SOME STATISTICAL METHODS FOR POLYGENIC MODELING AND GENETIC PREDICTION IN LARGE BIOBANKS

by

Jie Song

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: August 29th, 2022

The dissertation is approved by the following members of the Final Oral Committee:
Qiongshi Lu, Associate Professor, Biostatistics & Medical Informatics, UW-Madison
Lauren Schmitz, Assistant Professor, Robert M. La Follette School of Public Affairs, UW-Madison
Jason Fletcher, Professor, Robert M. La Follette School of Public Affairs, UW-Madison
Karl Broman, Professor, Biostatistics & Medical Informatics, UW-Madison
Colin Dewey, Professor, Biostatistics & Medical Informatics, UW-Madison

This thesis is dedicated to my parents and my best friend Minjie, who keep encouraging and supporting me during my doctoral study.

ACKNOWLEDGMENTS

First, I want to express my deep gratitude to my advisor Professor Qiongshi Lu for his expert guidance and mentoring. The vast majority of this thesis has grown out of regular discussions with Professor Lu. I am inspired by his passion for research and his many deep insights in statistical methods. I am also indebted to his character reference and his consistent funding support, which have led to exciting collaborative opportunities in several applied statistical genomics projects. It is my honor and pleasure to work with Professor Lu during my Ph.D study in Madison, WI.

I am also thankful to my committee members Drs. Jason Fletcher, Lauren Schimitz, Colin Dewey and Karl Broman for their many helpful comments and insights.

My appreciation also extends to Drs. Peter Chian, Menggang Yu, Jun Shao for their past guidance that has helped me in pursuing Statistics graduate study in the United States.

Finally, I want to thank my family and friends for their unconditional love and unwavering support throughout all these years, which have meant the world to me.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Polygenic risk scores (PRS) are predictors for individuals' genetic susceptibilities to disease. The standard protocol for generating PRS requires an external genome-wide association study (GWAS) independent from the target cohort, creating a challenge for calculating PRS in large population biobanks due to their frequent inclusion in GWAS. UK Biobank (UKB), one of the largest human genetic cohorts, is often used as GWAS. We investigate the performance of UKB Cross-Validated PRS (CV-PRS) which can be constructed within a single cohort through sample splitting. We explored type-I-error and predictive performance of CV-PRS for 9 UKB phenotypes including education attainment, height, and number of children. When sample size is greater than 100,000, CV-PRS shows little or no type-I-error inflation. We further investigated the effect of sample splitting strategies and find that the number of folds has minor effect on type-I-error and predictive performance. We demonstrate that such approach may be a reasonable strategy to produce PRS in large population cohorts when external GWAS are unavailable.

Genetic nurture refers to a phenomenon that parental genotypes could affect the family environment through their behaviors, and further affect their children's phenotypes. Thus, children's phenotype, such as education attainment, is influenced not only by their own genotypes, but also by their parents' genotypes. We use direct and indirect effect to refer to these two genetic effects. Under this situation, the genetic correlation for two traits can also be partitioned into correlation of direct effects and correlation of indirect effects, and other cross-term correlations. Being able to partition genetic correlation will allow us to get better understanding about the pathway of genetic correlation. We proposed a linear mixed model to estimate genetic covariance of direct/indirect effects, and applied it to UK Biobank siblings on 5 traits. We found significant covariance for educational attainment and household income indirect effects, height and overall health indirect effects, and body mass index and overall health direct effects.

1 INTRODUCTION

Genome-wide association studies (GWAS) have attracted widespread attention in the past two decades. Starting from the completion of Human Genome Project (HGP) in 2003 and the first publication of GWAS in 2005, genome-wide association approach has achieved remarkable success in identifying risk genes related to complex traits and revealing new pathways for disease etiology. Up to 2021, more than 5,700 GWAS have now been reported (Uffelmann et al., 2021), and over 55,000 unique loci have been identified for nearly 5,000 complex traits and diseases (Loos, 2020).

Except for being used to detect signals, GWAS also invoke various applications, including generating polygenic scores (PGS) for more downsteam analysis, *in silico* fine-mapping, pathway analysis, genetic correlation analysis, polygenic risk prediction, and other hypotheses testing analysis (Uffelmann et al., 2021). These applications provide insights into the basis of human genetics and shed light on possibility of clinical use. My thesis mainly focused on PGS and genetic correlation analysis.

PGS play an important role in measuring the genetic contribution to a complex trait and predicting genetic risk for a disease (also referred to polygenic risk scores, PRS), and are especially useful for highly polygenic traits. Rather than being affected by a few of SNPs or genes, some traits such as education, height, and Alzheimer's disease are affected by hundreds or thousands of SNPs. PGS aggregate the contribution of a large number of SNPs, and have been commonly applied in various studies. Till now, PGS studies have demonstrated reliable prediction for many complex genetic phenotypes (Duncan et al., 2019) including blood pressure (Hoffmann et al., 2017), diabetes (Qi et al., 2017) and height (Wood et al., 2014). The construction of PGS requires an appropriate large GWAS whose participants should not be overlapped with PGS samples. Thus, large biobanks such as UK Biobank (UKB) are often used to produce GWAS. However, generating PRS for large biobanks themselves become challenging due to limited external GWAS. One possible solution is to split the whole biobank into non-overlapping subsets, and use different sets for training GWAS and calculating PGS. In chapter 2, we explored the feasibility of this idea, and proposed suggestions for applications.

Besides single-trait analysis, multi-trait analysis based on GWAS also has been widely studied. Genetic correlation as a key metric to quantify the shared genetic basis for two traits has gained popularity in the field. It was traditionally used in animal studies (Lynch et al., 1998; Sodini et al., 2018), and introduced to human genetics in the 1970s-1980s. Except for improving our understanding of complex traits, genetic correlation is also utilized in numerous downstream analysis such as improving prediction accuracy of PRS and exploring possible causality between traits. It can be easily estimated using individual-level data, or GWAS data. Thus, the blooming of GWAS studies leads to a large number of genetic correlations being found, but this also create more challenges in interpreting these correlations, especially when we studied traits related to socioeconomic factors such as education, income and occupation. Interpreting the genetic basis of these traits can be difficult since they are highly affected by the environment or the behaviors from others.

To quantify the possible genetic effects via environment, recent studies have explored ′genetic nurture′ effects that quantify the effects of parental genotypes on their offspring phenotypes, which are also denoted as indirect effects. On the other hand, we denote the genetic effects from own genotypes as direct effects. As such, partitioning genetic effects into two pathways enables us to obtain better understanding of the interplay of genetics and environment. In this way, genetic correlation can also be partitioned into direct and indirect pathways. Chapter 3 of my thesis focused on the partition of genetic correlation using individual-level data.

Besides these major projects, my work also involves multiple interdisciplinary collaborations in a variety of biological fields, including parameter-tuning method using GWAS data (Zhao et al., 2021), review of PRS (Zijie Zhao, 2021), PRS-disease association in Vitamin D for children with cystic fibrosis (Lai et al., 2022), and study of gender differences in the association between parity and cognitive function (submitted).

2 GENERATING POLYGENIC RISK SCORES IN LARGE BIOBANKS THROUGH CROSS-VALIDATION

2.1 Background

Polygenic risk scores (PRS) are genome-wide summaries of genetic propensities for complex traits (Wray et al., 2007; Choi et al., 2020; Zijie Zhao, 2021). PRS constructed from well-powered genome-wide association studies (GWAS) have provided novel insights into the polygenic genetic architecture and pervasive gene-environment interplays underlying numerous diseases and traits (Purcell et al., 2009; Weiner et al., 2017; Barcellos Silvia et al., 2018; Shin and Lee, 2021; Arnau-Soler et al., 2019; Werme et al., 2021b). Recent statistical advances in variant effect estimation have substantially improved the predictive accuracy of PRS (Ge et al., 2019; Vilhjálmsson et al., 2015; Hu et al., 2017). Additionally, ever-growing GWAS sample size, particularly the emergence of large population biobanks, is another major achievement in the field and has greatly accelerated findings in human genetics research. For example, UK Biobank (UKB) (Bycroft et al., 2018) is one of the largest population cohorts in the world, consisting of 500,000 participants from England, Scotland, and Wales, and has contributed to hundreds of GWAS meta-analyses. It is also often of interest to produce PRS and perform analyses using these scores in UKB. For instance, a recent study demonstrated that UKB participants with high PRS for coronary artery disease show substantially elevated disease risk that is comparable to the risk conferred by monogenic mutations known to cause hypercholesterolemia (Khera et al., 2018). Another recent study leveraged the education reform information in UK to demonstrate that PRS can moderate the effects of education on health outcomes (Barcellos Silvia et al., 2018). However, UKB's frequent inclusion in GWAS creates a challenge for PRS analysis in UKB because generating PRS requires an independent GWAS. Although it may be possible to perform GWAS meta-analysis with UKB samples excluded, UKB-excluded summary association statistics are not typically

[0] Co-authors: Fengyi Zheng, Yuchang Wu, Jason M. Fletcher, Qiongshi Lu

provided by studies. Even when non-UKB GWAS exist, they are often from earlier studies with substantially reduced statistical power, which leads to poor predictive performance of PRS.

Cross-validation is a widely used method for estimating prediction error and choosing tuning parameters in predictive modeling (Bates et al., 2021; Shao, 1993; Sivula et al., 2020; Zhang, 1993; Dietterich, 1998; Xu and Liang, 2001; Yang, 2007; Arlot and Celisse, 2010). It provides prediction accuracy estimation by splitting the samples into non-overlapping training and validation sets. A similar data splitting procedure may be applied to biobank cohorts for PRS construction. In a nutshell, we can perform GWAS using training samples, and then use this GWAS to calculate PRS in the validation set. Through rotating the choice of training and validation sets, we will eventually produce PRS for all individuals in the dataset. Thus, this procedure allows PRS construction without an external GWAS. When external GWAS data are available, they can be meta-analyzed with the GWAS produced in the biobank training set to further improve PRS performance. Several studies have explored cross-validation-based PRS. Mefford et al. (Mefford et al., 2020) applied leave-one-out cross-validation to generate reference-free Linear Mixed Model (LMM) PRS. However, this method does not account for dependent samples such as siblings. Mak et al. (Mak et al., 2018) utilized cross-validation and split-validation methods to solve overfitting problem due to overlap between target and discovery data. They showed that these methods can lead to a desirable increase in predictive power than using external GWAS alone, but there is no discussion about the performance under different number of folds and possible overfitting issue due to correlation between different folds. Moreover, they did not mention possible issues related to dependent samples.

In this paper, we assess the empirical validity of cross-validated PRS (CV-PRS) using UKB data. We quantify the impact of sample size, sample relatedness, and number of cross-validation folds on the performance of CV-PRS, and provide several practical guidelines for applying CV-PRS in downstream analysis.

2.2 Results

Overview of the CV-PRS framework

PRS is a weighted sum of allele counts across many (from dozens to millions of) single-nucleotide polymorphisms (SNPs). Typically, its calculation requires summary association statistics from a GWAS as input. Marginal regression coefficients in GWAS summary statistics, in conjunction with linkage disequilibrium (LD) estimated from a reference panel, are often combined to estimate the SNP effect sizes. Once the model is trained (i.e., SNP effect sizes are estimated), these effect size values can be used as SNP weights to calculate PRS in a sample with individual-level genotype data. To avoid overfitting, it is critical that individuals used in PRS calculation do not overlap with the samples in GWAS.

CV-PRS differs from a traditional PRS in that it does not rely on any external GWAS. Instead, the whole sample is equally partitioned into K folds (e.g., K = 4 in Figure 2.1 for illustration). Samples in one fold are used as the target samples for PRS calculation while the remaining K-1 folds are used to perform GWAS. The procedure is then repeated by rotating the target fold until PRS is generated for all samples (Figure 2.1). This procedure is the same as a standard cross-validation exercise except that the primary goal here is to generate PRS for all samples rather than estimating prediction error or selecting the optimal PRS model. Although this method may be useful when no powerful external GWAS is available, it may also lead to sample PRS correlations caused by the partial sample overlap in GWAS used in different folds, resulting in biased standard errors and poorly calibrated type-I error in downstream association analysis. We investigate this issue using simulations.

CV-PRS produced in large samples have well-calibrated type-I error

We first examined whether PRS generated from cross-validation can become spuriously correlated between independent samples in different folds. We generated

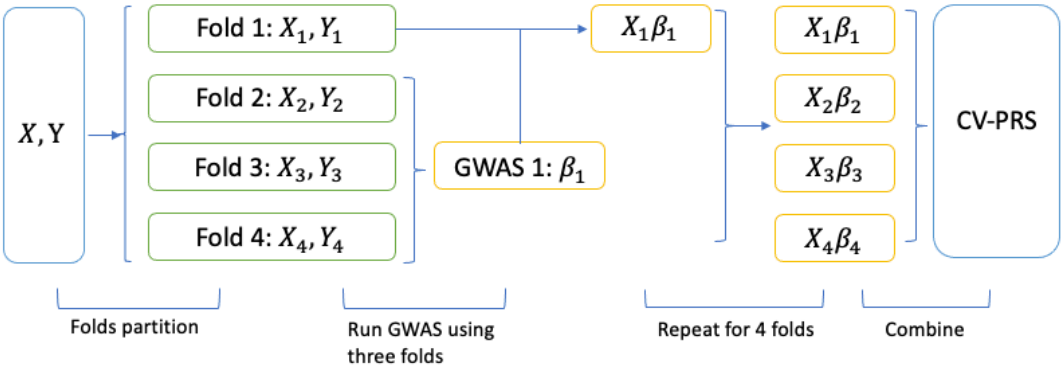

Figure 2.1: CV-PRS flowchart. Samples are partitioned into four folds. We perform a GWAS using data in three folds and calculate PRS for individuals in the remaining fold. Through rotating the choice of target fold and repeating the procedure, we obtain CV-PRS for the entire sample.

four-fold CV-PRS using independent individuals in UKB with sample sizes 1k, 5k, 10k, 50k, and 100k respectively, and compared them to a traditional PRS produced from an external GWAS (EX-PRS). In a total of 200 repeats, each time we simulated a set of phenotypes based on real genotype data, preset SNP effect sizes, and a random error term. We then calculated CV-PRS, and standardized them to have mean 0 and variance of 1 in each repeat. Meanwhile, this set of phenotypes and genotypes were used to generate a GWAS that was applied to a testing set with a sample size of 2k to obtain EX-PRS, and also standardized the EX-PRS to have mean 0 and variance 1. Detailed simulation settings are described in the Methods section.

We estimated the individual-level pair-wise correlations of PRS over 200 repeats among study samples. When the sample size was less than 10k, we observed a clear pattern of block-wise correlations for CV-PRS (Figure 2.2). These blocks corresponded to the four folds in CV-PRS. Notably, both between-fold and within-fold correlations reduced as the total sample size increased, and the pattern became closer to the correlation plots for EX-PRS (Figure A.1). An interpretation is that the estimation accuracy of GWAS for each fold improves as sample size increases, thus the effect sizes estimates are more consistent across the GWAS generated for different folds, i.e., more similar to an external GWAS.

