Edge, Pennell, and Mooney labs for helpful conversations about this work. RAP, JGS, and MDE were supported by NIH grant R35GM137758. MP was supported by NIH grant R35GM151348.

References

Abdellaoui, A., S. Sanchez-Roige, J. Sealock, J. L. Treur, J. Dennis, P. Fontanillas, S. Elson, The 23andme Research Team, M. G. Nivard, H. F. Ip, *et al.*, 2019 Phenome-wide investigation of health outcomes associated with genetic predisposition to loneliness. Human Molecular Genetics **28**: 3853–3865.

Adrion, J. R., C. B. Cole, N. Dukler, J. G. Galloway, A. L. Gladstein, G. Gower, C. C. Kyriazis, A. P. Ragsdale, G. Tsambos, F. Baumdicker, *et al.*, 2020 A community-maintained standard library of population genetic models. eLife **9**: e54967.

Baumdicker, F., G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, *et al.*, 2022 Efficient ancestry and mutation simulation with msprime 1.0. Genetics **220**: iyab229.

Belsky, D. W., T. E. Moffitt, D. L. Corcoran, B. Domingue, H. Harrington, S. Hogan, R. Houts, S. Ramrakha, K. Sugden, B. S. Williams, *et al.*, 2016 The Genetics of Success: How Single-Nucleotide Polymorphisms Associated With Educational Attainment Relate to Life-Course Development. Psychol Sci **27**: 957–972.

Berg, J. J., A. Harpak, N. Sinnott-Armstrong, A. M. Joergensen, H. Mostafavi, Y. Field, E. A. Boyle, X. Zhang, F. Racimo, J. K. Pritchard, *et al.*, 2019 Reduced signal for polygenic adaptation of height in UK Biobank. elife **8**.

Beutel, M. E., E. M. Klein, E. Brähler, I. Reiner, C. Jünger, M. Michal, J. Wiltink, P. S. Wild, T. Münzel, K. J. Lackner, *et al.*, 2017 Loneliness in the general population: prevalence, determinants and relations to mental health. BMC Psychiatry **17**: 97.

Border, R., S. O'Rourke, T. de Candia, M. E. Goddard, P. M. Visscher, L. Yengo, M. Jones, and M. C. Keller, 2022 Assortative mating biases marker-based heritability estimators. Nat Commun **13**: 660.

Border, R., J. Wang, C. Caggiano, S. Sankararaman, A. J. Schork, P. Turley, A. S. Young, D. J. Benjamin, A. W. Dahl, and N. Zaitlen, 2024 Simple models of non-random mating and environmental transmission bias standard human genetics statistical methods. Pages: 2024.10.16.618755 Section: New Results.

Bulik-Sullivan, B., 2015 Relationship between LD Score and Haseman-Elston Regression. Pages: 018283 Section: New Results.

Bulik-Sullivan, B. K., P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, 2015 LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet **47**: 291–295.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4**: s13742–015–0047–8.

Choi, D., K. J. Tsuchiya, and N. Takei, 2019 Interaction effect of oxytocin receptor (OXTR) rs53576 genotype and maternal postpartum depression on child behavioural problems. Sci Rep **9**: 7685.

Conley, D., 2016 Socio-Genomic Research Using Genome-Wide Molecular Data. Annu. Rev. Sociol. **42**: 275–299.

Cuevas, A. G., F. D. Mann, D. R. Williams, and R. F. Krueger, 2021 Discrimination and anxiety: Using multiple polygenic scores to control for genetic liability. Proceedings of the National Academy of Sciences **118**: e2017224118.

Dennis, J., J. Sealock, R. T. Levinson, E. Farber-Eger, J. Franco, S. Fong, P. Straub, D. Hucks, W.-L. Song, M. F. Linton, *et al.*, 2021 Genetic risk for major depressive disorder and loneliness in sex-specific associations with coronary artery disease. Mol Psychiatry **26**: 4254–4264.

DiPrete, T. A., C. A. P. Burik, and P. D. Koellinger, 2018 Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. Proceedings of the National Academy of Sciences **115**: E4970–E4979.

