UNIVERSITY OF CALIFORNIA

Los Angeles

Efficient Methods for Understanding the Genetic Architecture of Complex Traits

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Yue Wu

2022

ABSTRACT OF THE DISSERTATION

Efficient Methods for Understanding the Genetic Architecture of Complex Traits

by

Yue Wu
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2022
Professor Sriram Sankararaman, Co-Chair
Professor Eleazar Eskin, Co-Chair

Understanding the genetic architecture of complex traits is a central goal of modern human genetics. Recent efforts focused on building large-scale biobanks, that collect genetic and trait data on large numbers of individuals, present exciting opportunities for understanding genetic architecture. However, these datasets also pose several statistical and computational challenges. In this dissertation, we consider a series of statistical models that allow us to infer aspects of the genetic architecture of single and multiple traits. Inference in these models is computationally challenging due to the size of the genetic data – consisting of millions of genetic variants measured across hundreds of thousands of individuals. We propose a series of scalable computational methods that can perform efficient inference in these models and apply these methods to data from the UK Biobank to showcase their utility.

The dissertation of Yue Wu is approved.

Quanquan Gu

Bogdan Pasaniuc

Eleazar Eskin, Committee Co-Chair

Sriram Sankararaman, Committee Co-Chair

University of California, Los Angeles

2022

To my parents, my grandmother, and Poppy

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

On my Ph.D. journey, I have been fortunate enough to meet many people who have supported me along my journey to this point in various ways, and I am very grateful for their help.

Firstly I would like to express my sincere gratitude to my committee members. Professor Sriram Sankararaman, my advisor, deserves special thanks for his continuous support in directing me to the study. I would like to express my appreciation for his great patience and dedication. I consider myself privileged to have worked with him and learned from his profound knowledge of statistics, human genetics, and machine learning. Throughout my studies, he not only taught me scientific knowledge but also taught me a mindset to approach scientific problems. I learned a bright attitude to navigate through frustration from him. And most importantly, his enthusiasm for science is infectious, and it has encouraged me to continue pursuing the problem I am interested in.

Professor Eleazar Eskin's computation genetics class back in my undergraduate program opened this whole new world to me. His class inspired me to continue exploring the challenges and excitement in this field. He also encouraged me to pursue a Ph.D. and has supported me over the passing years. I still remember the question Professor Bogdan Pasaniuc asked when he came to my poster at my very first conference in the first year of my Ph.D. He has provided valuable and constructive advice to my work. Professor Quanquan Gu's class is the most theoretical class I have taken at UCLA and from which I learned the most. I learned a rigorous way of thinking about neural networks. His class also inspired me and provided a new perspective on my Ph.D. work.

I have had many opportunities to collaborate and brainstorm with brilliant people from Sriram Lab, Zar lab, and Bogdan Lab. I would like to thank Farhad Hormozdiari and Joanne Joon Wha J. Joo, who were mentors that held my hand through the beginning of my academic journey. I have had many group meetings and journal clubs with Arun Durvasula, Chris Robles, Alec Chiu, Ruth Johnson, Ali Pazoki, Erin Molloy, April Wei, and Boyang Fu from Sriram lab work on diverse topics. It has been a lot of fun exploring new things together. Further, the discussions I have had with Serghei Mangul, Rob Brown, Dat Duang, Lisa Gai, Jennifer Zou, Kodi Collins from Zar lab, Kathy Burch, and Kangchen Hou from Bogdan lab have been inspiring.

Last but not the least, I would like to thank my family, for their endless love and unshakable trust in me.

VITA

2011–2016	B.S. Computer Science, minor in Mathematics, UCLA, Los Angeles, California.
2017–2021	Teaching Assistant, Department of Computer Science, UCLA
2021	Research Intern, Health Next, Microsoft
2016–2022	Graduate Student Researcher, Department of Computer Science, UCLA

PUBLICATIONS

Parts of the work in this thesis have appeared in the following publications:

Yue Wu, Kathryn S. Burch, Andrea Ganna, Paivi Pajukanta, Bogdan Pasaniuc, Sriram Sankararaman. "Fast estimation of genetic correlation for Biobank-scale data." **The American Journal of Human Genetics**, 2021.