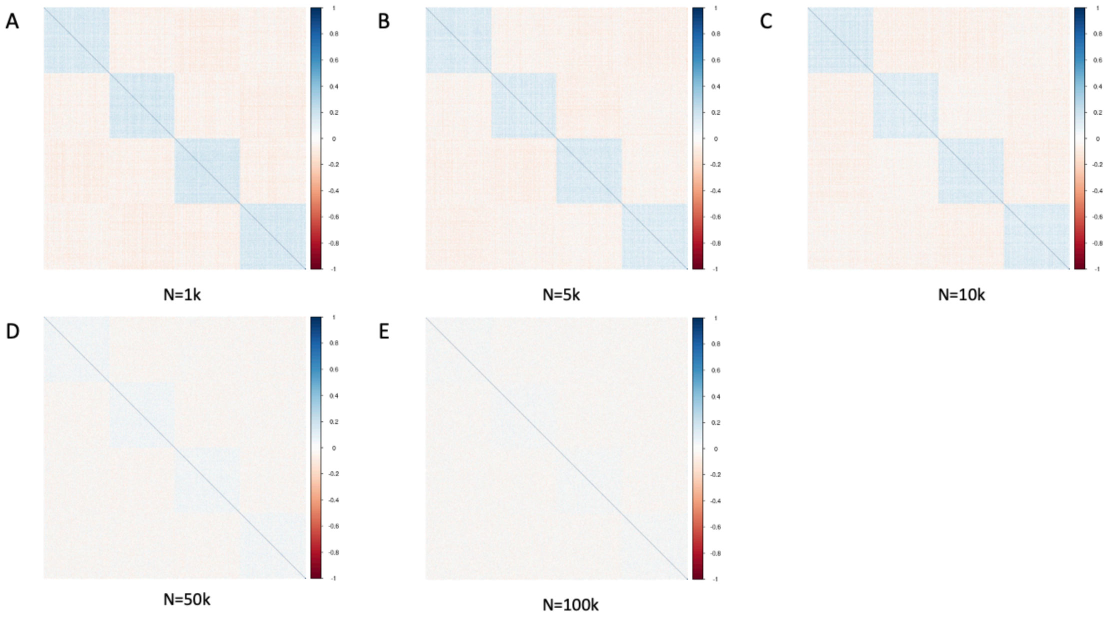

Figure 2.2: Individual-level pair-wise correlation of CV-PRS. Pairwise correlations of CV-PRS were calculated for all sample pairs over 200 random repeats. A-E: sample sizes are 1k, 5k, 10k, 50k, 100k, respectively.

Next, we performed simulations to investigate the type-I error rates in association analysis based on CV-PRS. We used educational attainment (EA) as an example trait to produce CV-PRS in the analysis (Methods). We simulated phenotypes that were genetically independent of EA by randomly selecting 1% SNPs (960k) as causal SNPs, generating their effect sizes from a normal distribution (h2 = 0.4), multiplying effect sizes by corresponding genotypes, and finally adding a random error term. Then, the simulated phenotypes were regressed on CV-PRS of EA to obtain association p-values. This procedure was repeated 200 times. We define type-I error rate as the proportion of p-values that are less than or equal to a specific threshold, i.e. 0.01 or 0.05 in our simulation. We conducted this simulation under sample sizes of 1k, 5k, 10k, 50k, and 100k. Figure 2.3 shows the type-I error results. When sample size was relatively small (e.g. N=1k), the type-I error rates tended to be moderately higher than expected when the threshold was 0.01. As sample size increased to 50k-100k, no type-I error inflation was observed. This trend is consistent with the results of individual-level correlation above. Thus, larger sample size is more desirable for applying CV-PRS. We suggest using a sample size of at least 50k.

Next, we explored the effect of number of cross-validation folds on the type-I error of CV-PRS. We calculated CV-PRS for EA in UKB (N = 376,729) with the number of folds ranging from 4 to 20 (Figure A.2). Regressing these CV-PRS to 200 simulated heritable phenotypes as described above, we did not observe substantial changes in PRS performance across all settings. Thus, we conclude that the number of folds has a minor role in CV-PRS and it may be reasonable to choose a lower number to reduce computational burden.

Sample relatedness affects CV-PRS performance

It is common that participants in large biobanks include related samples. For example, UKB involves about 17k sibling pairs (identified by KING (Manichaikul et al., 2010)) and other related samples. When related individuals are present, application of CV-PRS could be affected by sample splitting strategies. We illustrate this problem using sibling difference model as an example. Sibling difference

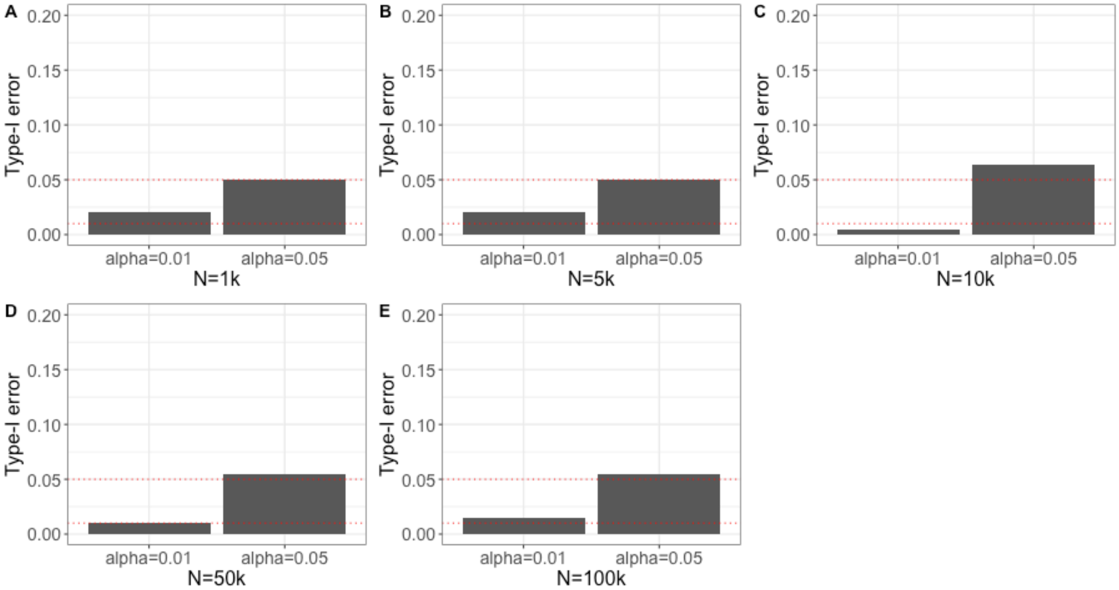

Figure 2.3: Type-I error results for EA CV-PRS based on various sample sizes. Sample sizes are 1k/5k/10k/50k/100k for panels A-E, respectively. Each bar shows the proportion of p-values that were less than or equal to the alpha value (i.e., 0.01 and 0.05).

model is a analytical strategy that regresses the phenotypic differences between siblings on their genotypic differences which is also equivalent to adjusting for family fixed effects in sibling-based regression (Fletcher et al., 2021; Metzger and McDade, 2010). We used CV-PRS of EA to illustrate this problem. We performed two separate regression analyses in UKB for EA phenotype and EA CV-PRS using 1) all independent individuals and sibling pairs (N = 404,435), and 2) only sibling pairs (N = 33,630). We found similar regression coefficients in these analyses (1.14 and 1.33 respectively). However, when we performed the sibling difference analysis on full sibling pairs, the coefficient estimate became substantially different with flipped direction (coefficient = -0.67, Figure 2.4A). As a comparison, we also performed the sibling difference analysis on UKB PRS that used external GWAS (Lee et al. 2018) and obtained a positive association coefficient 0.45 âŁ“ a weakened but still positive effect which is consistent with the literature on within-family PRS effect (Fletcher et al., 2021). We found that this issue was due to overfitting caused by related samples (i.e., siblings) split into different folds during cross-validation (see Supplementary Notes for detailed derivations).A simple fix of this issue is to re-partition samples by family identifier so that siblings from the same family always fall into the same fold in cross-validation. After doing this, we obtained a regression coefficient of 0.48 in sibling difference analysis (Figure 2.4A), which is consistent with population-based analysis results.

We also performed additional simulations to further explore the type-I error rate for CV-PRS when related samples are included. In addition to sample splitting strategy, the proportion of related samples in the analysis could be another factor that affects type-I error. In the analysis we described above, 8.3% of the total samples were sibling pairs. Will a higher proportion of related samples affect association testing even if we partitioned samples by family? With the total sample size set to be 50k, we randomly included 10% and 50% of the samples to be full siblings, and the remaining to be independent individuals. For each setting, we generated EA CV-PRS and 200 heritable phenotypes that were genetically independent from EA using the same procedure described previously. We partitioned samples using two different approaches: 1) all individuals were randomly assigned to different folds; and 2) siblings were partitioned into the same fold and independent samples were randomly assigned. Results were shown in Figures 4B-E. When samples were partitioned randomly with family structure ignored, we found type-I error inflation was observed even when the proportion of related samples was relatively low (10%). With samples partitioned by family structure, we did not observe type-I error inflation. Based on these simulations and results from UKB, we suggest partitioning samples by family in CV-PRS applications.

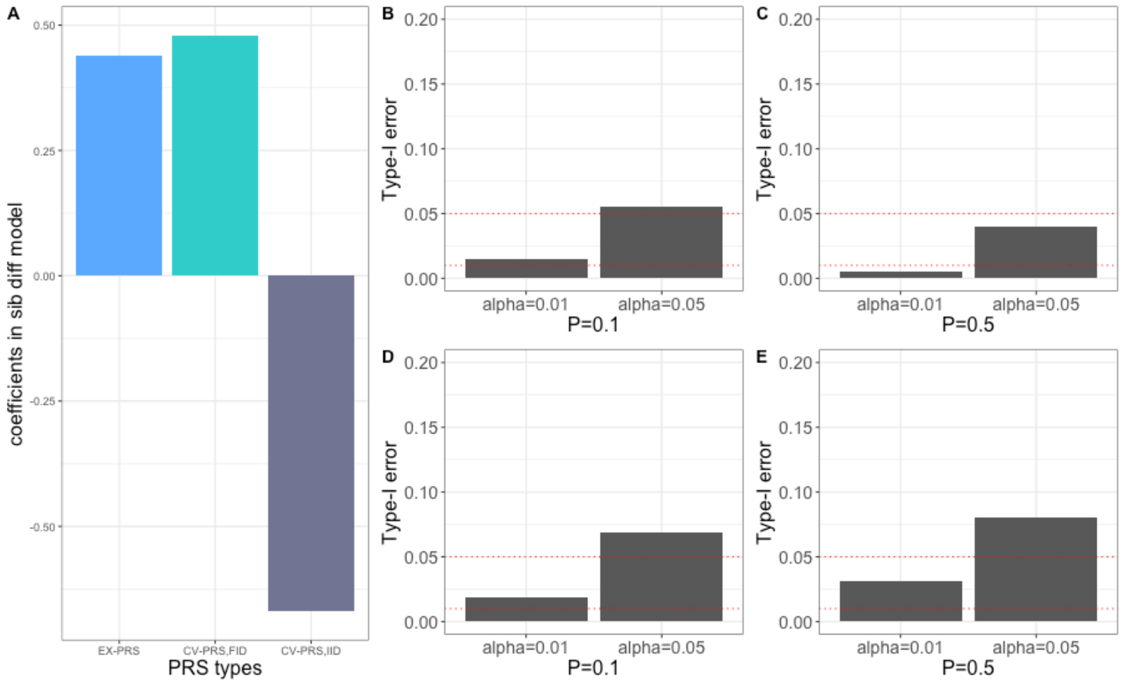

Figure 2.4: Type-I error rates for CV-PRS with different proportions of related samples. A-C: proportions of related samples are 10% and 50% respectively, and samples from the same family were partitioned into the same fold (family structure were identified by KING[25]). D-F: proportions of related samples are 10% and 50% respectively, and samples were randomly partitioned regardless of their family structures. The value of each bar represents the proportion of p-values that are less than or equal to 0.01 or 0.05. G: coefficients of PRS in sibling difference models. Coefficients from left to right are for sibling difference models using EX-PRS, CV-PRS with samples partitioned based on family structure, and CV-PRS with samples randomly partitioned regardless of family structure.

Adjusting for fold assignment in CV-PRS applications

Next, we explore the prediction performance of CV-PRS. Notice that CV-PRS is constructed by stacking the PRS from each fold. It is possible that the distributions of PRS are not identical across all folds, which may impact the predictive accuracy of CV-PRS. Still using EA as an example, the distributions of CV-PRS in different folds are normally distributed but with noticeably different means (Figure 2.5). To identify the source of this variation, we applied the same external GWAS to each fold to derive EX-PRS. We did not find differences in PRS distributions across folds (with averaged p-value of 0.46 in pairwise Kolmogorov-Smirnov tests). Thus, the variation in CV-PRS distribution is explained by the differences in GWAS produced in different folds.

A consequence of the between-fold CV-PRS variation is reduced predictive performance (quantified by R-squared values) in downstream regression. We suggest two approaches for improvement. The first one is to add the fold indicator as a covariate in regression and report partial R-squared that measures contribution of CV-PRS only. We note that this approach does not account for the differences in standard deviation of PRS across folds, but these differences are relatively minor compared to the variations in means. Another approach is to standardize PRS within each fold separately before stacking them together (Mefford et al., 2020). We used EA and height as examples to illustrate that these two methods can indeed improve predictive performance of CV-PRS. We produced CV-PRS of EA and height for 408,325 UKB participants, then regressed EA (and height) phenotype on its CV-PRS to obtain partial R-squared values. We observed that both strategies produced higher R-squared values than directly combining PRS from each fold (Figure 2.6A).

We have shown in the previous section that number of folds has no effect on type-I error rate. Now we investigate whether it influences the CV-PRS predictive performance using CV-PRS of height as an example. Figure 2.6 shows the R-squared values of linear association model between height and its CV-PRS with the number of folds ranging from 2 to 20. A visible increase in R-squared values was observed when the number of folds changed from 2 to 4. This is expected since the GWAS sample size substantially increased (from 50% to 75% of the total sample size). When the number of folds is greater than 4, the increase in R-squared values became relatively small. To balance computation burden and the predictive performance and also to avoid overfitting, we set the number of folds of 4 as the default choice in analyses presented in this paper.

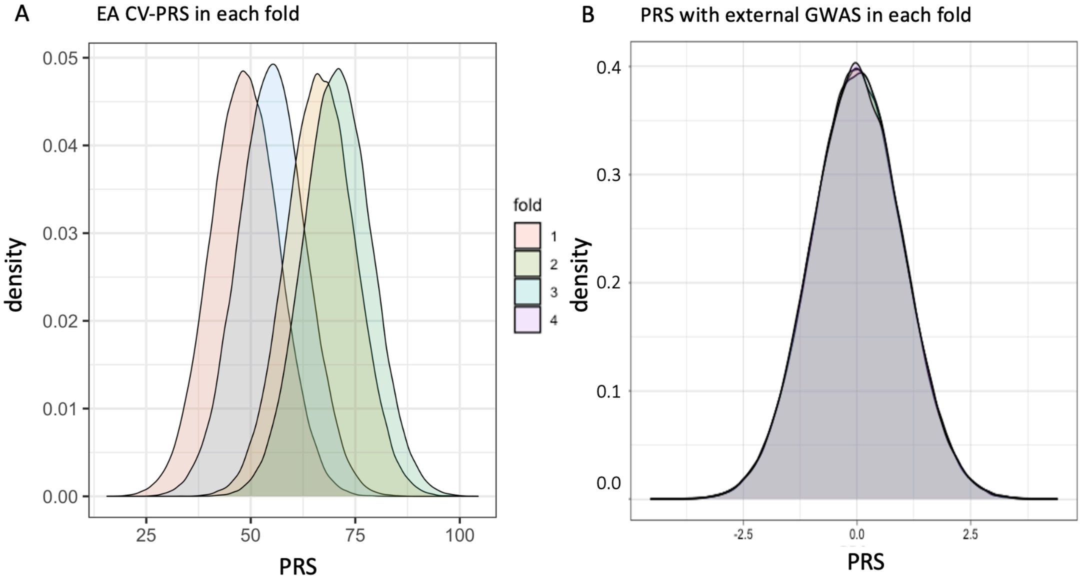

Figure 2.5: Exploratory analysis of EA CV-PRS. A: Distribution of CV-PRS for each of 4 folds. B: Distribution of PRS calculated by genotypes from each fold and an external GWAS.