D'Onofrio, B. M., A. Sjölander, B. B. Lahey, P. Lichtenstein, and A. S. Öberg, 2020 Accounting for Confounding in Observational Studies. Annu. Rev. Clin. Psychol. **16**: 25–48.

D'Onofrio, B. M., Q. A. Class, B. B. Lahey, and H. Larsson, 2014 Testing the Developmental Origins of Health and Disease Hypothesis for Psychopathology Using Family-Based Quasi-Experimental Designs. Child Dev Perspect **8**: 151–157.

Felsenstein, J., 1985 Phylogenies and the Comparative Method. The American Naturalist **125**: 1–15.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet **5**: e1000695.

Hakulinen, C., L. Pulkki-Råback, M. Virtanen, M. Jokela, M. Kivimäki, and M. Elovainio, 2018 Social isolation and loneliness as risk factors for myocardial infarction, stroke and mortality: UK Biobank cohort study of 479 054 men and women. Heart **104**: 1536–1542.

Harden, K. P. and P. D. Koellinger, 2020 Using genetics for social science. Nat Hum Behav **4**: 567–576.

Jaffee, S. R. and T. S. Price, 2012 The implications of genotype–environment correlation for establishing causal processes in psychopathology. Development and Psychopathology **24**: 1253–1264.

Lee, S., J. Yang, M. Goddard, P. Visscher, and N. Wray, 2012 Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics **28**: 2540–2542.

Lee, S. L., E. Pearce, O. Ajnakina, S. Johnson, G. Lewis, F. Mann, A. Pitman, F. Solmi, A. Sommerlad, A. Steptoe, *et al.*, 2021 The association between loneliness and depressive symptoms among adults aged 50 years and older: a 12-year population-based cohort study. The Lancet Psychiatry **8**: 48–57.

Lin, Y.-S., T. Tan, Y. Wang, B. Pasaniuc, A. R. Martin, and E. G. Atkinson, 2025 Differential performance of polygenic prediction across traits and populations depending on genotype discovery approach. Pages: 2025.03.18.644029 Section: New Results.

Liu, Z., Y. Song, S. Shi, Z. Zheng, W. Song, W. Zhang, and K. Dou, 2024 Association between loneliness and incident atrial fibrillation across different genetic predisposition: findings from the UK Biobank. Eur J Med Res **29**: 1–11.

Lynch, M., 1991 Methods for the Analysis of Comparative Data in Evolutionary Biology. Evolution **45**: 1065–1080.

Manousaki, D., A. Harroud, R. E. Mitchell, S. Ross, V. Forgetta, N. J. Timpson, G. D. Smith, C. Polychronakos, and J. B. Richards, 2021 Vitamin D levels and risk of type 1 diabetes: A Mendelian randomization study. PLoS Med **18**: e1003536.

Martin, A. R., M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, 2019 Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet **51**: 584–591.

Martins, E. P. and T. F. Hansen, 1997 Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. The American Naturalist **149**: 646–667.

Matthews, T., L. J. H. Rasmussen, A. Ambler, A. Danese, J. Eugen-Olsen, D. Fancourt, H. L. Fisher, K. K. Iversen, M. Schultz, K. Sugden, *et al.*, 2024 Social isolation, loneliness, and inflammation: A multi-cohort investigation in early and mid-adulthood. Brain Behav Immun **115**: 727–736.

Mills, M. C. and C. Rahal, 2019 A scientometric review of genome-wide association studies. Commun Biol **2**: 1–11.

Moshtael, R., M. E. Lynch, G. E. Duncan, and C. R. Beam, 2024 A Genetically Informed Study of the Association Between Perceived Stress and Loneliness. Behav Genet **54**: 268–277.

Mostafavi, H., A. Harpak, I. Agarwal, D. Conley, J. K. Pritchard, and M. Przeworski, 2020 Variable prediction accuracy of polygenic scores within an ancestry group. eLife **9**: e48376.

Nianogo, R. A., A. Rosenwohl-Mack, K. Yaffe, A. Carrasco, C. M. Hoffmann, and D. E. Barnes, 2022 Risk Factors Associated With Alzheimer Disease and Related Dementias by Sex and Race and Ethnicity in the US. JAMA Neurol **79**: 584–591.