Yue Wu, Sriram Sankararaman. "A scalable estimator of SNP heritability for Biobank-scale data." **Bioinformatics**, 2018.

Other relevant publications:

Ali Pazokitoroudi, **Yue Wu**, Kathryn S. Burch, Kangcheng Hou, Bogdan Pasaniuc, Sriram Sankararaman. "Efficient variance components analysis across millions of genomes." **Nature Communications**, 2020.

Yue Wu, Eleazar Eskin, Sriram Sankararaman. "A unifying framework of statistics

imputation." **Journal of Computational Biology**, 2020.

Kangcheng Hou, Kathryn S. Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, **Yue Wu**, Sriram Sankararaman, Bogdan Pasaniuc. "Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture." **Nature Genetics**, 2019.

Yue Wu *, Farhad Hormozdiari *, Joon Wha J. Joo, Eleazar Eskin. "Improving imputation accuracy by inferring causal variants in genetic studies." **Journal of Computational Biology**, 2018.

CHAPTER 1

Introduction

The main theme of this thesis is to develop efficient algorithms to enable the study of the genetic architecture of complex traits across large datasets. The key aspects of the genetic architecture that we are trying to understand in this thesis are the following:

- How much of the variation in a phenotype is explained by genetic variation?
- How much of the genetic effects of traits are shared? How do we quantify and identify the shared genetic architecture that reveals pleiotropy?
- Are the genetic effects shared between phenotypes distributed across the genome? Or could we identify hot spots for potential shared pathways? Can this relationship be explained by other traits?

1.1 Genetic and phenotypic variation

We first introduce the parameter termed *heritability*. Heritability is the proportion of variation in a trait that can be explained by genetic variation. Heritability is an important parameter in efforts to understand the genetic architecture of complex traits as well as in the design and interpretation of genome-wide association studies [1].

Attempts to understand the heritability of complex traits attributable to genome-wide SNP variation data have motivated the analysis of large datasets as well as the development of sophisticated tools to estimate heritability in these datasets.

In large-scale datasets such as the UK Biobank, the number of genetic variants M is substantially larger than the number of samples N, where $M >> N$. Thus, linear mixed models (LMM) have emerged as a key tool for heritability estimation [2, 3, 4, 5, 6]. In Chapter 2, we first introduce the concept of heritability and introduce the linear mixed model (LMM). We then discuss the computational difficulty of inference in an LMM and describe our method, RHE-reg, for efficient heritability estimation.

1.2 Shared genetic architecture among traits

We then try to understand the genetic architecture shared across multiple traits using the parameter termed *genetic correlation*. Genetic correlation is the correlation of the effect sizes across a set of genetic variants on a pair of traits [7]. For instance, if a genetic variant has exactly the same contribution to a pair of traits, the genetic correlation of the variant across the pair of traits would be 1. Genetic correlation can provide insights into shared genetic pathways and serve as a starting point to investigate pleiotropy and causal relationships among traits [8, 9, 10, 11].

In Chapter 3, we introduce two statistical models based on LMMs to estimate genetic correlations. The first models a pair of traits at a time while the second jointly models multiple traits. We discuss inference algorithms in these multi-variate LMMs and propose our method, SCORE, for efficiently estimating genetic correlations.

1.3 Localizing shared genetic architecture to specific regions

Having detected shared genetic effects across traits, the next question of interest is whether this relationship is uniformly distributed along the genome or whether it is enriched in certain regions [11, 12, 10]. In Chapter 4, we discuss extensions of the multivariate model developed in Chapter 3 into a multi-component multivariate mixed model.

CHAPTER 2

A scalable estimator of SNP heritability for Biobank-scale data

In this chapter, we endeavor to study *heritability*, *i.e.*, the proportion of variation in a trait that can be explained by genetic variation.

A central question in biology is to understand how much of the variation in a trait (phenotype) can be explained by genetics as opposed to environmental factors. The heritability of a trait is a central notion in quantifying the contribution of genetics to the variation in a trait. The heritability of a trait refers to the proportion of variation in the trait that can be explained by genetic variation [1]. The narrow-sense heritability (h^2) refers to the proportion of trait variation that can be explained by a linear function of genetic variation [13]. Beyond understanding the genetic basis of a phenotype, heritability determines the power of genetic association studies to detect genetic variants associated with a phenotype, the accuracy of using genetic data to predict a phenotype, as well as the response of a phenotype to natural and artificial selection [14].