Comparison of R-squared values for EX-PRS and CV-PRS

CV-PRS provides an alternative option when no external GWAS is available. We compared the predictive R-squared values of EX-PRS and CV-PRS on EA and height. For EX-PRS, we first performed GWAS in UKB, and then applied GWAS to a set of holdout samples (N=10k, 50k, and 100k respectively) to generate EX-PRS. The size of GWAS was chose to be 25%, 50%, 75% and 100% of the size of holdout sample. We also generated CV-PRS on a randomly drawn sample with N=10k, 50k, and 100k, respectively. Results are shown in Figure 2.6. The R-squared values of CV-PRS were comparable to R-squared values of EX-PRS when the external GWAS sample size is 50%-70% of the holdout sample size for EX-PRS. Thus, when external GWAS has a smaller size, CV-PRS may yield better predictive performance compared to EX-PRS.

When a well-powered external GWAS does exist, cross-validation could still be beneficial because we may combine the external GWAS and cross-validated GWAS for each fold through meta-analysis to further increase power. Based on the same settings described above, meta-analyzing external GWAS with cross-validated GWAS for each fold produced PRS that outperformed the CV-PRS without meta-analysis (Figure 2.6C-D) except for the case when the target size is 10k for EA due to limited power of UKB external GWAS. However, when the external GWAS is obtained from a different study, we generally expect a smaller gain in predictive power from meta-analysis due to effect heterogeneity across studies, although this may be compensated if the external GWAS is big. We investigated this using the same CV-PRS generated before and replacing the external GWAS with published GWAS not using UKB samples. More specifically, we used the height GWAS from the GIANT consortium (Wood et al., 2014) (N = 253,288) and the EA GWAS from Lee et al. (2018) with UKB sample excluded (N = 324,162) (Lee et al., 2018). Since these are fairly large external GWAS compared to the dataset we used for cross-validation, EX-PRS based on these external summary statistics always showed substantially higher predictive R2 compared to CV-PRS. But when the cross-validation sample size was big (e.g., 100k in our analysis), meta-analysis of external and cross-validated GWAS further improved PRS performance. However, when the sample size is small in the cross-validation cohort, the gain was negligible.

2.3 Discussion

The typical procedure to produce PRS requires the GWAS and target samples to be independent to avoid overfitting. This can be challenging for generating scores in large population cohorts since they are included in many large GWAS. Summary statistics independent from these cohorts are usually either unavailable or severely underpowered. In this paper, we investigated the feasibility of CV-PRS via a variety

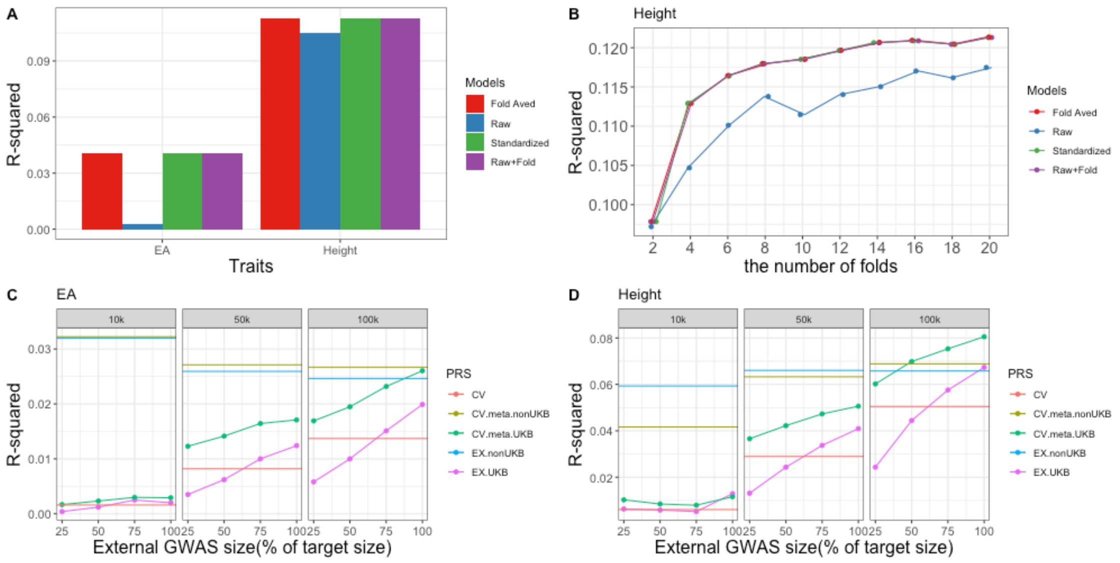

Figure 2.6: Performance of CV-PRS on EA and height. A: Bar plot for regression R^2 under various settings. Fold Aved: average regression R^2 across four folds of cross-validation. Raw: regressing phenotypes on CV-PRS in the combined sample of all folds. Standardized: standardizing CV-PRS in each fold separately, then regressing phenotypes on standardized CV-PRS in the combined sample. Raw+fold: regressing phenotypes on CV-PRS with the fold indicator added as a covariate. B: CV-PRS R^2 on height across a range of fold numbers. C, D: Predictive R^2 of CV-PRS, EX-PRS based on external GWAS performed in UKB, EX-PRS based on non-UKB GWAS, and CV-PRS based on meta-analysis of cross-validated and external GWAS. Panels C and D show results of EA and height, respectively.

of simulations and real data applications, shedding light on the influence of sample size, sample relatedness, number of folds, and data processing strategies on PRS performance. We found that when sample size is large (e.g., greater than 500k in UKB), individual-level CV-PRS correlations within and between folds become negligible and type-I error of CV-PRS appears under control. Additionally, number of folds in cross-validation does not have a big impact on the performance of CV-PRS. However, it is important to partition related study participants into the same fold to avoid overfitting in downstream analysis. We also recommend either adding the fold indicator as a covariate in PRS association analysis, or standardizing CV-PRS within each fold before stacking the scores across all samples. Finally, when an external GWAS does exist, performing meta-analysis of external and cross-validated GWAS will further improve PRS performance. These results provide important guidelines on producing and applying PRS in large cohorts through cross-validation.

Our study also has some limitations. First, we only studied a simple PRS model based on clumped GWAS summary statistics without any p-value thresholding. Whether cross-validation can be applied to more sophisticated PRS models (Ge et al., 2019) and parameter-tuning strategies (Zhao et al., 2021) remains to be studied in the future. Second, we only analyzed samples of European descent in this study and our results only provide guidance on a simple scenario where each ancestry group is analyzed separately. However, recent studies have convincingly demonstrated the lack of PRS portability across ancestral populations (Martin et al., 2019) and highlighted the importance of multi-ancestry GWAS joint modeling (Ruan et al., 2022; Miao et al., 2022). The best strategy to implement multi-population integrative analysis of CV-PRS remains elusive and needs to be studied in the future. Finally, compared to the standard PRS approach, producing CV-PRS can be time consuming. More work needs to be done to improve the computational efficiency of implementing cross-validation in large biobank datasets.

Taken together, we provided empirical evidence to support cross-validation as an alternative strategy for producing PRS in large biobanks. While results based on CV-PRS need to be interpreted with caution and will ultimately needed to be validated in external cohorts, these scores can sometimes be the only option in certain applications and are therefore crucial for hypothesis generation and exploratory analysis.

3 PARTITIONING HERITABILITY AND GENETIC COVARIANCE BY DIRECT AND INDIRECT EFFECT PATHS

3.1 Background

Genetic correlation is an effective metric for quantifying the shared genetic architecture of multiple complex traits (van Rheenen et al., 2019; Zhang et al., 2021a) and has quickly gained popularity in genome-wide association studies (GWAS) in the past few years. Genetic correlation is typically defined as the correlation between additive genetic components of two complex traits and can be estimated from genome-wide data of millions of single nucleotide polymorphisms (SNPs). Several methods have been developed to improve genetic correlation estimation using individual-level GWAS data (Yang et al., 2011; Loh et al., 2015), GWAS summary statistics (Bulik-Sullivan et al., 2015b; Lu et al., 2017), or both (Zhang et al., 2021b). Recent studies have also expanded this concept to quantify genetic correlations in local genomic regions (Guo et al., 2021b; Werme et al., 2021a; Shi et al., 2017; Zhang et al., 2021c), between human ancestral populations (Brown et al., 2016; Miao et al., 2022), and using other types of genetic variations (Guo et al., 2021a). Overall, these methods have become a routine component of complex trait genetic studies and provided crucial insights into the genetic basis of numerous human traits.

As genetic correlation analysis becomes standard in GWAS, challenges arise in interpreting the pervasive correlations observed across many phenotypes. In particular, interpretation of genetic correlations involving human behavioral phenotypes is challenging because most methods for estimating these correlations ignore the apparent gene-environment correlation underlying human behavior (Koellinger Philipp and Harden, 2018; Bates et al., 2018; Trejo and Domingue, 2018; Willoughby et al., 2021; de Zeeuw et al., 2020; Cheesman et al., 2020; Domingue and Fletcher, 2020; Young et al., 2019; Hwang et al., 2020). For example, we now know that parents' genomes can influence parental behaviors and family environment

[0] Co-authors: Yiqing Zou, Yuchang Wu, Jiacheng Miao, Ze Yu, Jason M. Fletcher, Qiongshi Lu

which, in turn, shape the phenotypes of their children. Additionally, because parent and offspring genotypes are correlated, typical GWAS can be severely confounded by parental genotypes, and effect estimates obtained in GWAS are mixtures of direct genetic effects (i.e., how one's own genotypes affect his/her phenotype) and indirect genetic effects (i.e., how parents' genotypes affect their children's phenotypes; this is also referred to as genetic nurture) (Kong et al., 2018; Wang et al., 2021). Being able to decompose shared genetics between two traits by direct and indirect effect paths is crucial for understanding the respective roles of genetics and environments. Although methods have been developed for separating SNP effects into direct/indirect paths, and follow-up genetic correlation analyses performed on direct/indirect association statistics can be a feasible strategy in some cases (Wu et al., 2021), these GWAS typically require genotype-phenotype data from a large number of families which can be difficult to obtain even in large biobanks.

In this paper, we introduce a statistical framework to estimate the variance and covariance components of direct/indirect genetic effects on trait pairs using data from a limited number of families. This framework enables us to partition both heritability and genetic covariance into direct and indirect pathways, and gain better knowledge of how genetics and environment together shape people's phenotypes as well as the correlation between a pair of traits. We employ the method of moments to obtain accurate estimates of variance and covariance parameters and use a Jackknife approach to obtain confidence intervals. We apply our framework to five complex traits in UK Biobank (UKB) (Bycroft et al., 2018) for conducting two-trait analysis for 10 trait pairs in total, and partition the trait-specific and shared genetic components by direct and indirect effect paths.

3.2 Results

Methods overview

Our approach is built on a pair of linear models for a pair of standardized traits Y_1 and Y_2 , standardized own- genotypes X, and standardized parental genotypes

$X_p + X_m$ as follows:

$$Y_1 = X\alpha_1 + (X_p + X_m)\beta_1 + \epsilon_1$$

$$Y_2 = X\alpha_2 + (X_p + X_m)\beta_2 + \epsilon_2$$

Error term ϵ_1 and ϵ_2 are assumed to behave mean 0 and variance-covariance matrix Σ. α_k and β_k ($k = 1, 2$) are random effect vectors denoting direct (of own- genotypes) and indirect effects (of parental genotypes) on two traits. They share the variance-covariance structure below with M being the number of SNPs and I_M being the identity matrix.

$$var \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \frac{1}{M} \begin{bmatrix} \sigma^2_{\alpha_1} I_M & \rho_\alpha I_M & \rho_{\alpha_1\beta_1} I_M & \rho_{\alpha_1\beta_2} I_M \\ \rho_\alpha I_M & \sigma^2_{\alpha_2} I_M & \rho_{\alpha_2\beta_1} I_M & \rho_{\alpha_2\beta_2} I_M \\ \rho_{\alpha_1\beta_1} I_M & \rho_{\alpha_2\beta_1} I_M & \sigma^2_{\beta_1} I_M & \rho_\beta I_M \\ \rho_{\alpha_1\beta_2} I_M & \rho_{\alpha_2\beta_2} I_M & \rho_\beta I_M & \sigma^2_{\beta_2} I_M \end{bmatrix}$$

Here, ρ_α and ρ_β are two key parameters we focus on, although other cross terms are also estimated and may be of interest in practice. These two parameters quantify the covariance of direct effects and covariance of indirect effects on two traits, respectively. We apply the method of moments to produce parameter estimates and employ a Jackknife approach to quantify the standard error of parameter estimates. More statistical details are described in the Methods section.

Simulation results

We first performed simulations to demonstrate that our method produces unbiased estimates with well-calibrated confidence intervals. We randomly selected 2000 families from UKB with data on full sibling pairs and imputed parental genotypes to perform simulations. The imputation is based on expectation of the average of parental genotypes using two or more sibling genotypes (Young et al., 2020). We explored multiple parameter settings. Each setting contained 200 repeats. Details on simulation settings are described in the Methods section and in Table B.1. Results

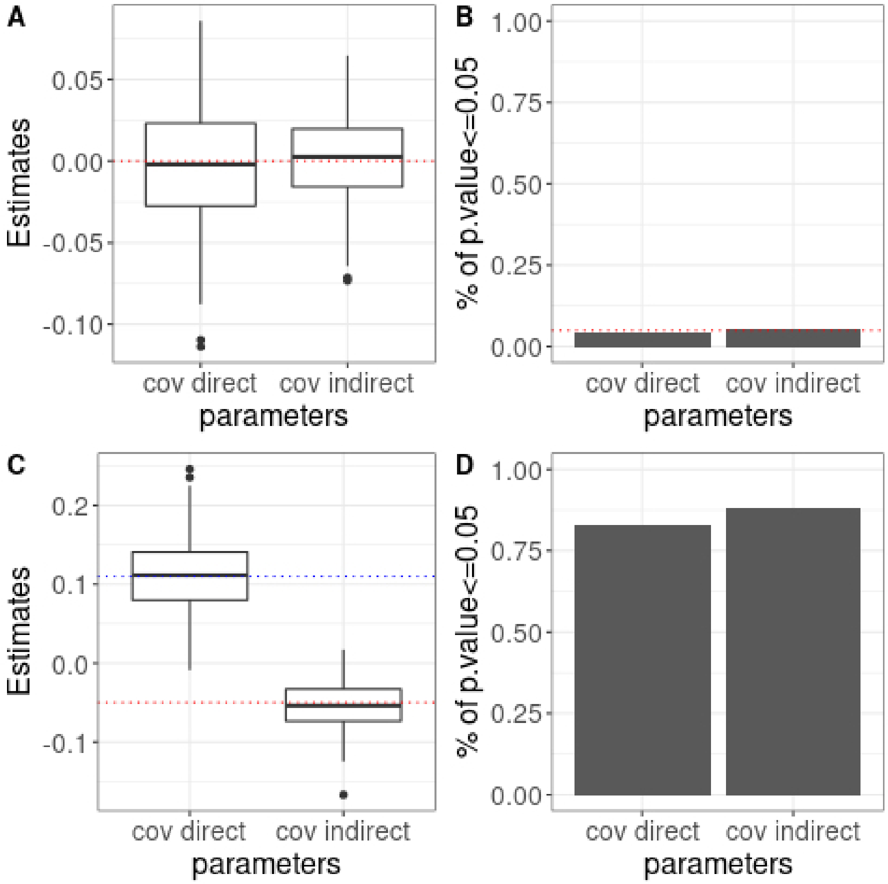

Figure 3.1: Simulation results on point estimates, type-I error, and statistical power for covariance of direct/indirect effects. Panels A,C: box plot for point estimates of ρ_α (i.e., covariance of direct effects) and ρ_β (i.e., covariance of indirect effects). Red dotted lines indicate the true values of ρ_α. Blue dotted line show the true value of ρ_β. Penals B,D: proportion of p-values $\leqslant 0.05$ across 200 repeats. In panel B, the dotted line highlights the type-I error threshold 0.05.

related to covariance of direct effects and covariance of indirect effects are shown in Figure 3.1. Results for the estimates of all 14 parameters are shown in Figure S1. We observed unbiased estimates and well-controlled confidence interval coverage rates for all parameters across various settings.