Office of the Surgeon General, 2023 Our Epidemic of Loneliness and Isolation. Technical report, US Department of Health and Human Services, Washington, D.C.

Patel, R. A., S. A. Musharoff, J. P. Spence, H. Pimentel, C. Tcheandjieu, H. Mostafavi, N. Sinnott-Armstrong, S. L. Clarke, C. J. Smith, V.A. Million Veteran Program, *et al.*, 2022 Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. Am J Hum Genet **109**: 1286–1297.

Patel, R. A., C. L. Weiß, H. Zhu, H. Mostafavi, Y. B. Simons, J. P. Spence, and J. K. Pritchard, 2025 Characterizing selection on complex traits through conditional frequency spectra. Genetics **229**: iyae210.

Paul, E., F. Bu, and D. Fancourt, 2021 Loneliness and Risk for Cardiovascular Disease: Mechanisms and Future Directions. Curr Cardiol Rep **23**: 1–7.

Pingault, J.-B., A. G. Allegrini, T. Odigie, L. Frach, J. R. Baldwin, F. Rijsdijk, and F. Dudbridge, 2022 Research Review: How to interpret associations between polygenic scores, environmental risks, and phenotypes. Journal of Child Psychology and Psychiatry **63**: 1125–1139.

Pingault, J.-B., F. Rijsdijk, T. Schoeler, S. W. Choi, S. Selzam, E. Krapohl, P. F. O'Reilly, and F. Dudbridge, 2021 Genetic sensitivity analysis: Adjusting for genetic confounding in epidemiological associations. PLOS Genetics **17**: e1009590.

Proto, E. and C. Quintana-Domeque, 2021 COVID-19 and mental health deterioration by ethnicity and gender in the UK. PLOS ONE **16**: e0244419.

Rødevand, L., S. Bahrami, O. Frei, A. Lin, O. Gani, A. Shadrin, O. B. Smeland, K. S. O. Connell, T. Elvsåshagen, A. Winterton, *et al.*, 2021 Polygenic overlap and shared genetic loci between loneliness, severe mental disorders, and cardiovascular disease risk factors suggest shared molecular mechanisms. Transl Psychiatry **11**: 3.

Schmitz, L. and D. Conley, 2017 Modeling Gene-Environment Interactions With Quasi-Natural Experiments. Journal of Personality **85**: 10–21.

Schraiber, J. G., M. D. Edge, and M. Pennell, 2024 Unifying approaches from statistical genetics and phylogenetics for mapping phenotypes in structured populations. PLOS Biology **22**: e3002847.

Sinnott-Armstrong, N., S. Naqvi, M. Rivas, and J. K. Pritchard, 2021 GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. eLife **10**: e58615.

Smith, G. D. and S. Ebrahim, 2003 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol **32**: 1–22.

Smith, G. D. and G. Hemani, 2014 Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet **23**: R89–R98.

Sohail, M., R. M. Maier, A. Ganna, A. Bloemendal, A. R. Martin, M. C. Turchin, C. W. Chiang, J. Hirschhorn, M. J. Daly, N. Patterson, *et al.*, 2019 Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. elife **8**.

Speed, D. and D. J. Balding, 2015 Relatedness in the post-genomic era: is it still useful? Nat Rev Genet **16**: 33–44.

Speed, D., N. Cai, UCLEB Consortium, M. R. Johnson, S. Nejentsev, and D. J. Balding, 2017 Reevaluation of SNP heritability in complex human traits. Nat Genet **49**: 986–992.

Uddin, M. J., C. Hjorthøj, T. Ahammed, M. Nordentoft, and C. T. Ekstrøm, 2022 The use of polygenic risk scores as a covariate in psychological studies. Methods in Psychology **7**: 100099.

Uyeda, J. C., R. Zenil-Ferguson, and M. W. Pennell, 2018 Rethinking phylogenetic comparative methods. Syst Biol **67**: 1091–1109.

Van Bogart, K., C. G. Engeland, M. J. Sliwinski, K. D. Harrington, E. L. Knight, R. Zhaoyang, S. B. Scott, and J. E. Graham-Engeland, 2022 The Association Between Loneliness and Inflammation: Findings From an Older Adult Sample. Front Behav Neurosci **15**: 801746.