While family-based studies enabled the estimation of heritability of a wide variety of traits, the availability of genome-wide genetic variation data has enabled a direct estimation of the heritability associated with genotyped SNPs, termed *SNP*

heritability. Initial attempts to estimate heritability from genomic data focused on the variation in a trait that could be explained by SNPs that were discovered to be significantly associated with the trait in a genome-wide association study (GWAS). These estimates were found to severely under-estimate the narrow-sense heritability, a phenomenon known as *missing heritability*. A major insight into the mystery of missing heritability emerged in [15] who showed that using all genotyped SNPs jointly to explain variation in a trait led to a substantially larger estimate of heritability than from SNPs that were found to be associated in GWAS. Subsequent analyses suggest that much of the missing heritability could be explained by the presence of a large number of SNPs of weak effects that has, in turn, motivated analyses of larger datasets.

Linear Mixed Models (LMMs) have emerged as a key analytical technique for estimating the heritability of complex traits using genome-wide SNP variation data. Beyond their application in estimating SNP heritability, LMMs are widely used in association tests where they are used to control for population stratification [2, 3, 4, 5, 6], in phenotype and disease risk prediction [16, 17, 18, 19, 15], and in understanding the relative contribution of genomic regions to variation in a trait of interest [15, 17, 18]. A key step in the application of LMMs is the estimation of their parameters, *i.e.*, often referred to as variance components. Estimation of variance components is a computationally challenging problem on genomic datasets containing large numbers of individuals and SNPs. The most commonly used method for variance components estimation in LMMs relies on maximizing the likelihood of the parameters. Often, a related estimator, known as the restricted maximum likelihood (REML) estimator, is preferred due to a reduced bias relative to maximum likelihood estimators. Both maximum likelihood as well as REML estimation, however, rely on computationally

intensive optimization problems. While a number of methods have been proposed to improve the computational efficiency of REML estimators [20, 21, 4, 5, 22, 23], all of these methods rely on iterative optimization algorithms that do not scale well to Biobank-scale datasets consisting of millions of individuals genotyped at tens of millions of SNPs. Further, REML has been shown to yield biased estimates of heritability in ascertained case-control studies [24, 25].

In this chapter, we propose a scalable randomized algorithm to estimate variance components of a linear mixed model. Our method is based on Haseman-Elston (HE) regression [26, 27, 28, 29], a Method-of-Moments (MoM) estimator of the heritability of a phenotype. The HE regression estimator, like other MoM estimators, tends to be statistically less efficient compared to REML. On the other hand, HE regression is computationally attractive as it leads to a set of linear equations in the variance components that can be solved analytically. While this property of HE regression is appealing, a key computational bottleneck in the application of HE regression is the computation of an $N \times N$ matrix that summarizes the relationship between all N pairs of individuals in the dataset. As a result, the computation and memory requirements of HE scale quadratically with the number of individuals.

Our randomized HE regression (RHE-reg) estimator relies on the observation that the key bottleneck in HE regression can be replaced by multiplying the $N \times M$ (individuals $\times$ SNPs) matrix of genotypes with a small number, B, of random vectors. This leads to a randomized estimator with runtime $\mathcal{O}(NMB)$ and memory requirements $\mathcal{O}(NM)$. Further, we leverage the observation that the genotype matrix has entries in a finite set, *i.e.*, $\{0, 1, 2\}$ so that the time complexity of the matrix-vector multiplication reduces to $\mathcal{O}(\frac{NMB}{max(\log_3(N), \log_3(M))})$ [30]. This additional gain in efficiency can be substantial when the number of SNPs or individuals is large. For example, in

the UK Biobank, N is of the order of 10^5 while M is of the order of 10^6. Thus, we propose an estimator of variance components with runtime $\mathcal{O}(\frac{NMB}{\max(\log_3(N),\log_3(M))} + NM)$ and memory requirement $\mathcal{O}(NM)$.