Partitioning heritability by direct/indirect effect paths for five complex traits

We applied our model to 12,571 families in UKB with full sibling data available and parental genotypes imputed (Methods). We analyzed five quantitative traits: height, body mass index (BMI), overall health, educational attainment (EA; quantified as years of education), and household income. For each trait pair, we estimated 14 parameters including trait-specific parameters (which partition heritability), trait pair parameters (which partition genetic covariance), and parameters for error terms. The detailed data processing procedure is described in the Methods section.

First, we examined parameter estimates for the variance of genetic direct effects, genetic variance of indirect effects, and genetic covariance between direct and indirect effects on the same trait. These parameters are the partitioned terms of trait heritability; therefore, we can reconstruct total heritability estimates by linearly combining these three terms. For comparison, we also calculated LDSC heritability estimates using GWAS performed on the same data. We observed a high correlation (correlation = 0.962, slope=0.749; Figure 3.2A).

Results of these three parameters are shown in Figure 3.3A and Table B.2. We found a substantial and statistically significant direct effect component on BMI ($p = 4.9 \times 10^{-5}$). The BMI indirect effect was closed to zero and not significant. Similarly, we found a significant direct genetic component ($p = 4.8 \times 10^{-3}$) and a weaker indirect component for height. Notably, the direct and indirect genetic components for height showed a significant positive covariance ($p = 7.7 \times 10^{-3}$). The direct and indirect genetic components for EA were similar in size but only the indirect component reached statistical significance. Consistent with previous studies (Wu et al., 2021), we did not find evidence for genetic correlation between the EA direct and indirect effects. We found null results for household income and overall health.

We also obtained the estimated covariance between siblings' error terms for each trait (Figure 3.3B and Table B.2). Error term covariances are significant for all 5 phenotypes, suggesting a substantial contribution of shared environment on

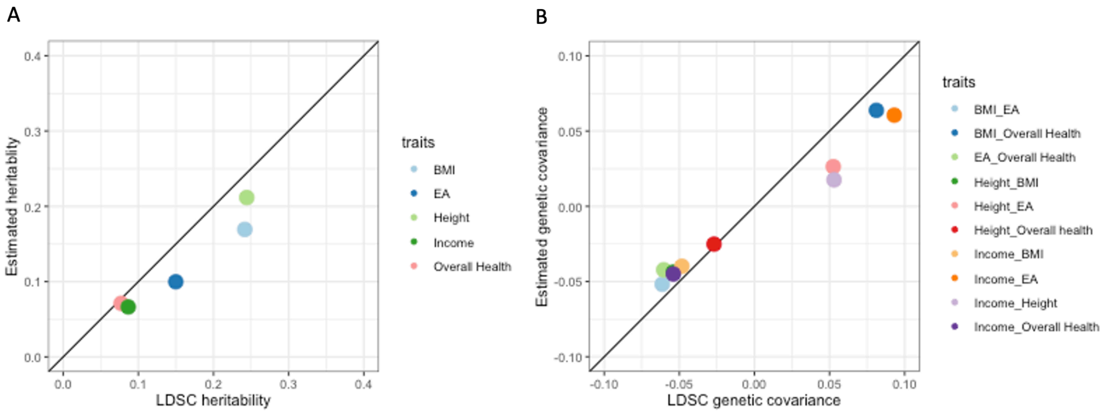

Figure 3.2: Total heritability and genetic covariance can be recovered from partitioned parameter estimates. A: estimates for total heritability. B: estimates for genetic covariance. X-axis: LDSC estimates. Y-axis: Estimates reconstructed from parameter estimates in our analysis.

siblings' phenotypes.. The findings for EA are consistent with Branigan et al. study (Branigan et al., 2013) that measures variance contribution of shared environment factors using ACE model based on twin studies. They obtained the estimates from multiple cohorts, and the results for two UK cohorts showed statistical evidence of the shared environment variance components with estimates were about 0.3, and our estimates for the shared environment variance is 0.26 (as a linear combination of variance of indirect effects, covariance of direct and indirect effects, and covariance of errors for two siblings).

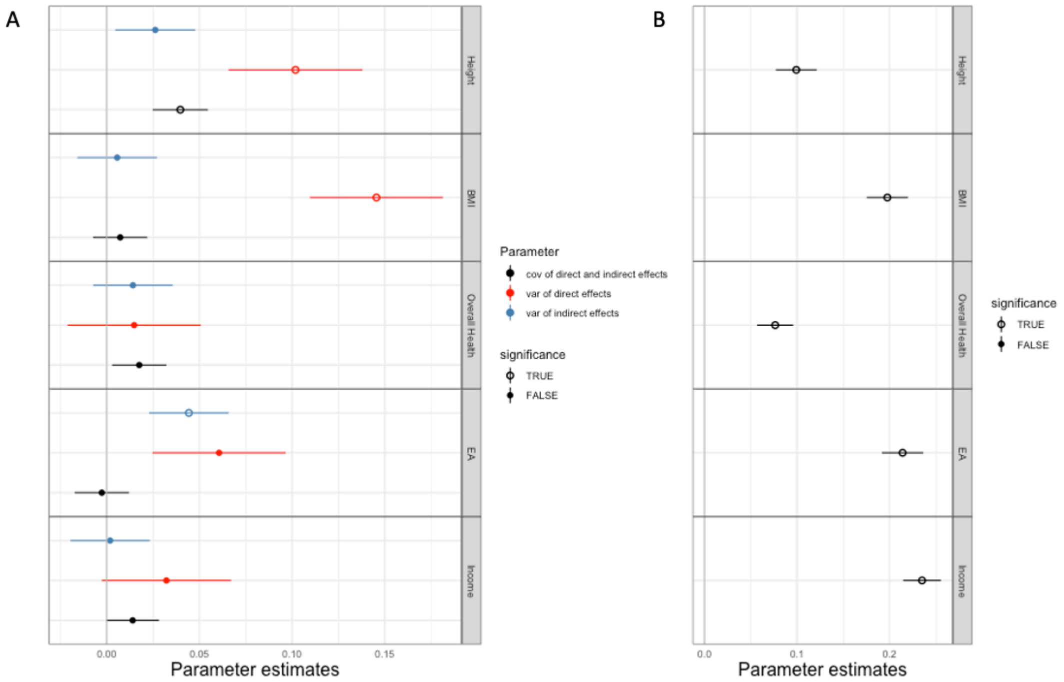

Figure 3.3: Estimation results for 5 complex traits in UKB. Each interval shows the point estimate and standard error for one parameter. Circles denote significant parameter estimates at the 0.05 level. Other estimates are denoted by solid circles. A: variance of direct effects, variance of indirect effects, and covariance of direct and indirect effects for height, BMI, overall health, EA and income. B: covariance of error terms for sibling pairs for each trait.

Partitioning genetic covariance for 10 trait pairs

Our approach also partitions the total genetic covariance between two traits into direct and indirect effect paths. Similar to the analysis for total heritability, we reconstructed the total genetic covariance estimates based on partitioned parameters (Methods) and compared them to LDSC estimates. We found a high correlation between these estimates (Figure 3.2B, correlation = 0.977, slope = 0.703).

Estimation results for ρ_α (i.e., genetic covariance of direct effects on two traits) and ρ_β (i.e., genetic covariance of indirect effects on two traits) are shown in Figure 3.4A-B (also see Table B.3). We found a significant direct effect covariance (ρ_α= 0.044) between BMI and worse overall health (a higher value in this trait indicates worse health; see Methods). This contributes to 69% of the total genetic covariance between BMI and overall health. Several other trait pairs showed significant covariances of indirect genetic components, including height and BMI (ρ_β= - 0.02, 48.7% of genetic covariance), height and income (ρ_β= 0.032, 182% of genetic covariance), and EA and overall health (ρ_β= - 0.018, 40.8% of genetic covariance). EA and income have substantial and statistically significant covariances in both direct and indirect genetic components (ρ_α= 0.044, 72.9% of genetic covariance, ρ_β= 0.032, 53.5% of genetic covariance).

We also estimated other cross-term covariances, including the covariance of direct effect on one trait and indirect effect on another trait (Figure 3.4C-D). We identified a number of significant correlations. For example, we found highly significant correlations between the indirect effect of BMI and the direct effect of both EA ($\rho_{\alpha_2\beta_1}$ = - 0.037, 66.2% of genetic covariance) and income ($\rho_{\alpha_2\beta_1}$= - 0.028, 70.5% of genetic covariance). The direct effect of income is also negatively correlated with the indirect effects of overall health ($\rho_{\alpha_2\beta_1}$= - 0.02, 45.5% of genetic covariance). The direct effect of height is correlated with the indirect effect of better overall health ($\rho_{\alpha_1\beta_2}$= - 0.02, 74.7% of genetic covariance), higher EA ($\rho_{\alpha_1\beta_2}$= 0.02, 80.0% of genetic covariance), and lower income ($\rho_{\alpha_1\beta_2}$= - 0.029, 165% of genetic covariance).

3.3 Discussion

In this paper, we introduced a new statistical framework to decompose heritability and genetic covariance of complex traits by direct and indirect effect paths. Through simulations and application to five traits using full siblings and their imputed parental genotypes in UKB, we demonstrated that our model produces unbiased estimates with well-controlled confidence interval coverage.

Partitioned variance and covariance components can be combined to recover traditional heritability and genetic covariance estimates. But stratified heritability

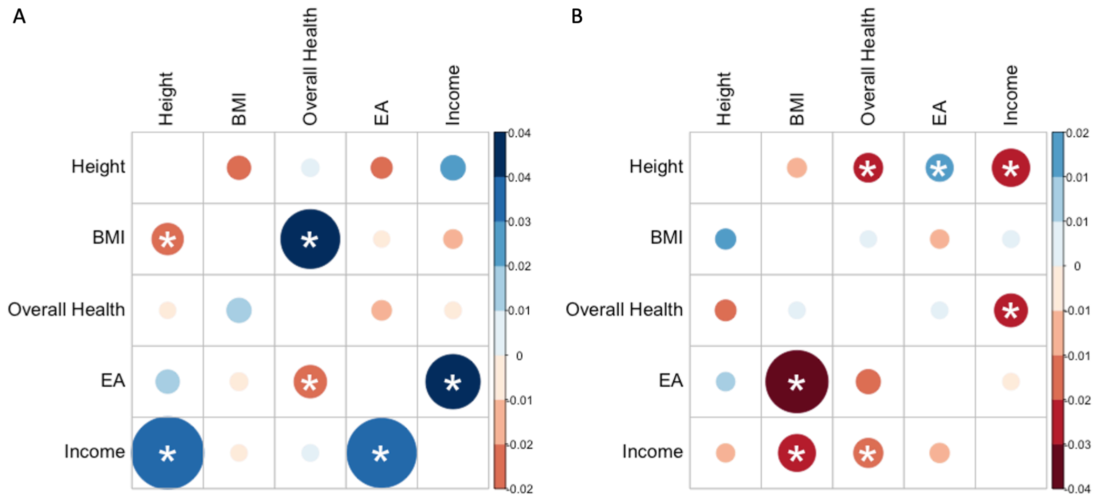

Figure 3.4: Estimation results for covariance of direct/indirect effects on five complex traits. A: upper triangle: heatmap of direct effect genetic covariance (i.e., (ρ_α); lower triangle: heatmap of indirect effect genetic covariance (i.e., ρ_β). B: covariance between direct effects of trait 1 (rows) and indirect effects of trait 2 (columns). Size of the circles visualizes false discovery rate (FDR). Smaller size refers to larger FDR. Significant correlations with FDR<0.05 are marked by asterisks.

and genetic covariance estimates provided crucial new insights into the genetic basis of complex traits. First, our analysis confirmed that traits like EA have a substantial indirect genetic component. That is, a substantial proportion of trait heritability is mediated through parents and the family environment they create. For some other traits (e.g., BMI), the contribution of indirect genetic effect is negligible, and heritability is primarily explained by the direct genetic component. More importantly, our approach allowed decomposing genetic covariance, which is highly novel compared to other approaches in the literature. Similar to findings in single-trait analysis, our results highlight the importance of considering indirect genetic components when estimating and interpreting the shared genetic basis between complex traits. Several trait pairs in our analysis (e.g., height and income) only showed significant genetic covariance in their indirect genetic components. Some other traits (e.g., BMI and overall health) have substantial genetic sharing of the direct effect component. Then, for EA and income, two important socioeconomic GWAS phenotypes, genetic covariance of both direct and indirect genetic components were substantial and highly statistically significant. Omitting either one of them in genetic correlation analysis will likely lead to biased interpretation. For example, if we ignore the covariance of indirect effects, we may conclude that EA and income are correlated mainly due to the shared genetic basis. However, the indirect paths play an important role in determining the relatedness of EA and income. Without noticing this, people may pay little attention in creating a better family environment to their children's education if they wish their children to gain more income in the future. With our approach, it is now possible to carefully decompose heritability or shared genetic components into direct and family-mediated paths, which will be beneficial for post-GWAS analysis in general and analysis of human sociobehavioral phenotypes in the future.

Our study also has several limitations. First, parental genotypes used in our analysis were not directly measured but imputed based on siblings' genotypes. The imputation accuracy depends on several factors including the number of siblings in each family and assortative mating. SNIPar adjusted for bias due to assortative mating, and more siblings for a single family could leads to more reliable imputation of their parental genotypes. These may affect the accuracy of parameter estimates. However, we note that our approach can be applied to measured genotypes and phenotypes in parent-offspring trios if such data are available, but our accessible trios cohorts have smaller sample sizes compared to UKB full siblings data. Second, although it is also possible to calculate genetic correlations by combining stratified genetic covariance and heritability estimates for direct/indirect effects, these estimates for genetic correlations are numerically unstable at the sample size in our analysis, which is why we based our inference of shared genetics on genetic covariance instead. Furthermore, our analyses were limited to quantitative traits. For applications involving binary outcomes, future work is needed to expand the statistical model.

Taken together, our method is a first step to explore the covariance structure of direct and indirect effects for complex human traits. It marks an important methodological innovation and may have broad and impactful applications in future GWAS analysis for traits that are highly affected by parental behaviors and the environment, such as birthweight, the number of children, and some neurological diseases.

4 METHOD DETAILS FOR CHAPTER 2 AND CHAPTER 3

4.1 Methods for chapter 2

UKB data processing

We used UKB samples with European ancestry, identified from principal component analysis (data field 22006), to calculate CV-PRS and perform regression analyses. We used KING (Manichaikul et al., 2010) to infer the pairwise family kinship and created family identifiers in UKB. Standing height is obtained from data field 50 in UKB, measured in centimeters and NA values are omitted. For EA, we used the âŁœqualificationâŁž (data field 6138) to compute the years of schooling as the EA phenotype following Lee et al. (Lee et al., 2018).