Vilhjálmsson, B., J. Yang, H. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, *et al.*, 2015 Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet **97**: 576–592.

Wang, J. Y., N. Lin, M. Zietz, J. Mares, V. M. Narasimhan, P. J. Rathouz, and A. Harpak, 2024 Three Open Questions in Polygenic Score Portability.

Wang, Y., J. Guo, G. Ni, J. Yang, P. M. Visscher, and L. Yengo, 2020 Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun **11**: 3865.

Westoby, M., L. Yates, B. Holland, and B. Halliwell, 2023 Phylogenetically conservative trait correlation: Quantification and interpretation. Journal of Ecology **111**: 2105–2117.

Williams, M. T., L. K. Chapman, J. Wong, and E. Turkheimer, 2012 The role of ethnic identity in symptoms of anxiety and depression in African Americans. Psychiatry Research **199**: 31–36.

Xia, N. and H. Li, 2018 Loneliness, Social Isolation, and Cardiovascular Health. Antioxidants & Redox Signaling **28**: 837–851.

Yair, S. and G. Coop, 2022 Population differentiation of polygenic score predictions under stabilizing selection. Phil. Trans. R. Soc. B **377**: 20200416.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat Genet **42**: 565–569.

Yengo, L., S. Vedantam, E. Marouli, J. Sidorenko, E. Bartell, S. Sakaue, M. Graff, A. U. Eliasen, Y. Jiang, S. Raghavan, *et al.*, 2022 A saturated map of common genetic variants associated with human height. Nature **610**: 704–712.

Zhao, Z., X. Yang, S. Dorn, J. Miao, S. H. Barcellos, J. M. Fletcher, and Q. Lu, 2024 Controlling for polygenic genetic confounding in epidemiologic association studies. Proc. Natl. Acad. Sci. U.S.A. **121**: e2408715121.

Zheng, B., Q. Lu, and J. Fletcher, 2023 Estimating Causal Effects of Fertility on Life Course Outcomes: Evidence Using A Dyadic Genetic Instrumental Variable Approach. Working Paper w30955, National Bureau of Economic Research, Cambridge, MA.

Supplementary Note 1

We will show that the generative statistical model described in the main text underlies existing methods to control genetic confounding in observational epidemiological studies. As in the main text, suppose we have two traits X and Y, such that:

$$X = G\beta_X + e_X \tag{3}$$

$$Y = G\beta_Y + e_Y + be_X \tag{4}$$

where G is the centered genotype matrix; β_X and β_Y are the additive SNP effects; and e_X and e_Y are the non-genetic effects on the traits. We will refer to X as the exposure and Y as the outcome. We assume e_X and e_Y are independently and identically distributed across individuals with a mean of 0, and we further assume that non-genetic effects e_X and e_Y are independent of G. We assume β_X and β_Y have a mean of 0, and make no other assumptions, such that β_X and β_Y could be independent or correlated.

Under this model, the desired estimate of exposure effect is given by:

$$b = \frac{\mathrm{Cov}(Y, e_X)}{\mathrm{Var}(e_X)} \tag{5}$$

Given that e_X is generally unmeasurable, we approximate b by using the GRM to estimate $\mathrm{Cov}(Y, e_X)$ and $\mathrm{Var}(e_X)$ via bivariate variance components analysis.

It is straightforward to show that this is the same model assumed by the PENGUIN statistical framework (Zhao *et al.*, 2024). The PENGUIN estimator is introduced as:

$$b = \frac{\mathrm{Cov}(Y, X) - \mathrm{Cov}(G\beta_Y, G\beta_X)}{\mathrm{Var}(X) - \mathrm{Var}(G\beta_X)}$$

Using the independence of β_X, β_Y, e_X, and e_Y, we have

$$\begin{aligned} b &= \frac{\mathrm{Cov}(Y, X) - \mathrm{Cov}(G\beta_Y, G\beta_X) - \mathrm{Cov}(be_X + e_Y, G\beta_X)}{\mathrm{Var}(X) - \mathrm{Var}(G\beta_X)} \\ &= \frac{\mathrm{Cov}(Y, X - G\beta_X)}{\mathrm{Var}(X) - \mathrm{Var}(G\beta_X)} \\ &= \frac{\mathrm{Cov}(Y, e_X)}{\mathrm{Var}(e_X)}, \end{aligned}$$

and we have thus recovered the interventional effect of the exposure, as defined in Equation 5.