We apply the RHE-reg estimator to the problem of estimating SNP heritability. We show that our method yields unbiased SNP heritability estimates. While our method is statistically inefficient compared to REML (both because it is moment-based as well as the added randomization), we show in practice that the statistical inefficiency is minimal, particularly for large sample sizes. Further, our method is substantially more computationally efficient so that it can be effectively applied to whole-genome genotype data from hundreds of thousands of individuals. REML has been shown to yield biased estimates of heritability in ascertained case-control studies [28, 24] while the RHE-reg estimator can also be applied in this setting.

Finally, since variance component analysis is of interest beyond heritability estimation, the RHE-reg estimator can enable rapid estimation of variance components in all of the settings in which LMMs are used.

2.1 Models for quantifying the contribution of genetic variation to trait variation

Assume that we observe genotypes from N individuals at M SNPs. Typically, $M >> N$. Let $\boldsymbol{G}$ denote the matrix of genotypes on the traits measured. We define $\boldsymbol{X}$ to be the $N \times M$ matrix of standardized genotypes obtained by centering and scaling each column of $\boldsymbol{G}$ so that $\sum_n x_{n,m} = 0$ for all $m \in \{1, \ldots, M\}$. Let $\boldsymbol{y}$ denote the vector of phenotype of size N.

2.1.1 Linear Mixed Model

We assume the vector of phenotypes $\boldsymbol{y}$ is related to the genotypes by a linear mixed model (LMM):

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \epsilon & (2.1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_N) & (2.2) \\ \boldsymbol{\beta} &\sim \mathcal{N}(0, \frac{\sigma_g^2}{M} \boldsymbol{I}_M) & (2.3) \end{aligned}$$

Here $\boldsymbol{y}$ is centered so that $\sum_n y_n = 0$. σ_e^2 is the residual variance while σ_g^2 is the variance component corresponding to the M SNPs. The SNP heritability is defined as $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$.

In this model, we have $\mathbb{E}[\boldsymbol{y}] = 0$ while the population covariance of the phenotype vector $\boldsymbol{y}$ is:

$$\begin{aligned} cov(\boldsymbol{y}) &= \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\right] - \mathbb{E}[\boldsymbol{y}]\,\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}} \\ &= \sigma_g^2 \frac{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}}{M} + \sigma_e^2 \boldsymbol{I}_N \\ &= \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I}_N \qquad (2.4) \end{aligned}$$

Here $\boldsymbol{K} = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ is the genetic relatedness matrix (GRM) computed from all SNPs.

2.2 RHE-reg: A scalable estimator of heritability

2.2.1 Method of Moments

The method of moments principle obtains estimates of the model parameters such that the theoretical moments match the sample moments. In our model, the first theoretical moment, $\mathbb{E}[\boldsymbol{y}]$, is $\mathbf{0}$ by definition while the corresponding sample moment is also zero since we standardized the phenotypes. The second sample moment is $\boldsymbol{y}\boldsymbol{y}^T$ and the second theoretical moment is $cov(\boldsymbol{y}) = \sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}_N$. Thus, the method-of-moments (MoM) estimator of (σ_g^2, σ_e^2) is obtained by searching for values of σ_g^2, σ_e^2 such that the sample and theoretical moments are close, *i.e.*, by solving an ordinary least squares (OLS) problem:

$$(\widehat{\sigma_g^2}, \widehat{\sigma_e^2}) = argmin_{\sigma_g^2, \sigma_e^2} ||\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I})||_F^2 \tag{2.5}$$

Since the Frobenius norm of a matrix $\boldsymbol{A}$, $||\boldsymbol{A}||_F = \sqrt{\text{tr}\left[\boldsymbol{A}\boldsymbol{A}^T\right]}$, the OLS problem can be re-written as:

$$(\widehat{\sigma_g^2}, \widehat{\sigma_e^2}) = argmin_{\sigma_g^2, \sigma_e^2} \text{tr}\left[(\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}))(\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}))^T\right] \tag{2.6}$$

.