GWAS analysis and PRS calculation

We constructed CV-PRS using 1000 Genomes Project Phase III European samples as the reference (Auton et al., 2015) for linkage disequilibrium (LD). All GWAS summary statistics were clumped using PLINK (Purcell et al., 2007) to remove variants with high LD. We used an LD window size of 1Mb and a pairwise r2 threshold of 0.1. In addition, strand-ambiguous SNPs and SNPs that do not exist in the UKB imputed genotype data were removed. No additional filtering based on GWAS p-values was applied. Genetic principal components (PCs) were computed using flashPCA2 (Abraham et al., 2017). All GWAS were performed using Hail, with year of birth, sex, genotyping array, and 20 PCs added as covariates. PRS calculation was implemented in PRSice-2 (Choi and O'Reilly, 2019).

Simulation settings

For the simulations involving individual-level PRS correlation, we simulated heritable phenotypes with UKB genotype data. We randomly specified 4k causal SNPs. Effect size of each causal SNP was sampled from a normal distribution with mean zero and total heritability of 0.4. The simulations were performed based on sample sizes of 1k, 10k, 50k, 100k, and 200k, each of which involves 200 repeats. In each repeat, we multiplied genotype data by SNP effect sizes, and then added a random error term with a mean of 0 and a variance of 0.6 to obtain the phenotype values. These genotype data and simulated phenotype data were used to generate CV-PRS. For EX-PRS, we performed GWAS on the same data used to produce CV-PRS, and produced PRS on another set of holdout samples with a size of 2k in each repeat. By repeating the whole process, we obtained 200 sets of CV-PRS and EX-PRS which we used to calculate individual-level PRS correlations. The number of folds for CV-PRS was set to be 4 in these analyses.

We used EA as an example phenotype in the type-I error simulations. First, we simulated phenotypes that were genetically independent from EA. 1% of SNPs (960k) were randomly selected to be causal with effect sizes randomly sampled from a normal distribution and a total heritability of 0.4. Then, we multiplied genotype data with simulated effect sizes and added a random error term simulated from a normal distribution of mean 0 and variance 0.6 to produce phenotypes. These simulated phenotypes were regressed on CV-PRS to obtain association p-values. We repeated this procedure 200 times to calculate type-I error rate.

For the analysis of number of folds, we chose the cross-validation fold numbers to be 4/6/8/10/12/14/16/18/20. We used the same phenotypes from previous simulations to calculate type-I error rates.

For the simulations about sample relatedness, we set the total sample size to be 50k and chose the size of dependent samples to be 10%, 30%, and 50% of the total sample size. For example, in the setting where 10% of the total sample are related, we randomly selected 2.5k full sibling pairs (5k samples), and 45k independent individuals from UKB. The whole dataset was then split into 4 folds 1) by family identifiers (identified by KING) and 2) completely randomly without considering sample relatedness. Phenotypes were simulated in the same approach as previous simulations. Finally, we calculated CV-PRS based on genotype data and simulated phenotype data, and performed association analysis using CV-PRS and real traits (i.e., height and EA).

Real data application in UKB

We used all UKB individuals to investigate CV-PRS performance under various settings and number of folds. For comparison of CV-PRS, EX-PRS, and PRS produced from meta-analysis of external and cross-validated GWAS, we set the sample sizes to be 10k, 50k, and 100k. For each sample size setting, we calculated 1) CV-PRS; 2) EX-PRS based on an external GWAS whose size equals to 25%, 50%, 75% and 100% of the total sample size; 3) CV-PRS that involves meta-analysis with EX-PRS in 2); 4) CV-PRS that involves meta-analysis with non-UKB external GWAS. The non-UKB external GWAS for height and EA were obtained from the GIANT consortium (Wood et al., 2014) and Lee et al. (2018) (Lee et al., 2018) (with UKB samples excluded).

4.2 Methods for chapter 3

Model for decomposing heritability and genetic covariance (degC)

We assume a pair of linear genetic models as follows:

$$Y_1 = X\alpha_1 + (X_p + X_m)\beta_1 + \epsilon_1$$

$$Y_2 = X\alpha_2 + (X_p + X_m)\beta_2 + \epsilon_2$$

where Y_1 and Y_2 denote two complex traits, X denotes own- genotypes, and X_p and X_m denote paternal and maternal genotypes. Since we used sibling pairs data as input (that will be described later in this section), we assume the total number of sibling pairs is N.Then Y_1 and Y_2 are vectors with length 2N, and X,X_p and X_m are 2N-by-M matrices, where M is the total number of SNPs. We assume both genotypes and phenotypes (Y_1, Y_2, X, X_p, and X_m) to be standardized. α_1 and α_2 are random effect terms that quantify the direct genetic effects on Y_1 and Y_2, i.e., how someone's own genotypes affect their own phenotype. Following conventions in the linear mixed model literature on modeling heritability (Yang et al.), we also assume these effects to follow normal distributions.

$$\alpha_k \sim N(0, \frac{\sigma^2_{\alpha_k}}{M}), k = 1, 2$$

Here, $\sigma^2_{\alpha_1}$ and $\sigma^2_{\alpha_2}$ are the direct genetic variance components for two traits respectively. Similarly, β_1 and β_2 are random effect terms for quantifying indirect genetic effects on two traits, i.e., how the parents' genotypes affect someone's phenotype. These random effects also follow normal distributions.

$$\beta_k \sim N(0, \frac{\sigma^2_{\beta_k}}{M}), k = 1, 2$$

Importantly, direct and indirect effects on the same trait (i.e., α_k and β_k) can be correlated. Direct effects on two different traits (i.e., α_1 and α_2), or indirect effects on two traits (i.e., β_1 and β_2), can both be correlated. Fairly generally, we assume α_1, α_2, β_1, and β_2 to have a joint distribution:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} \sim (0, \frac{1}{M} \begin{bmatrix} \sigma^2_{\alpha_1} I_M & \rho_\alpha I_M & \rho_{\alpha_1\beta_1} I_M & \rho_{\alpha_1\beta_2} I_M \\ \rho_\alpha I_M & \sigma^2_{\alpha_2} I_M & \rho_{\alpha_2\beta_1} I_M & \rho_{\alpha_2\beta_2} I_M \\ \rho_{\alpha_1\beta_1} I_M & \rho_{\alpha_2\beta_1} I_M & \sigma^2_{\beta_1} I_M & \rho_\beta I_M \\ \rho_{\alpha_1\beta_2} I_M & \rho_{\alpha_2\beta_2} I_M & \rho_\beta I_M & \sigma^2_{\beta_2} I_M \end{bmatrix})$$

Here, ρ_α quantifies the covariance of direct effects on two traits and ρ_β is the covariance of indirect effects on two traits. Similarly, $\rho_{\alpha_1\beta_1}$, $\rho_{\alpha_1\beta_2}$, $\rho_{\alpha_2\beta_1}$, and $\rho_{\alpha_2\beta_2}$ are pairwise covariance parameters between α and β. We note that this model is similar to what is used in the genetic nurture literature(Kong et al., 2018; Wu et al., 2021) in that it quantifies how parental genotypes shapes children's phenotypes. But a difference is that it is a polygenic model that uses random variables α_k and β_k to characterize genome-wide effects. This is motivated by the linear mixed model literature on heritability and genetic covariance estimation.

Based on the model formulation, naturally one would assume that the data required to fit this model would be genotypes of parents-offspring trios (i.e., X, X_p, and X_m) and offspring phenotypes on two traits (i.e., Y_1 and Y_2). While this is true, the number of trios available even in large population biobanks such as UKB is limited. Therefore, we made two important adjustments. One is that in our analyses, we leveraged the relatively large number of full sibling pairs available in UKB and recent statistical genetic advances in parental genotype imputation(Young et al., 2020). We begin with N sibling pairs (thus the total sample size is 2N), then impute the sum of parental genotypes $X_p + X_m$. These become the input data in our analysis. This also leads to the second adjustment we made which has also been shown above - we do not distinguish paternal and maternal indirect effects and instead focus on their average. This is similar to what was done in the indirect effect GWAS literature when sample size was small(Wu et al., 2021). In practice, in cases where a large number of trios are available, paternal and maternal indirect effects may be denoted as separate random variables in this framework.

One implication of having sibling pairs instead of independent samples or trios in the analysis is that we also need to consider the shared environmental effects between siblings. Error terms on two traits (i.e., ϵ_1 and ϵ_2) are both normally distributed random variables. But we allow 1) correlation of errors on two traits for the same individual, which is a common assumption in the genetic correlation estimation literature (Lu et al., 2017; Bulik-Sullivan et al., 2015a), and importantly, we also allow 2) correlation of error terms between siblings due to their shared environments. More specifically, we assume:

$$\epsilon_k \sim N(0, I_N \otimes \Sigma_k), \Sigma_1 = \begin{bmatrix} \sigma^2_{\epsilon_k} & \rho_{\epsilon_k} \\ \rho_{\epsilon_k} & \sigma^2_{\epsilon_k} \end{bmatrix}, k = 1, 2$$

We denote ϵ_{1i},ϵ_{2i} as such family i sibling pair error terms for trait 1 and 2 respectively, assume

$$\begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix} \sim N(0, \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{bmatrix} S), where \Sigma_{12} = \begin{bmatrix} \eta & \delta \\ \delta & \eta \end{bmatrix}$$

Parameter estimation

For simplicity, we denote the whole set of 14 parameters as

$$\Theta = \{\sigma^2_{\alpha_1}, \sigma^2_{\beta_1}, \rho_{\epsilon_1}, \rho_{\alpha}, \rho_{\beta}, \eta, \delta, \sigma^2_{\alpha_2}, \sigma^2_{\beta_2}, \rho_{\epsilon_2}, \rho_{\alpha_1\beta_1}, \rho_{\alpha_1\beta_1}, \rho_{\alpha_2\beta_2}, \rho_{\alpha_2\beta_1}\}$$

, and $\{\Theta\}_k$ stands for the kth parameter. Rather than utilizing an iterative restricted maximum likelihood algorithm to solve the linear mixed model, we employ the method of moments to improve computational efficiency. This approach produces parameter estimates by minimizing the distance between the model-derived and empirical variance-covariance matrix of phenotypes, i.e. cov(Y), where $Y = [Y_1^T, Y_2^T]^T$. Since we assumed an additive penetrance model, we can easily show that the model-derived covariance also follows an additive form:

$$var(Y) = I + \Sigma_k V_k \{\Theta\}_k$$

where V_k denote the sample relatedness matrix corresponding to kth parameter. For example, $V_1 = \frac{1}{M}\begin{bmatrix} XX^T & 0 \\ 0 & 0 \end{bmatrix} - I_{4N}$ and $V_2 = \frac{1}{M}\begin{bmatrix} (X_m + X_p)(X_m + X_p)^T & 0 \\ 0 & 0 \end{bmatrix} - I_{2N}$. We list all the relatedness matrices in Table B.4. The empirical variance of phenotypes is the estimates of cov(Y) based on real data, i.e., $\hat{cov}(y) = yy^T$. We estimate all parameters by minimizing the following function

$$L(\Theta) = \| yy^T - I - \Sigma_k \{\Theta\}_k V_k \|_F^2$$

where $\| \dot{} \|_F$ is the Frobenius norm. To minimize this function, we can use gradient descent method. That is, we let $\frac{\partial L}{\partial \{\Theta\}_k} = 0$ for all k and obtain a linear system of the form $A\Theta = B$,where A is an matrix and B is a vector. Then we could obtain parameter estimates by solving this linear equation.

While the variance of each parameter estimate can be derived under this statistical framework, they can be numerically unstable (and even turn negative) when sample size is limited. Therefore, we applied a resampling-based block Jackknife approach (Efron, 1982) to quantify the variance of parameter estimates. Block

Jackknife generates variance estimates by dividing the whole dataset into B blocks, holding out one block and producing parameter estimates using the other B-1 blocks in each time, then repeating this procedure for B times. Finally, we estimate variances using following formula:

$$v\hat{a}r(\{\Theta\}_k) = \frac{B-1}{B}\Sigma_{b=1}^{B}(\{\hat{\Theta}\}_{k,b} - \{\bar{\Theta}\}_k)^2$$

where $\{\bar{\Theta}\}_{k,b}$ denotes the estimate for kth parameter in bth repeat, and $\{\bar{\Theta}\}_k$ denotes denotes the averaged value across B estimates for parameter Θ_k. To account for multiple testing in real data applications, we calculated false discovery rate (FDR) based on all estimates for all pairs of traits.

Our model can also incorporate fixed effect covariates(Ahn et al.). For simplicity, we ignore the subscripts in the proposed model and add fix effects as follows:

$$Y = Z\gamma + X\alpha + (X_p + X_m)\beta + \epsilon$$

where Z stands for the covariate data matrix and Î³ denotes their fixed effects. To account for fixed effects, we can multiply a matrix Q that satisfies $QZ = 0$ to both sides of the equation. Matrix Q could be obtained by deriving the orthogonal vector spaces of the singular value decomposition of matrix Z. After multiplying matrix Q on both sides, the equation becomes the following without a fixed effect term.

$$QY = QX\alpha + Q(X_p + X_m)\beta + Q\epsilon$$

Then, the estimation procedure for all 14 parameters is the same as before.

Reconstructing heritability and genetic covariance using decomposed variance components

Following the model described above, total heritability h^2 and genetic covariance ρ_{gc} can be recovered from the decomposed variance component parameters as follows:

$$h_i^2 = \sigma^2_{\text{Î±}_i} + \sigma^2_{\text{Î²}_i} + 2\rho_{\alpha_i\beta_i}, i = 1, 2$$

$$\rho_{gc} = \rho_\alpha + \rho_\beta + \rho_{\alpha_1\beta_2} + \rho_{\alpha_2\beta_1}$$

By plugging in empirical estimates on the right-hand side of these equations, we could obtain estimates for total heritability and genetic covariance (Figure 3.2).

Data processing

UKB samples with European ancestry were identified from principal component analysis (data field 22006). We used KING (Manichaikul et al., 2010) to infer the pairwise family kinship and created family identifiers in UKB. Sum of parental genotypes were imputed by SNIPar (Young et al., 2020). Only autosomal SNPs with minor allele frequencies (MAF) greater than or equal to 0.05 and missing rate less than 0.01 were used for the analysis. Genotypes were standardized to have mean 0 and variance 1 using estimated minor allele frequencies, and missing values were imputed as 0. The sum of parental genotypes was standardized to have a variance of 2.

For phenotypes, participants were selected as overlapping samples of five phenotypes: height, BMI, EA, income, and overall health. We obtained height and BMI phenotypes from UKB data fields 12144 and 21001. Following previous work (Lee et al., 2018), we used data field 6138 to compute the EA phenotype as years of schooling. Household income data were obtained from UKB data field 738. The answers were coded as follows: (1) - Less than 18,000 (Pounds), (2) - 18,000 to 30,999, (3) - 31,000 to 51,999, (4) - 52,000 to 100,000, (5) - Greater than 100,000. Overall health was defined based on data field 100508. The answers were coded as follows: (1) - Excellent, (2) - Good, (3) - Fair, (4) - Poor. Note that a higher value indicates poorer health. We removed individuals with missing phenotype values and standardized all 5 traits to have mean 0 and variance 1. The final dataset we used for the analysis includes 4,748,473 SNPs and 12,571 sibling pairs (25,142 individuals and their imputed parents).