We will next show that this is the same model underlying PGS-based methods to control genetic

confounding. In other words, we will show that PGS-corrected estimates of the exposure effect seek to approximate b as defined above. If we suppose X is a function of additive genetic effects, $G\beta_X$, and non-genetic effects, e_X, a perfect PGS for X is given by $G\beta_X$. PGS-based methods for controlling genetic confounding include the exposure PGS as a covariate in the regression of outcome on exposure:

$$Y = b_X X + b_{PGS} \cdot (G\beta_X) + e_Y$$

Solving for b_X in this regression,

$$b_X = \frac{\mathrm{Var}(G\beta_X)\mathrm{Cov}(X,Y) - \mathrm{Cov}(X,G\beta_X)\mathrm{Cov}(G\beta_X,Y)}{\mathrm{Var}(G\beta_X)\mathrm{Var}(X) - \mathrm{Cov}(X,G\beta_X)^2}$$

Assuming independence of genetic and non-genetic effects,

$$\begin{aligned} b_X &= \frac{\mathrm{Var}(G\beta_X)\mathrm{Cov}(X,Y) - \mathrm{Var}(G\beta_X)\mathrm{Cov}(G\beta_X,Y)}{\mathrm{Var}(G\beta_X)\mathrm{Var}(X) - \mathrm{Var}(G\beta_X)^2} \\ &= \frac{\mathrm{Cov}(X,Y) - \mathrm{Cov}(G\beta_X,Y)}{\mathrm{Var}(X) - \mathrm{Var}(G\beta_X)} \\ &= \frac{\mathrm{Cov}(X - G\beta_X,Y)}{\mathrm{Var}(X) - \mathrm{Var}(G\beta_X)} \\ &= \frac{\mathrm{Cov}(e_X,Y)}{\mathrm{Var}(e_X)}, \end{aligned}$$

and we have again recovered the interventional effect of the exposure, as defined in Equation 5. Thus, PGS-corrected estimates are equivalent to b, the interventional effect of the exposure when the exposure PGS is identical to the true genetic value. However, note that this is not true in the absence of the perfect PGS for the exposure. For example, regressing on the perfect PGS for the outcome does not control genetic confounding, as previously suggested (Zhao *et al.*, 2024).

Supplementary Note 2

We will evaluate the portability of PENGUIN-corrected estimates by deriving an analytical expression for the PENGUIN estimator when GWAS summary statistics are obtained from one cohort and individual-level data is obtained from a second cohort.

As before, suppose we have an exposure X and an outcome Y influenced by genetic factors G

and non-genetic factors e_X and e_Y:

$$
\begin{aligned}
X_{\text{GWAS}} &= G_{\text{GWAS}}\beta_X + e_X \\
Y_{\text{GWAS}} &= G_{\text{GWAS}}\beta_Y + be_X + e_Y \\
X_{\text{ind}} &= G_{\text{ind}}\beta_X + e_X \\
Y_{\text{ind}} &= G_{\text{ind}}\beta_Y + be_X + e_Y
\end{aligned}
$$

We will assume that the non-genetic effect of the exposure, b, is constant across both cohorts in absolute units. We will also assume that the total variance contributed by non-genetic factors (e.g. $\text{Var}(e_X)$ and $\text{Var}(e_Y)$ is shared across both cohorts. (Note, of course, that in real-world situations, environmental variance can differ among groups (Mostafavi *et al.*, 2020; Wang *et al.*, 2024).)
In a scenario where the PENGUIN estimator is constructed from two different cohorts, the estimator is:

$$
\hat{b}_{\text{PENG}} = \frac{\text{Cov}(X_{\text{ind}}, Y_{\text{ind}}) - \text{Cov}(G_{\text{GWAS}}\beta_X, G_{\text{GWAS}}\beta_Y)}{\text{Var}(X_{\text{ind}}) - \text{Var}(G_{\text{GWAS}}\beta_X)}
$$