The MoM estimator satisfies the normal equations:

$$\begin{bmatrix} \text{tr}\left[\boldsymbol{K}^2\right] & \text{tr}\left[\boldsymbol{K}\right] \\ \text{tr}\left[\boldsymbol{K}\right] & N \end{bmatrix} \begin{bmatrix} \widehat{\sigma_g^2} \\ \widehat{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{2.7}$$

Solving the normal equations requires computing $\text{tr}\left[\boldsymbol{K}^2\right] = \sum_{i,j} K_{i,j}^2$, $\text{tr}\left[\boldsymbol{K}\right] = \sum_i K_{i,i}$, $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} = \sum_{i,j} K_{i,j}y_iy_j$ and $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = \sum_{n=1}^N y_n^2$. The GRM $\boldsymbol{K}$ can be computed in time $\mathcal{O}(MN^2)$ and requires $\mathcal{O}(N^2)$ memory. Given the GRM, computing each of

the coefficients for the normal equation requires $\mathcal{O}(N^2)$ time. Finally, given each of the coefficients, we can solve analytically solve for the $\widehat{\sigma_g^2}$ and $\widehat{\sigma_e^2}$. Indeed, we can write

$$\widehat{\sigma_g^2} = \frac{\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{y}}{\mathrm{tr}\left[\boldsymbol{K}^2\right] - N} \tag{2.8}$$

Thus, the key bottleneck in solving the HE regression lies in computing the GRM.

2.2.2 RHE-reg

Given that $\boldsymbol{K} = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, we can compute the quantities $\mathrm{tr}\left[\boldsymbol{K}\right] = \frac{1}{M}\sum_{i,j} X_{i,j}^2$, $\boldsymbol{w} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, $\mathrm{tr}\left[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}\right] = \frac{1}{m}\sum_{m=1}^{M} w_m^2$. For standardized genotypes, $\mathrm{tr}\left[\boldsymbol{K}\right] = N$ while $\mathrm{tr}\left[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}\right]$ can be computed in $\mathcal{O}(MN)$ time.

The one remaining quantity that we need to compute efficiently is $\mathrm{tr}\left[\boldsymbol{K}^2\right]$. Given a $N \times N$ matrix $\boldsymbol{A}$ and a random vector $\boldsymbol{z}$ with mean zero and covariance $\boldsymbol{I}_N$, we use the following identity to construct a randomized estimator of the trace of matrix $\boldsymbol{A}$ (see Appendix B.1 for a proof):

$$\mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}\right] = \mathrm{tr}\left[\boldsymbol{A}\right] \tag{2.9}$$

Equation 2.9 leads to the following unbiased estimator of the trace of $\boldsymbol{K}^2$ given B random vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$, drawn independently from a distribution with zero mean and identity covariance matrix $\boldsymbol{I}_N$:

$$\begin{aligned} L_B \equiv \widehat{\mathrm{tr}\left[\boldsymbol{K}^2\right]} &= \frac{1}{B}\sum_b \boldsymbol{z}_b^{\mathrm{T}}\boldsymbol{K}\boldsymbol{K}\boldsymbol{z}_b \\ &= \frac{1}{B}\frac{1}{M^2}\sum_b \boldsymbol{z}_b^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}_b \\ &= \frac{1}{B}\frac{1}{M^2}\sum_b \left\|\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}_b\right\|_2^2 \end{aligned} \tag{2.10}$$

In practice, we draw each entry of $\boldsymbol{z}$ independently from a standard normal distribution. We note that the estimator L_B involves two matrix-vector multiplications of $N \times M$ matrix repeated B times for a total runtime of $\mathcal{O}(NMB)$.

The RHE-reg estimator $(\widetilde{\sigma_g^2}, \widetilde{\sigma_e^2})$ is obtained by solving the Normal equations (Equation 2.7) by replacing $\mathrm{tr}\,[K^2]$ with L_B.

$$\begin{bmatrix} L_B & \mathrm{tr}\,[\boldsymbol{K}] \\ \mathrm{tr}\,[\boldsymbol{K}] & N \end{bmatrix} \begin{bmatrix} \widetilde{\sigma_g^2} \\ \widetilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{2.11}$$

The RHE-reg estimator of the SNP heritability is then obtained by $h^2_{rhe} = \frac{\widetilde{\sigma^2}}{s_y^2}$ where $s_y^2 = \frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{N-1}$ is the unbiased estimator of the phenotypic variance.

2.2.3 Sub-linear computations

The key bottleneck in the RHE-reg is the computation of L_B which involves repeated multiplication of the normalized genotype matrix $\boldsymbol{X}$ by a real-valued vector. Leveraging the fact that each element of the genotype matrix $\boldsymbol{G}$ takes values in the set $\{0, 1, 2\}$, we can improve the complexity of these multiplication operations from $\mathcal{O}(NM)$ to $\mathcal{O}(\frac{NM}{\max(\log_3 N, \log_3 M)})$ using the Mailman Algorithm [30].