Simulation settings

We randomly selected 2000 full sibling pairs in UKB to perform simulations. A total of four settings were explored. The parameter values for each setting are listed in Table B.1. All settings were repeated 200 times. In each repeat, we simulated phenotypes by generating indirect effects, direct effects, and error terms based on the proposed penetrance model using the true parameters. Then, we produced parameter estimates using our statistical framework, and obtained variance estimates using block Jackknife.

Code availability

Implemented software of our approach is freely available at (https://github.com/qlu-lab/GV-partition).

A APPENDIX OF "GENERATING POLYGENIC RISK SCORES THROUGH CROSS-VALIDATION"

A.1 Supplementary Notes

CV-PRS in sibling difference model

We observed that when we randomly partition individuals into folds without considering sample relatedness, the association coefficient of CV-PRS in a sibling difference model showed flipped effect direction. Here we investigate the statistical reason behind this observation. Assume we have following penetrance model for phenotype Y and genotype X,

$$Y = X\beta + \epsilon$$

Without loss of generality, we assume genotypes X are centered at 0. β is genetic effect term, and ϵ is the error term and assumed to have mean of 0. A sibling difference model regresses the phenotypic difference in of siblings on the PRS difference of siblings, i.e.,

$$Y_{sib1} - Y_{sib2} \ \Sigma_i X_{sib1}\hat{\beta}_1 - \Sigma_i X_{sib2}\hat{\beta}_2$$

Notice that in sibling difference model, we only consider sibling pairs. When we randomly partition all samples into K folds without considering their family structure, it is possible that two siblings from the same family are partitioned into different folds. We will focus on this subset of sibling pairs (that two siblings lie in two folds) and see how they affect the estimated effect for sibling difference model. For each sibling pair, we randomly assign one sibling as sib1, and the other as sib2. Then we denote the phenotypes for all sib1's as Y^*_{sib1}, and phenotypes for all sib2's as Y^*_{sib2}. Similarly, we denote their genotypes as X^*_{sib1} and X^*_{sib2} respectively.

For each SNP, the effect size for one fold is estimated using data from the other K-1 folds. Since sib1 and sib2 lie in different fold, sib1 must be used for deriving effect sizes for sib2, and sib2 must be used for deriving effect sizes for sib1. Notice that except for sib2, there are other samples that are also used for calculation effect sizes for sib1 (i.e., those sibling pairs that are not be partitioned into different folds and lie in folds that do not contain sib1). Denote the genotypes for these samples as X_2^*, and denote their phenotypes as Y_2^*. Similarly, for samples that are not sib1 and will be used to calculate effect sizes of sib2, we denote their genotypes to be X_1^*, and their corresponding phenotypes as Y_1^*. Then the estimated effect size of SNP i for X_{sibi}^* is

$$\hat{\beta}_{1,i} = \frac{\begin{bmatrix} X_{sib2,i}^* \\ X_{2,i}^* \end{bmatrix}^T \begin{bmatrix} Y_{sib2}^* \\ Y_2^* \end{bmatrix}}{\begin{bmatrix} X_{sib2,i}^* \\ X_{2,i}^* \end{bmatrix}^T \begin{bmatrix} X_{sib2,i}^* \\ X_{2,i}^* \end{bmatrix}} = \frac{(X_{sib2,i}^*)^T Y_{sib2}^* + (X_{2,i}^*)^T Y_2^*}{(X_{sib2,i}^*)^T X_{sib2}^* + (X_{2,i}^*)^T X_{2,i}^*}$$

Then, the PRS for X_{sib1}^* are

$$\begin{aligned} PRS_{sib1}^* &= \Sigma_i X_{sib1}^* \hat{\beta}_{1,i} \\ &= \Sigma_i X_{sib1}^* \frac{(X_{sib2,i}^*)^T Y_{sib2}^* + (X_{2,i}^*)^T Y_2^*}{(X_{sib2,i}^*)^T X_{sib2}^* + (X_{2,i}^*)^T X_{2,i}^*} \\ &= \frac{X_{sib1}^* (X_{sib2,i}^*)^T Y_{sib2}^*}{(X_{sib2,i}^*)^T X_{sib2}^* + (X_{2,i}^*)^T X_{2,i}^*} + \frac{X_{sib1}^* (X_{2,i}^*)^T Y_2^*}{(X_{sib2,i}^*)^T X_{sib2}^* + (X_{2,i}^*)^T X_{2,i}^*} \end{aligned}$$

The second term does not contain information of Y_{sib2}^*, so let's denote it as a number $a_{1,i}$. For the first term, the denominator is a positive number, and the numerator is approximately $0.5X_{sib2}^*\sqrt{2p_i(1-p_i)}$, where p_i denote the allele frequency of SNP i. Similarly, we can obtain expression for the PRS of X_{sib2}^*. Thus, the difference in PRS for sib1's and sib2's is

$$\begin{aligned}
&PRS^*_{sib1} - PRS^*_{sib2}\\
\approx&\Sigma_i(\frac{0.5Y^*_{sib2}\sqrt{2p_i(1-p_i)}}{(X^*_{sib2,i})^T X^*_{sib2,i} + (X^*_{2,i})^T X^*_{2,i}} + a_{1,i}) - \Sigma_i(\frac{0.5Y^*_{sib1}\sqrt{2p_i(1-p_i)}}{(X^*_{sib1,i})^T X^*_{sib1,i} + (X^*_{1,i})^T X^*_{1,i}} + a_{2,i})\\
=&(\Sigma_i(\frac{0.5\sqrt{2p_i(1-p_i)}}{(X^*_{sib2,i})^T X^*_{sib2,i} + (X^*_{2,i})^T X^*_{2,i}} + a_{1,i}))Y^*_{sib2} - \Sigma_i(\frac{0.5\sqrt{2p_i(1-p_i)}}{(X^*_{sib1,i})^T X^*_{sib1,i} + (X^*_{1,i})^T X^*_{1,i}} + a_{2,i})Y^*_{sib1}\\
&+ \Sigma_i(a_{1,i} - a_{2,i})\\
\approx&(\Sigma_i(\frac{0.5\sqrt{2p_i(1-p_i)}}{\frac{k-1}{k}n\sqrt{2p_i(1-p_i)}})(Y^*_{sib2} - Y^*_{sib1}) + c\\
\approx&\frac{0.5km}{n(k-1)}(Y^*_{sib2} - Y^*_{sib1}) + c
\end{aligned}$$

The last second approximation is based on random sampling. m denotes the number of SNPs that are used to calculate PRS, n denotes the total sample size, and $c = \Sigma_i(a_{1,i} - a_{2,i})$ is the term that does not contain Y^*_{sib1} and Y^*_{sib2}.

Based on above, for sibling pairs that two siblings were partitioned to two folds, their phenotypic difference will be negatively related to the difference in their CV-PRS. Such sibling pairs occupied 75% of the total. Thus, the effect of sibling difference model was dominated by the negative association due to these sibling pairs.

A.2 Supplementary Figures

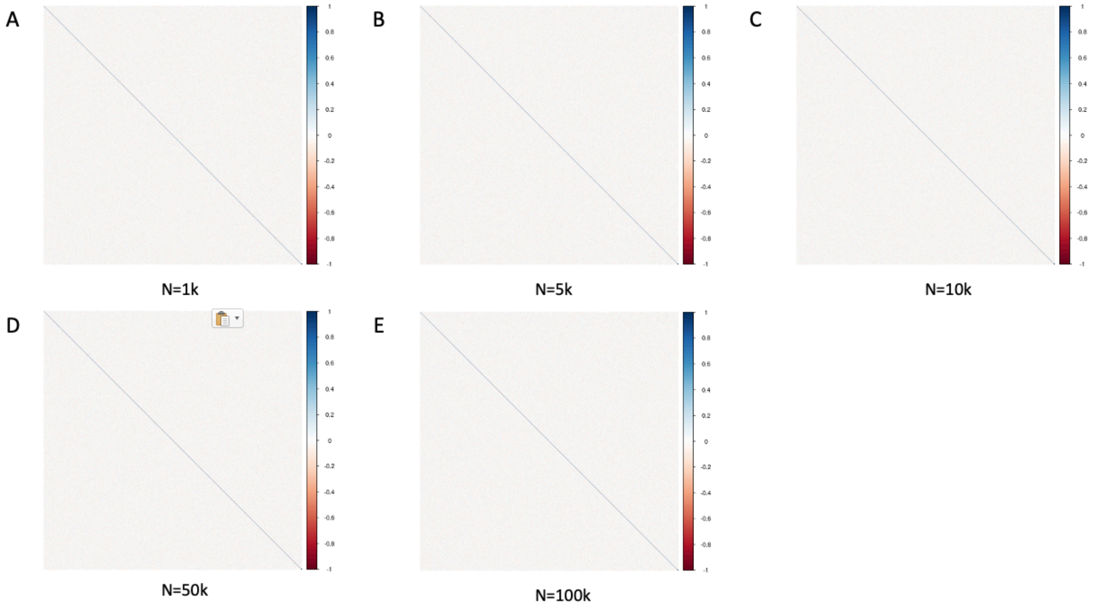

Figure A.1: Individual-level pair-wise correlation of EX-PRS. Panels A-E: sample sizes are 1k/5k/10k/50k/100k, respectively.

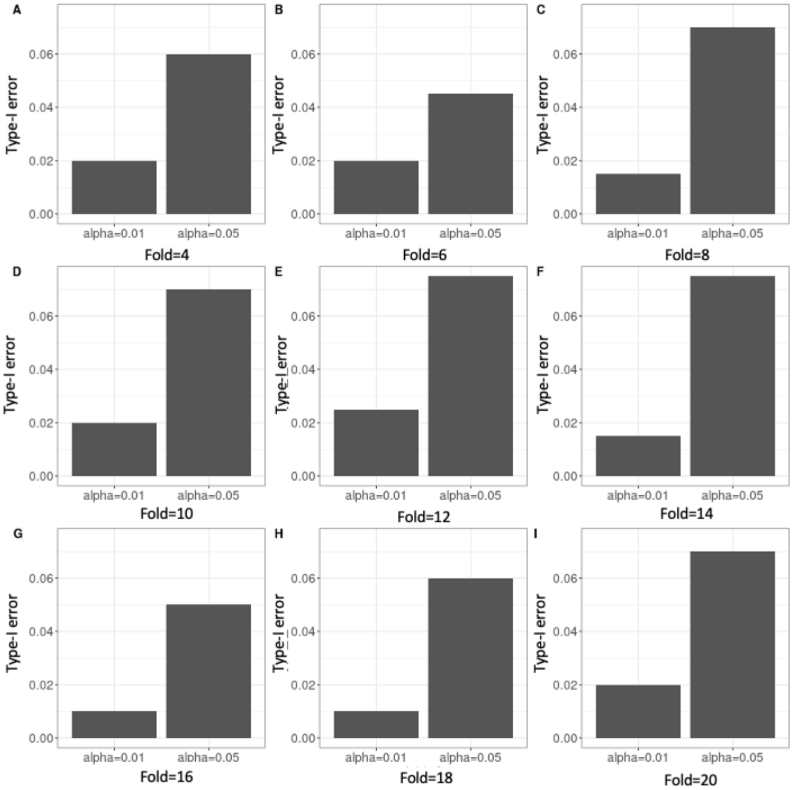

Figure A.2: No type-I error inflation across various choices of number of folds. Panels A-I illustrate type-I error of EA CV-PRS with the number of folds being 4/6/8/10/12/14/16/18/20 respectively.

B APPENDIX OF "PARTITIONING HERITABILITY AND GENETIC COVARIANCE BY DIRECT AND INDIRECT EFFECT PATHS"

B.1 Supplementary Tables

Parameters	Setting1	Setting2
$\sigma^2_{\alpha_1}$	0.202	0.180
$\sigma^2_{\beta_1}$	0.056	0.025
ρ_{ϵ_1}	0.099	0.099
ρ_{α}	0.000	0.110
ρ_{β}	0.000	-0.050
η	0.038	0.038
δ	0.007	0.007
$\sigma^2_{\alpha_2}$	0.246	0.22
$\sigma^2_{\beta_2}$	0.057	0.024
ρ_{ϵ_2}	0.198	0.198
$\rho_{\alpha_1\beta_2}$	0.012	0.012
$\rho_{\alpha_1\beta_1}$	0.040	0.060
$\rho_{\alpha_2\beta_2}$	0.007	0.03
$\rho_{\alpha_2\beta_1}$	0.015	0.015

Table B.1: Parameter settings for simulations.

	height		BMI		overall health		EA		income	
	EST	SE	EST	SE	EST	SE	EST	SE	EST	SE
$\sigma^2_{\alpha_1}$	0.102	0.036	0.146	0.036	0.015	0.036	0.061	0.036	0.032	0.035
$\sigma^2_{\beta_1}$	0.026	0.022	0.006	0.021	0.014	0.021	0.044	0.021	0.002	0.021
ρ_{ϵ_1}	0.099	0.022	0.198	0.022	0.076	0.020	0.214	0.022	0.235	0.020
$\sigma^2_{\alpha_1\beta_1}$	0.040	0.015	0.007	0.015	0.018	0.015	-0.003	0.015	0.014	0.014

Table B.2: Estimates for single traits analysis.

Point estimates										
Trait 1	height	height	height	height	BMI	BMI	BMI	EA	overall health	overall health
Trait 2	BMI	EA	overall health	income	EA	overall health	income	income	EA	income
$\sigma^2_{\alpha_1}$	0.102	0.102	0.102	0.102	0.146	0.146	0.146	0.061	0.015	0.015
$\sigma^2_{\beta_1}$	0.026	0.026	0.026	0.026	0.006	0.006	0.006	0.044	0.014	0.014
ρ_{ϵ_1}	0.099	0.099	0.099	0.099	0.198	0.198	0.198	0.214	0.076	0.076
ρ_α	-0.024	-0.020	0.008	0.024	-0.001	0.044	-0.015	0.044	-0.014	-0.005
ρ_β	-0.020	0.015	-0.001	0.032	-0.007	0.015	-0.002	0.032	-0.019	0.003
η	0.038	0.117	0.024	0.161	-0.044	0.205	-0.060	0.284	-0.088	-0.166
δ	0.007	0.018	-0.006	0.062	-0.042	0.058	-0.054	0.166	-0.062	-0.068
$\sigma^2_{\alpha_2}$	0.146	0.061	0.015	0.032	0.061	0.015	0.032	0.032	0.061	0.032
$\sigma^2_{\beta_2}$	0.006	0.044	0.014	0.002	0.044	0.014	0.002	0.002	0.044	0.002
ρ_{ϵ_2}	0.198	0.214	0.076	0.235	0.214	0.076	0.235	0.235	0.214	0.235
$\rho_{\alpha_1\beta_2}$	-0.012	0.022	-0.023	-0.029	-0.011	0.002	0.005	-0.003	0.003	-0.023
$\rho_{\alpha_1\beta_1}$	0.040	0.040	0.040	0.040	0.007	0.007	0.007	-0.003	0.018	0.018
$\rho_{\alpha_2\beta_2}$	0.007	-0.003	0.018	0.014	-0.003	0.018	0.014	0.014	-0.003	0.014
$\rho_{\alpha_2\beta_1}$	0.015	0.010	-0.015	-0.010	-0.037	0.003	-0.028	-0.012	-0.016	-0.020
SE										
$\sigma^2_{\alpha_1}$	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036
$\sigma^2_{\beta_1}$	0.022	0.022	0.022	0.022	0.021	0.021	0.021	0.021	0.021	0.021
ρ_{ϵ_1}	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.020	0.020
ρ_α	0.012	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.009	0.009
ρ_β	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.006	0.006
η	0.016	0.015	0.016	0.016	0.015	0.017	0.015	0.017	0.015	0.016
δ	0.016	0.016	0.014	0.014	0.014	0.016	0.014	0.019	0.014	0.015
$\sigma^2_{\alpha_2}$	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035
$\sigma^2_{\beta_2}$	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021
ρ_{ϵ_2}	0.022	0.022	0.020	0.020	0.022	0.020	0.020	0.020	0.022	0.020
$\rho_{\alpha_1\beta_2}$	0.009	0.009	0.009	0.009	0.009	0.008	0.010	0.009	0.007	0.008
$\rho_{\alpha_1\beta_1}$	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015
$\rho_{\alpha_2\beta_2}$	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014
$\rho_{\alpha_2\beta_1}$	0.009	0.009	0.009	0.009	0.009	0.008	0.009	0.009	0.007	0.008