We will show that in expectation, $\hat{b}_{\text{PENG}}$ is not necessarily equal to b. If we take the expectation of $\hat{b}_{\text{PENG}}$ over G, e_X, and e_Y, and assuming independence of G, e_X, and e_Y, we have:

$$
\mathbb{E}[\hat{b}_{\text{PENG}}] \approx \frac{\text{Cov}(G_{\text{ind}}\beta_X, G_{\text{ind}}\beta_Y) - \text{Cov}(G_{\text{GWAS}}\beta_X, G_{\text{GWAS}}\beta_Y) + b\sigma_e^2}{\text{Var}(G_{\text{ind}}\beta_X) - \text{Var}(G_{\text{GWAS}}\beta_X) + \sigma_e^2}
$$

If we assume that genotypes at causal loci are independent within each cohort, we can simplify further:

$$
\begin{aligned}
\mathbb{E}[\hat{b}_{\text{PENG}}] &\approx \frac{b\sigma_e^2 + \sum_{k=1}^{M} \beta_{X,k}\beta_{Y,k}\Big(\text{Var}(G_{k,\text{ind}}) - \text{Var}(G_{k,\text{GWAS}})\Big)}{\sigma_e^2 + \sum_{k=1}^{M} \beta_{X,k}^2\Big(\text{Var}(G_{k,\text{ind}}) - \text{Var}(G_{k,\text{GWAS}})\Big)} \\
&\approx \frac{b\sigma_e^2 + \sum_{k=1}^{M} 2\beta_{X,k}\beta_{Y,k}\Big(p_{k,\text{ind}}(1-p_{k,\text{ind}}) - p_{k,\text{GWAS}}(1-p_{k,\text{GWAS}})\Big)}{\sigma_e^2 + \sum_{k=1}^{M} 2\beta_{X,k}^2\Big(p_{k,\text{ind}}(1-p_{k,\text{ind}}) - p_{k,\text{GWAS}}(1-p_{k,\text{GWAS}})\Big)}
\end{aligned}
$$

It should be fairly straightforward that, if $p_{k,\text{GWAS}}$ and $p_{k,\text{ind}}$ are identical, $\mathbb{E}[\hat{b}_{\text{PENG}}] = b$. Alternately, if $p_{k,\text{GWAS}}$ and $p_{k,\text{ind}}$ are not identical, and per-SNP heritabilities differ between the two cohorts, $\mathbb{E}[\hat{b}_{\text{PENG}}] \neq b$.

Supplementary Figures

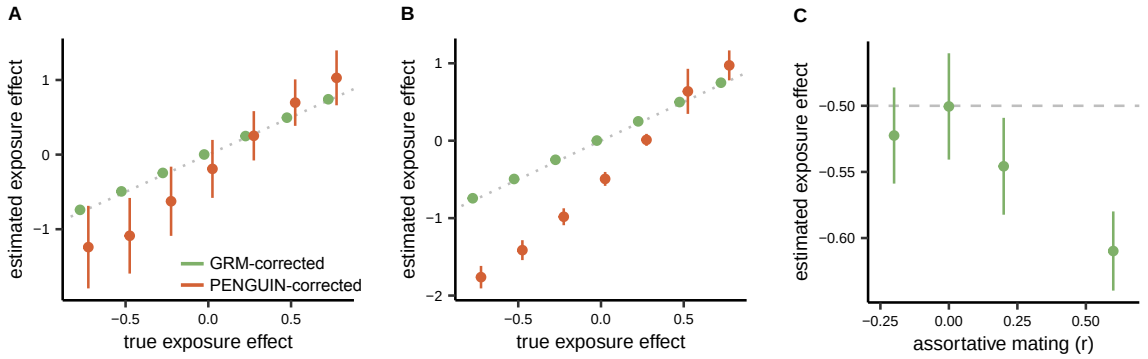

Figure S1: **Non-random mating. A)** GRM- and PENGUIN-corrected estimates of exposure effect in simulations of population structure, but no mean differences in environmental effect. **B)** GRM- and PENGUIN-corrected estimates of exposure effect in simulations of population stratification, i.e. a mean difference of 0.5 units in the environmental component for both exposure and outcome. **C)** GRM-corrected estimates of exposure effect across simulations of assortative mating; the true exposure effect is simulated at -0.5. Across all three simulations, exposure heritabilities were simulated at 0.5 in the absence of assortment and structure. Error bars correspond to mean ± 1 standard deviation.