2.2.3.1 The Mailman Algorithm

Consider a $M \times N$ matrix $\boldsymbol{A}^T$ whose entries take values in $\{0, 1, 2\}$. Assume that the number of SNPs $M = \log_3(N)$. The naive way to compute the product $\boldsymbol{A}^T\boldsymbol{b}$ for any real-valued vector $\boldsymbol{b}$ takes $O(log_3(N) * N)$ time.

The mailman algorithm decomposes the matrix $\boldsymbol{A}$ as $\boldsymbol{A}^T = \boldsymbol{U}_n\boldsymbol{P}$. atU_n is a $log_3(N) \times N$ matrix whose column contains all possible vectors over $\{0, 1, 2\}$ of length

$log_3(N)$. And P is an indicator matrix, where entry $P_{i,j} = 1$ if the i^{th} column is the same as j^{th} column in matrix $\boldsymbol{A} : \boldsymbol{A}^{(j)} = \boldsymbol{U}_n^{(i)}$. The decomposition of matrix $\boldsymbol{A}$ takes $\mathcal{O}(N \log_3(N))$ time. The desired product $\boldsymbol{A}^T\boldsymbol{b}$ is computed in two steps as $\boldsymbol{c} = \boldsymbol{P}\boldsymbol{b}$ followed by $\boldsymbol{U}_n\boldsymbol{c}$, each of which can be computed in only $\mathcal{O}(N)$ operations [30].

For a matrix $\boldsymbol{A}^{\mathrm{T}}$ with $M > \lceil \log_3(N) \rceil$, we partition $\boldsymbol{A}^{\mathrm{T}}$ into $\lceil \frac{M}{\lceil \log_3(N) \rceil} \rceil$ submatrices each of size $\lceil log_3(N) \times N \rceil$ each of which can be multiplied in time $\mathcal{O}(N)$ for a total computational cost of $\mathcal{O}(\frac{NM}{\log_3(N)})$.

2.2.3.2 Application of the Mailman algorithm to RHE-reg

Now consider the standardized genotype $\boldsymbol{X}$, which could be written as $\boldsymbol{X} = (\boldsymbol{G} - \boldsymbol{M})\boldsymbol{\Sigma}$, where $\boldsymbol{M}$ is a matrix where the i^{th} column contains the sample mean of the i^{th} SNP ($\boldsymbol{M} = \boldsymbol{1}_N\overline{\boldsymbol{g}}^{\mathrm{T}}$), and $\boldsymbol{\Sigma}$ is an $M \times M$ diagonal matrix, with the inverse of variance of each SNP as the diagonal entries.

Thus, when we compute $\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} = \frac{1}{M}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{M}\|\boldsymbol{\Sigma}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{M}^{\mathrm{T}}\boldsymbol{y})\|_2{}^2$ in Equation 2.15, computing $\boldsymbol{G}^T\boldsymbol{y}$ using the Mailman algorithm takes $\mathcal{O}(\frac{NM}{max(log_3M, log_3N)})$ operations. Similarly, to compute each term in the sum of the randomized estimator of $\mathrm{tr}\left[\boldsymbol{K}^2\right]$ (Equation 2.10), $\boldsymbol{X}^T\boldsymbol{z_b}$, we can substitute $\boldsymbol{X}^T\boldsymbol{z_b}$ with $\boldsymbol{\Sigma}\boldsymbol{G}^T\boldsymbol{z_b} - \boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{z_b}$. The first term $\boldsymbol{\Sigma}\boldsymbol{G}^T\boldsymbol{z_b}$ can again be computed using $\mathcal{O}(\frac{NM}{max(log_3M, log_3N)})$ using the Mailman algorithm, and the second term $\boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{z_b}$ is equivalent to scaling the N-vector $\boldsymbol{z_b}$ which can be computed in time $\mathcal{O}(N + M)$.