Table B.3: Estimates for trait pair analysis

Notations	Formula
V_1	$\frac{1}{M}\begin{bmatrix} XX^T & 0 \\ 0 & 0 \end{bmatrix} - I_{4N}$
V_2	$\frac{1}{M}\begin{bmatrix} (X_m + X_p)(X_m + X_p)^T & 0 \\ 0 & 0 \end{bmatrix} - 2I_{4N}$
V_3	$\begin{bmatrix} I_N \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix}$
V_4	$\frac{1}{M}\begin{bmatrix} 0 & XX^T \\ XX^T & 0 \end{bmatrix}$
V_5	$\frac{1}{M}\begin{bmatrix} 0 & (X_m + X_p)(X_m + X_p)^T \\ (X_m + X_p)(X_m + X_p)^T & 0 \end{bmatrix}$
V_6	$\begin{bmatrix} 0 & I_{2N} \\ I_{2N} & 0 \end{bmatrix}$
V_7	$\begin{bmatrix} 0 & I_N \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ I_N \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & 0 \end{bmatrix}$
V_8	$\frac{1}{M}\begin{bmatrix} 0 & 0 \\ 0 & XX^T \end{bmatrix} - I_{4N}$
V_9	$\frac{1}{M}\begin{bmatrix} 0 & 0 \\ 0 & (X_m + X_p)(X_m + X_p)^T \end{bmatrix} - 2I_{4N}$
V_{10}	$\begin{bmatrix} 0 & 0 \\ 0 & I_N \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{bmatrix}$
V_{11}	$\frac{1}{M}\begin{bmatrix} 0 & X(X_m + X_p)^T \\ (X_m + X_p)X^T & 0 \end{bmatrix}$
V_{12}	$\frac{1}{M}\begin{bmatrix} X(X_m + X_p)^T + (X_m + X_p)X^T & 0 \\ 0 & 0 \end{bmatrix} - 2I_{4N}$
V_{13}	$\frac{1}{M}\begin{bmatrix} 0 & 0 \\ 0 & X(X_m + X_p)^T + (X_m + X_p)X^T \end{bmatrix} - 2I_{4N}$
V_{14}	$\frac{1}{M}\begin{bmatrix} 0 & (X_m + X_p)X^T \\ X(X_m + X_p)^T & 0 \end{bmatrix}$

Table B.4: Sample relatedness matrices for cov(Y)

B.2 Supplementary Figures

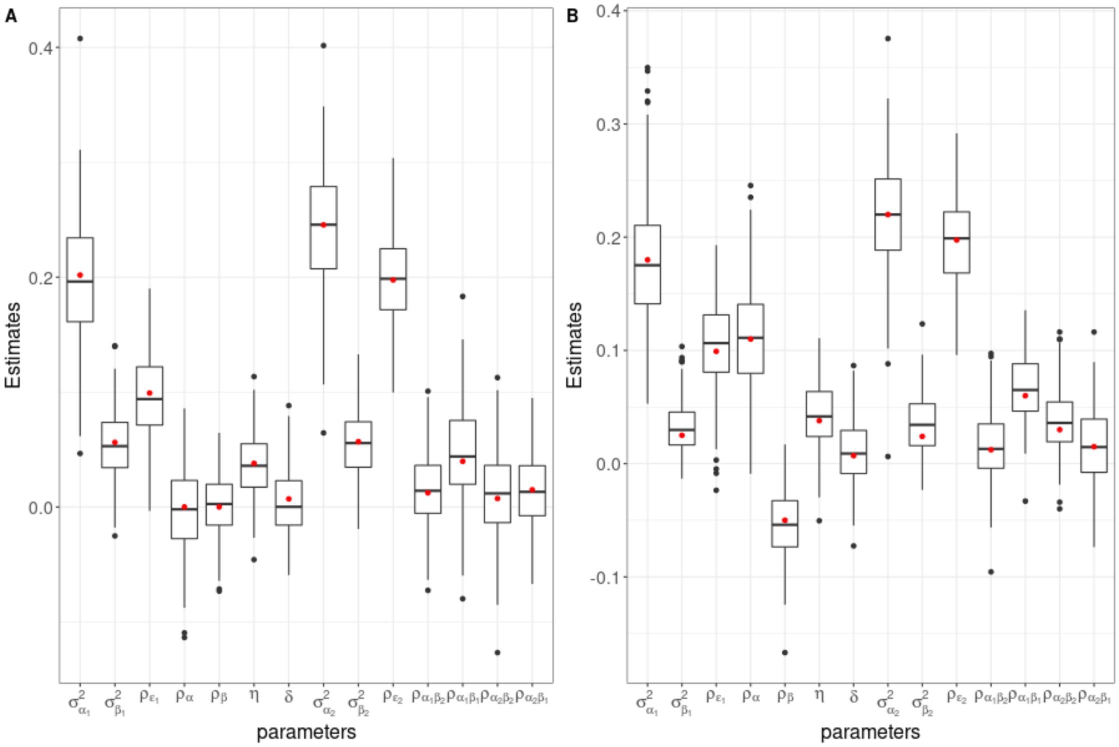

Figure B.1: Box plots for estimates of simulation. Red dots: pre-determined true values. A: boxplot for setting 1. B: boxplot for setting 2.

REFERENCES

Abraham, Gad, Yixuan Qiu, and Michael Inouye. 2017. Flashpca2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics* 33(17): 2776–2778.

Ahn, M., Wenbin Zhang Hh Fau Lu, and W. Lu. Moment-based method for random effects selection in linear mixed models (1017-0405 (Print)).

Arlot, Sylvain, and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys* 4:40–79.

Arnau-Soler, Aleix, Erin Macdonald-Dunlop, Mark J. Adams, Toni-Kim Clarke, Donald J. MacIntyre, Keith Milburn, Lauren Navrady, Caroline Hayward, Andrew M. McIntosh, and Pippa A. Thomson. 2019. Genome-wide by environment interaction studies of depressive symptoms and psychosocial stress in uk biobank and generation scotland. *Translational psychiatry* 9(1):1–13.

Auton, Adam, GonÃ§alo R. Abecasis, David M. Altshuler, Richard M. Durbin, GonÃ§alo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W. Albrecht, et al. 2015. A global reference for human genetic variation. *Nature* 526(7571):68–74.

Barcellos Silvia, H., S. Carvalho Leandro, and Patrick Turley. 2018. Education can reduce health differences related to genetic risk of obesity. *Proceedings of the National Academy of Sciences* 115(42):E9765–E9772. Doi: 10.1073/pnas.1802909115.

Bates, Stephen, Trevor Hastie, and Robert Tibshirani. 2021. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*.

Bates, Timothy C., Brion S. Maher, Sarah E. Medland, Kerrie McAloney, Margaret J. Wright, Narelle K. Hansell, Kenneth S. Kendler, Nicholas G. Martin, and Nathan A. Gillespie. 2018. The nature of nurture: Using a virtual-parent design to test parenting effects on children's educational attainment in genotyped families. *Twin Research and Human Genetics* 21(2):73–83.

Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. Variation in the heritability of educational attainment: An international meta-analysis. *Social Forces* 92(1):109–140.

Brown, B. C., C. J. Ye, A. L. Price, and N. Zaitlen. 2016. Transethnic genetic-correlation estimates from summary statistics. *Am J Hum Genet* 99(1):76–88. 1537-6605 Brown, Brielin C Asian Genetic Epidemiology Network Type 2 Diabetes Consortium Ye, Chun Jimmie Price, Alkes L Zaitlen, Noah K25 HL121295/HL/NHLBI NIH HHS/United States R01 HG006399/HG/NHGRI NIH HHS/United States Journal Article 2016/06/21 Am J Hum Genet. 2016 Jul 7;99(1):76-88. doi: 10.1016/j.ajhg.2016.05.001. Epub 2016 Jun 16.

Bulik-Sullivan, Brendan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, Po-Ru Loh, Laramie Duncan, John R. B. Perry, Nick Patterson, Elise B.

Robinson, Mark J. Daly, Alkes L. Price, Benjamin M. Neale, Consortium ReproGen, Consortium Psychiatric Genomics, and Consortium Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control. 2015a. An atlas of genetic correlations across human diseases and traits. *Nature Genetics* 47(11):1236–1241.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, Benjamin M. Neale, and Consortium Schizophrenia Working Group of the Psychiatric Genomics. 2015b. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 47(3):291–295.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, and Jared O'Connell. 2018. The uk biobank resource with deep phenotyping and genomic data. *Nature* 562(7726):203–209.

Cheesman, Rosa, Avina Hunjan, Jonathan RI Coleman, Yasmin Ahmadzadeh, Robert Plomin, Tom A McAdams, Thalia C Eley, and Gerome Breen. 2020. Comparison of adopted and nonadopted individuals reveals geneâŁ"environment interplay for education in the uk biobank. *Psychological science* 31(5):582–591.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F O'Reilly. 2020. Tutorial: a guide to performing polygenic risk score analyses. *Nature protocols* 15(9):2759–2772.

Choi, Shing Wan, and Paul F O'Reilly. 2019. Prsice-2: Polygenic risk score software for biobank-scale data. *Gigascience* 8(7):giz082.

Dietterich, Thomas G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10(7):1895–1923.

Domingue, Benjamin W, and Jason Fletcher. 2020. Separating measured genetic and environmental effects: Evidence linking parental genotype and adopted child outcomes. *Behavior genetics* 50(5):301–309.

Duncan, Laramie, Hanyang Shen, Bizu Gelaye, J Meijsen, K Ressler, M Feldman, R Peterson, and Ben Domingue. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications* 10(1):1–9.

Efron, Bradley. 1982. *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, Pa. : Society for Industrial and Applied Mathematics, 1982. Bibliography: pages 91-92.

Fletcher, Jason, Yuchang Wu, Tianchang Li, and Qiongshi Lu. 2021. Interpreting polygenic score effects in sibling analysis. *bioRxiv* 2021.07.16.452740.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W. Smoller. 2019. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications* 10(1):1776.

Guo, Hanmin, Lin Hou, Yu Shi, Sheng Chih Jin, Xue Zeng, Boyang Li, Richard P. Lifton, Martina Brueckner, Hongyu Zhao, and Qiongshi Lu. 2021a. Quantifying concordant genetic effects of de novo mutations on multiple disorders. *bioRxiv* 2021.06.13.448234.

Guo, Hanmin, James J. Li, Qiongshi Lu, and Lin Hou. 2021b. Detecting local genetic correlations with scan statistics. *Nature Communications* 12(1):2033.

Hoffmann, Thomas J, Georg B Ehret, Priyanka Nandakumar, Dilrini Ranatunga, Catherine Schaefer, Pui-Yan Kwok, Carlos Iribarren, Aravinda Chakravarti, and Neil Risch. 2017. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature genetics* 49(1):54–64.

Hu, Yiming, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran Xu, and Hongyu Zhao. 2017. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLOS Computational Biology* 13(6): e1005589.

Hwang, Liang-Dar, Justin D Tubbs, Justin Luong, Mischa Lundberg, Gunn-Helen Moen, Geng Wang, Nicole M Warrington, Pak C Sham, Gabriel Cuellar-Partida, and David M Evans. 2020. Estimating indirect parental genetic effects on offspring phenotypes using virtual parental genotypes derived from sibling and half sibling pairs. *PLoS genetics* 16(10):e1009154.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor, and Sekar Kathiresan. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* 50(9):1219–1224.

Koellinger Philipp, D., and K. Paige Harden. 2018. Using nature to understand nurture. *Science* 359(6374):386–387. Doi: 10.1126/science.aar6429.

Kong, Augustine, Gudmar Thorleifsson, L. Frigge Michael, J. Vilhjalmsson Bjarni, I. Young Alexander, E. Thorgeirsson Thorgeir, Stefania Benonisdottir, Asmundur Oddsson, V. Halldorsson Bjarni, Gisli Masson, F. Gudbjartsson Daniel, Agnar Helgason, Gyda Bjornsdottir, Unnur Thorsteinsdottir, and Kari Stefansson. 2018. The nature of nurture: Effects of parental genotypes. *Science* 359(6374):424–428. Doi: 10.1126/science.aan6877.

Lai, HuiChuan J., Jie Song, Qiongshi Lu, Sangita G. Murali, Manavalan Gajapathy, Brandon M. Wilk, Donna M. Brown, Elizabeth A. Worthey, and Philip M. Farrell. 2022. Genetic factors help explain the variable responses of young children with cystic fibrosis to vitamin d supplements. *Clinical Nutrition ESPEN*.

Lee, James J., Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson LinnÃ©r, Mark Alan Fontana, Tushar Kundu, Chanwook Lee, Hui Li, Ruoxi Li, Rebecca Royer, Pascal N. Timshel, Raymond K. Walters, Emily A. Willoughby, LoÃ¯c Yengo, Team andMe Research, Cogent, Consortium Social Science Genetic Association, Maris Alver, Yanchun Bao, David W. Clark, Felix R. Day, Nicholas A. Furlotte, Peter K. Joshi, Kathryn E. Kemper, Aaron Kleinman, Claudia Langenberg, Reedik MÃ¤gi, Joey W. Trampush, Shefali Setia Verma, Yang

Wu, Max Lam, Jing Hua Zhao, Zhili Zheng, Jason D. Boardman, Harry Campbell, Jeremy Freese, Kathleen Mullan Harris, Caroline Hayward, Pamela Herd, Meena Kumari, Todd Lencz, Jian'an Luan, Anil K. Malhotra, Andres Metspalu, Lili Milani, Ken K. Ong, John R. B. Perry, David J. Porteous, Marylyn D. Ritchie, Melissa C. Smart, Blair H. Smith, Joyce Y. Tung, Nicholas J. Wareham, James F. Wilson, Jonathan P. Beauchamp, Dalton C. Conley, TÃµnu Esko, Steven F. Lehrer, Patrik K. E. Magnusson, Sven Oskarsson, Tune H. Pers, Matthew R. Robinson, Kevin Thom, Chelsea Watson, Christopher F. Chabris, Michelle N. Meyer, David I. Laibson, Jian Yang, Magnus Johannesson, Philipp D. Koellinger, Patrick Turley, Peter M. Visscher, Daniel J. Benjamin, and David Cesarini. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics* 50(8):1112–1121. (Cognitive Genomics Consortium) 30038396[pmid] PMC6393768[pmcid] 10.1038/s41588-018-0147-3[PII].

Loh, Po-Ru, Gaurav Bhatia, Alexander Gusev, Hilary K. Finucane, Brendan K. Bulik-Sullivan, Samuela J. Pollack, Consortium Schizophrenia Working Group of Psychiatric Genomics, Teresa R. de Candia, Sang Hong Lee, Naomi R. Wray, Kenneth S. Kendler, Michael C. O'Donovan, Benjamin M. Neale, Nick Patterson, and Alkes L. Price. 2015. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics* 47(12):1385–1392. 26523775[pmid] PMC4666835[pmcid] ng.3431[PII].

Loos, Ruth J. F. 2020. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications* 11(1):5900.