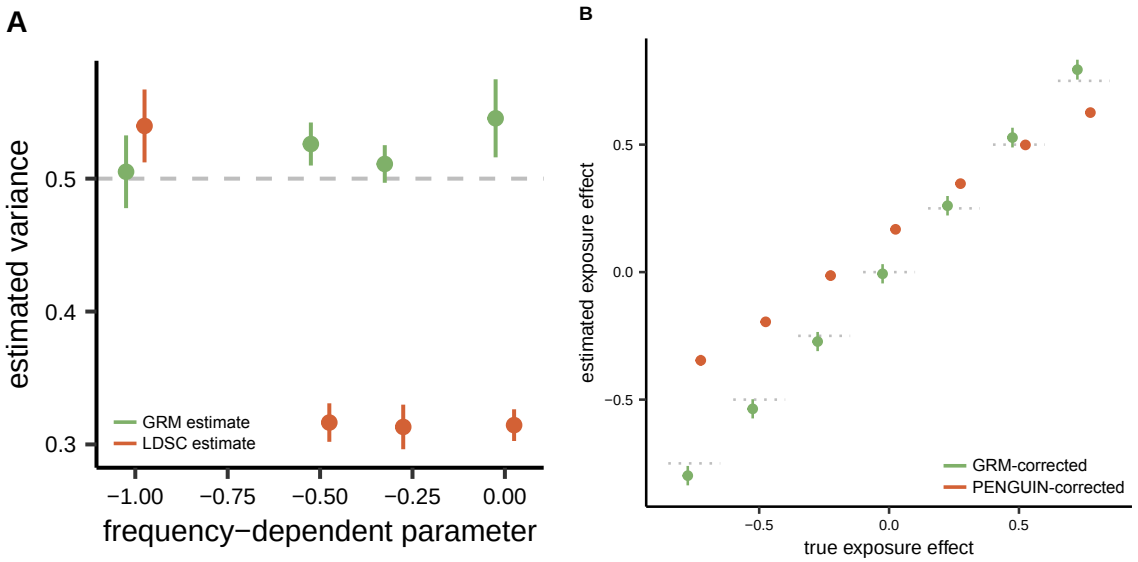

Figure S2: **Heritability models. A)** Estimates of heritability from the GRM and LDSC across simulations with different parameterizations between allele frequency and effect size. The x-axis corresponds to the parameter α (see Methods), such that $\alpha = -1$ represents the default assumption of the GRM and PENGUIN, and $\alpha = 0$ represents frequency-independent effects. The true simulated heritability is 0.5. **B)** GRM- and PENGUIN-corrected estimates of exposure effect when $\alpha = 0$ and allelic effects are independent of frequency. Error bars correspond to mean ± 1 standard deviation.

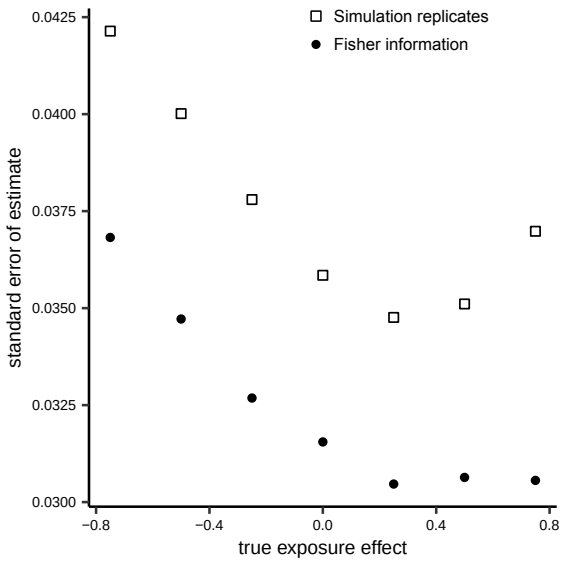

Figure S3: **Comparison of standard errors.** Standard errors of GRM-corrected estimates obtained from 10 simulation replicates compared to standard errors computed via the Fisher information matrix associated with a single simulation replicate.