2.2.4 Computing the Standard Error

We show in the Appendix (Section C.2) that the variance of the RHE-reg estimator of σ_g^2 can be approximated by the variance of the exact HE-regression estimator with

an additional contribution due to the randomization:

$$\text{Var}\left[\widetilde{\sigma_g^2}\right] \approx \text{Var}\left[\widehat{\sigma_g^2}\right] + \frac{1}{BN^2\left(\frac{\text{tr}[K^2]}{N} - 1\right)^2}\left(\sigma_g^2\right)^2 \text{Var}\left[\boldsymbol{z}^{\text{T}}\boldsymbol{K}^2\boldsymbol{z}\right] \tag{2.12}$$

Here B is the number of samples used and $\boldsymbol{z}$ is a random vector with mean zero and identity covariance matrix. For samples with low-levels of relatedness, we can assume $\boldsymbol{K} \approx \boldsymbol{I}$ and our estimates of σ_g^2 and $\text{tr}\left[\boldsymbol{K}^2\right]$ to estimate the variance.

2.2.5 Some remarks on the RHE-reg estimator

1. Equation 2.3 assumes an infinitesimal model for the phenotype. However, all our results only depend on the second moment of the SNP effect sizes. Thus, the RHE-reg estimator can yield valid estimates for non-infinitesimal architectures.

2. In a number of settings, it is desirable to include covariates, such as age or sex, in the analysis. This changes the model in Equation 2.3 to:

$$\boldsymbol{y}| = \boldsymbol{W}\alpha + \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{2.13}$$

   Here $\boldsymbol{W}$ is a $N \times C$ matrix of covariates while α is a C-vector of coefficients. In this setting, we transform Equation 2.13 by multiplying by the projection matrix $\boldsymbol{V} = \boldsymbol{I}_N - \boldsymbol{W}(\boldsymbol{W}\boldsymbol{W}^{\text{T}})^{-1}\boldsymbol{W}^{\text{T}}$:

$$\boldsymbol{V}\boldsymbol{y} = \boldsymbol{V}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}\epsilon \tag{2.14}$$

   The RHE regression estimator applied to Equation 2.14 then must satisfy the

following moment conditions:

$$\begin{bmatrix} J_B & \operatorname{tr}[\boldsymbol{V}\boldsymbol{K}] \\ \operatorname{tr}[\boldsymbol{V}\boldsymbol{K}] & N-C \end{bmatrix} \begin{bmatrix} \widetilde{\sigma}_g^2 \\ \widetilde{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{K}\boldsymbol{V}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{y} \end{bmatrix} \tag{2.15}$$

Here J_B is a randomized estimator of $\operatorname{tr}[\boldsymbol{V}\boldsymbol{K}\boldsymbol{V}\boldsymbol{K}]$ analogous to Equation 2.10. The cost of computing the RHE-reg estimator now includes the cost of computing the inverse of $\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}$ as well as multiplying $\boldsymbol{W}$ by a real-valued vector for an added computational cost of $\mathcal{O}(C^3 + NC)$. Typically, the number of covariates C is small (tens to hundreds) so that the presence of covariance does not greatly increase the computational burden.

3. The variance components model (Equation 2.3 and 2.4) can be extended in a straightforward manner to more than two variance components. A number of recent studies have explored the utility of these models to partition heritability based on functional annotations as well as other categories. Such extensions have been considered in recent work [31].

2.3 Experiments

We performed simulations to measure the performance of RHE-reg to other methods for heritability estimation in terms of accuracy, running time and memory usage. We compared RHE-reg to two methods for computing REML estimates: GCTA [20] and BOLT-REML [5] as well as implementations of Haseman-Elston regression.

2.3.1 Accuracy and robustness

In our first set of simulations, we compared the accuracy of RHE-reg to our implementation of exact HE-regression as well as GCTA, an implementation that computes the REML. We simulated genotypes assuming each SNP is drawn independently from a Binomial distribution with allele frequency that is sampled uniformly from the interval $(0, 1)$. Given the genotypes, we simulated phenotypes under an infinitesimal model,*i.e.*, with effect size at each SNP drawn independently from a normal distribution with mean zero and variance equal to the heritability divided by the number of SNPs. We considered different values for the true SNP heritability of the phenotype to be 0.2, 0.5, and 0.8.