Lu, Qiongshi, Boyang Li, Derek Ou, Margret Erlendsdottir, Ryan L. Powles, Tony Jiang, Yiming Hu, David Chang, Chentian Jin, Wei Dai, Qidu He, Zefeng Liu, Shubhabrata Mukherjee, Paul K. Crane, and Hongyu Zhao. 2017. A powerful approach to estimating annotation-stratified genetic covariance via gwas summary statistics. *American journal of human genetics* 101(6):939–964. 29220677[pmid] PMC5812911[pmcid] S0002-9297(17)30453-6[PII].

Lynch, Michael, Bruce Walsh, et al. 1998. *Genetics and analysis of quantitative traits*, vol. 1. Sinauer Sunderland, MA.

Mak, Timothy Shin Heng, Robert Milan Porsch, Shing Wan Choi, and Pak Chung Sham. 2018. Polygenic scores for uk biobank scale data. *BioRxiv* 252270.

Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, MichÃ¨le Sale, and Wei-Min Chen. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–2873.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* 51(4):584–591.

Mefford, Joel, Danny Park, Zhili Zheng, Arthur Ko, Mika Ala-Korpela, Markku Laakso, Päivi Pajukanta, Jian Yang, John Witte, and Noah Zaitlen. 2020. Efficient estimation and applications of cross-validated genetic predictions to polygenic risk scores and linear mixed models. *Journal of Computational Biology* 27(4):599–612.

Metzger, Molly W., and Thomas W. McDade. 2010. Breastfeeding as obesity prevention in the united states: A sibling difference model. *American Journal of Human Biology* 22(3):291–296. Https://doi.org/10.1002/ajhb.20982.

Miao, Jiacheng, Hanmin Guo, Gefei Song, Zijie Zhao, Lin Hou, and Qiongshi Lu. 2022. Quantifying portable genetic effects and improving cross-ancestry genetic prediction with gwas summary statistics. *bioRxiv* 2022.05.26.493528.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* 81(3):559–575.

Purcell, Shaun M., Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, Patrick F. Sullivan, Pamela Sklar, Shaun M. Purcell, Jennifer L. Stone, Patrick F. Sullivan, Douglas M. Ruderfer, Andrew McQuillin, Derek W. Morris, Colm T. O’Dushlaine, Aiden Corvin, Peter A. Holmans, Michael C. O’Donovan, Pamela Sklar, Naomi R. Wray, Stuart Macgregor, Pamela Sklar, Patrick F. Sullivan, Michael C. O’Donovan, Peter M. Visscher, Hugh Gurling, Douglas H. R. Blackwood, Aiden Corvin, Nick J. Craddock, Michael Gill, Christina M. Hultman, George K. Kirov, Paul Lichtenstein, Andrew McQuillin, Walter J. Muir, Michael C. O’Donovan, Michael J. Owen, Carlos N. Pato, Shaun M. Purcell, Edward M. Scolnick, David St Clair, Jennifer L. Stone, Patrick F. Sullivan, Pamela Sklar, Michael C. O’Donovan, George K. Kirov, Nick J. Craddock, Peter A. Holmans, Nigel M. Williams, Lyudmila Georgieva, Ivan Nikolov, N. Norton, H. Williams, Draga Toncheva, Vihra Milanova, Michael J. Owen, Christina M. Hultman, Paul Lichtenstein, Emma F. Thelander, Patrick Sullivan, Derek W. Morris, Colm T. O’Dushlaine, Elaine Kenny, Emma M. Quinn, Michael Gill, Aiden Corvin, Andrew McQuillin, Khalid Choudhury, Susmita Datta, Jonathan Pimm, Srinivasa Thirumalai, Vinay Puri, Robert Krasucki, Jacob Lawrence, Digby Quested, Nicholas Bass, Hugh Gurling, Caroline Crombie, Gillian Fraser, Soh Leh Kuan, Nicholas Walker, David St Clair, Douglas H. R. Blackwood, Walter J. Muir, Kevin A. McGhee, Ben Pickard, Pat Malloy, Alan W. Maclean, Margaret Van Beck, Naomi R. Wray, Stuart Macgregor, Peter M. Visscher, Michele T. Pato, Helena Medeiros, Frank Middleton, Celia Carvalho, Christopher Morley, Ayman Fanous, David Conti, James A. Knowles, Carlos Paz Ferreira, et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256):748–752. (Leader).

Qi, Qibin, Adrienne M Stilp, Tamar Sofer, Jee-Young Moon, Bertha Hidalgo, Adam A Szpiro, Tao Wang, Maggie CY Ng, Xiuqing Guo, MEta analysis of type 2 DIabetes in African Americans (MEDIA) Consortium, et al. 2017. Genetics of type 2 diabetes in us hispanic/latino individuals: results from the hispanic community health study/study of latinos (hchs/sol). *Diabetes* 66(5):1419–1425.

van Rheenen, Wouter, Wouter J. Peyrot, Andrew J. Schork, S. Hong Lee, and Naomi R. Wray. 2019. Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics* 20(10):567–581.

Ruan, Yunfeng, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Yong Min Ahn, Kazufumi Akiyama, Makoto Arai, Ji Hyun Baek, Wei J. Chen, Young-Chul Chung, Gang Feng, Kumiko Fujii, Stephen J. Glatt, Kyooseob Ha, Kotaro Hattori, Teruhiko Higuchi, Akitoyo Hishimoto, Kyung Sue Hong, Yasue Horiuchi, Hai-Gwo Hwu, Masashi Ikeda, Sayuri Ishiwata, Masanari Itokawa, Nakao Iwata, Eun-Jeong Joo, Rene S. Kahn, Sung-Wan Kim, Se Joo Kim, Se Hyun Kim, Makoto Kinoshita, Hiroshi Kunugi, Agung Kusumawardhani, Jimmy Lee, Byung Dae Lee, Heon-Jeong Lee, Jianjun Liu, Ruize Liu, Xiancang Ma, Woojae Myung, Shusuke Numata, Tetsuro Ohmori, Ikuo Otsuka, Yuji Ozeki, Sibylle G. Schwab, Wenzhao Shi, Kazutaka Shimoda, Kang Sim, Ichiro Sora, Jinsong Tang, Tomoko Toyota, Ming Tsuang, Dieter B. Wildenauer, Hong-Hee Won, Takeo Yoshikawa, Alice Zheng, Feng Zhu, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang, Tian Ge, and Initiatives Stanley Global Asia. 2022. Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics* 54(5):573–580.

Shao, Jun. 1993. Linear model selection by cross-validation. *Journal of the American statistical Association* 88(422):486–494.

Shi, Huwenbo, Nicholas Mancuso, Sarah Spendlove, and Bogdan Pasaniuc. 2017. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics* 101(5):737–751.

Shin, Jisu, and Sang Hong Lee. 2021. Gxesum: a novel approach to estimate the phenotypic variance explained by genome-wide gxe interaction based on gwas summary statistics for biobank-scale data. *Genome Biology* 22(1):183.

Sivula, Tuomas, Måns Magnusson, and Aki Vehtari. 2020. Uncertainty in bayesian leave-one-out cross-validation based model comparison. *arXiv preprint arXiv:2008.10296*.

Sodini, Sebastian M, Kathryn E Kemper, Naomi R Wray, and Maciej Trzaskowski. 2018. Comparison of genotypic and phenotypic correlations: Cheverud's conjecture in humans. *Genetics* 209(3):941–948.

Trejo, Sam, and Benjamin W Domingue. 2018. Genetic nature or genetic nurture? introducing social genetic parameters to quantify bias in polygenic score analyses. *Biodemography and Social Biology* 64(3-4):187–215.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. Genome-wide association studies. *Nature Reviews Methods Primers* 1(1):59.

Vilhjálmsson, B. J., J. Yang, H. K. Finucane, A. Gusev, S. LindstrÃ¶m, S. Ripke, G. Genovese, P. R. Loh, G. Bhatia, R. Do, T. Hayeck, H. H. Won, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P. M. Visscher, P. Kraft, N. Patterson, and A. L. Price. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet* 97(4):576–92.

Wang, Biyao, Jessie R. Baldwin, Tabea Schoeler, Rosa Cheesman, Wikus Barkhuizen, Frank Dudbridge, David Bann, Tim T. Morris, and Jean-Baptiste Pingault. 2021. Robust genetic nurture effects on education: A systematic review and meta-analysis based on 38,654 families across 8 cohorts. *The American Journal of Human Genetics* 108(9):1780–1791.

Weiner, Daniel J., Emilie M. Wigdor, Stephan Ripke, Raymond K. Walters, Jack A. Kosmicki, Jakob Grove, Kaitlin E. Samocha, Jacqueline I. Goldstein, Aysu Okbay, Jonas Bybjerg-Grauholm, Thomas Werge, David M. Hougaard, Jacob Taylor, Marie BÃ¦kvad-Hansen, Ashley Dumont, Christine Hansen, Thomas F. Hansen, Daniel Howrigan, Manuel Mattheisen, Jennifer Moran, Ole Mors, Merete Nordentoft, Bent NÃ¸rgaard-Pedersen, Timothy Poterba, Jesper Poulsen, Christine Stevens, Verneri Anttila, Peter Holmans, Hailiang Huang, Lambertus Klei, Phil H. Lee, Sarah E. Medland, Benjamin Neale, Lauren A. Weiss, Lonnie Zwaigenbaum, Timothy W. Yu, Kerstin Wittemeyer, A. Jeremy Willsey, Ellen M. Wijsman, Thomas H. Wassink, Regina Waltes, Christopher A. Walsh, Simon Wallace, Jacob A. S. Vorstman, Veronica J. Vieland, Astrid M. Vicente, Herman van Engeland, Kathryn Tsang, Ann P. Thompson, Peter Szatmari, Oscar Svantesson, Stacy Steinberg, Kari Stefansson, Hreinn Stefansson, Matthew W. State, Latha Soorya, Teimuraz Silagadze, Stephen W. Scherer, Gerard D. Schellenberg, Sven Sandin, Evald Saemundsen, Guy A. Rouleau, Bernadette RogÃ©, Kathryn Roeder, Wendy Roberts, Jennifer Reichert, Abraham Reichenberg, Karola RehnstrÃ¶m, Regina Regan, Fritz Poustka, Christopher S. Poultney, Joseph Piven, Dalila Pinto, Margaret A. Pericak-Vance, Milica Pejovic-Milovancevic, Marianne G. Pedersen, Carsten B. Pedersen, Andrew D. Paterson, Jeremy R. Parr, Alistair T. Pagnamenta, Guiomar Oliveira, John I. Nurnberger, Michael T. Murtha, Susana Mouga, Eric M. Morrow, Daniel Moreno De Luca, Anthony P. Monaco, Nancy Minshew, Alison Merikangas, William M. McMahon, Susan G. McGrew, Igor Martsenkovsky, Donna M. Martin, Shrikant M. Mane, Pall Magnusson, Tiago Magalhaes, Elena Maestrini, Jennifer K. Lowe, Catherine Lord, Pat Levitt, et al. 2017. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature Genetics* 49(7):978–985.

Werme, J., S. van der Sluis, D. Posthuma, and C. A. de Leeuw. 2021a. Lava: An integrated framework for local genetic correlation analysis. *bioRxiv* 2020.12.31.424652.

Werme, Josefin, Sophie van der Sluis, Danielle Posthuma, and Christiaan A. de Leeuw. 2021b. Genome-wide gene-environment interactions in neuroticism: an exploratory study across 25 environments. *Translational Psychiatry* 11(1):180.

Willoughby, Emily A, Matt McGue, William G Iacono, Aldo Rustichini, and James J Lee. 2021. The role of parental genotype in predicting offspring years of education: Evidence for genetic nurture. *Molecular psychiatry* 26(8):3896–3904.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics* 46(11):1173–1186.

Wray, Naomi R, Michael E Goddard, and Peter M Visscher. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research* 17(10):1520–1528.

Wu, Yuchang, Xiaoyuan Zhong, Yunong Lin, Zijie Zhao, Jiawen Chen, Boyan Zheng, J. Li James, M. Fletcher Jason, and Qiongshi Lu. 2021. Estimating genetic nurture with summary statistics of multigenerational genome-wide association studies. *Proceedings of the National Academy of Sciences* 118(25):e2023184118. Doi: 10.1073/pnas.2023184118.

Xu, Qing-Song, and Yi-Zeng Liang. 2001. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* 56(1):1–11.

Yang, J., Brian P. Benyamin B Fau McEvoy, Scott McEvoy Bp Fau Gordon, Anjali K. Gordon S Fau Henders, Dale R. Henders Ak Fau Nyholt, Pamela A. Nyholt Dr Fau Madden, Andrew C. Madden Pa Fau Heath, Nicholas G. Heath Ac Fau Martin, Grant W. Martin Ng Fau Montgomery, Michael E. Montgomery Gw Fau Goddard, Peter M. Goddard Me Fau Visscher, and P. M. Visscher. Common snps explain a large proportion of the heritability for human height (1546-1718 (Electronic)).

Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. Gcta: a tool for genome-wide complex trait analysis. *American journal of human genetics* 88(1):76–82. 21167468[pmid] PMC3014363[pmcid] S0002-9297(10)00598-7[PII].

Yang, Yuhong. 2007. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics* 35(6):2450–2473.

Young, Alexander I, Stefania Benonisdottir, Molly Przeworski, and Augustine Kong. 2019. Deconstructing the sources of genotype-phenotype associations in humans. *Science* 365(6460):1396–1400.

Young, Alexander I, Seyed Moeen Nehzati, Chanwook Lee, Stefania Benonisdottir, David Cesarini, Daniel J Benjamin, Patrick Turley, and Augustine Kong. 2020. Mendelian imputation of parental genotypes for genome-wide estimation of direct and indirect genetic effects. *BioRxiv*.

de Zeeuw, Eveline L, Jouke-Jan Hottenga, Klaasjan G Ouwens, Conor V Dolan, Erik A Ehli, Gareth E Davies, Dorret I Boomsma, and Elsje van Bergen. 2020. Intergenerational transmission of education and adhd: effects of parental genotypes. *Behavior Genetics* 50(4):221–232.

Zhang, Ping. 1993. Model selection via multifold cross validation. *The annals of statistics* 299–313.

Zhang, Yiliang, Youshu Cheng, Wei Jiang, Yixuan Ye, Qiongshi Lu, and Hongyu Zhao. 2021a. Comparison of methods for estimating genetic correlation between complex traits using gwas summary statistics. *Briefings in Bioinformatics* 22(5): bbaa442.

Zhang, Yiliang, Youshu Cheng, Yixuan Ye, Wei Jiang, Qiongshi Lu, and Hongyu Zhao. 2021b. Estimating genetic correlation jointly using individual-level and summary-level gwas data. *bioRxiv* 2021.08.18.456908.

Zhang, Yiliang, Qiongshi Lu, Yixuan Ye, Kunling Huang, Wei Liu, Yuchang Wu, Xiaoyuan Zhong, Boyang Li, Zhaolong Yu, Brittany G. Travers, Donna M. Werling, James J. Li, and Hongyu Zhao. 2021c. Supergnova: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome Biology* 22(1):262.

Zhao, Zijie, Yanyao Yi, Jie Song, Yuchang Wu, Xiaoyuan Zhong, Yupei Lin, Timothy J Hohman, Jason Fletcher, and Qiongshi Lu. 2021. Pumas: fine-tuning polygenic risk scores with gwas summary statistics. *Genome biology* 22(1):1–19.

Zijie Zhao, Tuo Wang Qiongshi Lu, Jie Song. 2021. Polygenic risk scores: effect estimation and model optimization. *Quantitative Biology* 9(2):133.