In our first series of experiments, we fixed the number of SNPs at $M = 10,000$ and varied the number of individuals $N = 1k, 2k \ldots 10k$. In the second series of experiments, we varied the number of SNPs $M = 1k, 2k \ldots 10k$ while fixing the number of individuals to be $N = 10,000$. We repeated each experiment 100 times in order to assess the variance of each of the estimators. We estimated heritability using RHE-reg with $B = 100$ random vectors.

Figure 2.1 compares the estimates of each of the three methods (RHE-reg, HE-regression, and GCTA) to the true heritability. First, we observe that all three methods obtain estimates of heritability that are quite close to each other as well as to the true heritability across the range of parameters explored. Second, RHE-reg and HE-regression are virtually indistinguishable in the variance of their estimates in each configuration. This suggests that the randomization makes a negligible contribution to the statistical accuracy of the MoM estimators. In some cases, RHE-reg even has a smaller variance than HE-regression. Thirdly, as expected, REML obtains estimators

that are closer to the true heritability compared to either of the MoM estimators for a high value of true heritability. For lower values of true heritability ($h^2 = 0.20, h^2 = 0.50$), the estimates from REML, HE-regression ,and RHE-reg are comparable. This result is also expected given that REML is asymptotically equivalent to MoM when the phenotypic correlation between individuals is small [32, 33]. Finally, the sample size has a bigger effect than the number of SNPs on the accuracy of each of the methods, consistent with theory.

Table 2.1: **The estimates of heritability from RHE-reg are consistent with those from GCTA and BOLT-REML on the NFBC data while RHE-reg is substantially faster**.

	Method					
	GCTA		BOLT-REML		RHE-reg	
	Runtime (min)	h_g^2 (se)	Runtime (min)	h_g^2 (se)	Runtime (min)	h_g^2 (se)
TG	11.28	0.145 (0.051)	8.87	0.148 (0.051)	1.61	0.145 (0.052)
HDL	10.81	0.325 (0.051)	9.72	0.326 (0.051)	1.30	0.349 (0.052)
BMI	10.85	0.237 (0.051)	9.29	0.235 (0.051)	1.29	0.200 (0.052)

Table 2.2: **Understanding the computational efficiency of RHE-reg**

	Runtime (min)	No Mailman (min)	No randomized trace estimate (min)
TG	1.61	3.70	38.5
HDL	1.30	2.60	36.2
BMI	1.29	2.68	36.7

2.3.2 Scalability

In the second set of simulations, we compared the runtime and memory usage of different methods. We compared RHE-reg to two REML methods, GCTA [20] and BOLT-REML [22] (a computationally efficient approximate method to compute the REML) as well as an exact MoM method MMHE [34]. In this experiment, we simulated genotype data consisting of $100,000$ SNPs over sample sizes of $N = 10k, 20k, 30k, 50k, 100k$ and $500k$ and then simulated phenotypes corresponding to the genotype data. For each dataset, we ran RHE-reg with $B = 100$ random vectors. We performed all comparisons on an Intel(R) Xeon(R) CPU 2.10GHz server with 128 GB RAM. All computations were restricted to a single core, capped to a maximum runtime of 12 hours and a maximum memory of 128 GB.

Figure 2.2 shows that both GCTA and MMHE do not scale to large sample sizes due to the requirement of computing and operating on a genetic relatedness matrix (GRM) that scales quadratically with N. GCTA could not complete its computation when running on $N = 100K$ individuals while MMHE did not complete its computation on $N = 50K$. BOLT-REML and RHE-reg scale linearly with sample size.

However, RHE-reg is an order of magnitude faster than BOLT-REML. For example, on a dataset of a size of $500K$ individuals, RHE-reg computed the heritability in about 30 minutes compared to 400 minutes for BOLT-REML. Figure 2.2 shows that RHE-reg is memory efficient as well.

2.3.3 Understanding the computational efficiency of RHE-reg

Our implementation of RHE-reg relies on two ideas to obtain computational efficiency: i) the use of a randomized estimator of the trace, and ii) the Mailman algorithm for fast matrix-vector multiplication. To explore the contribution of each of these ideas, we compared the runtimes of a MoM estimator with no randomization (HE-reg), RHE-reg using standard matrix-vector multiplication and RHE-reg using the Mailman algorithm. Table 2.2 shows the runtimes of each of these variants on the NFBC data. We see that the biggest runtime gain arises from applying the randomized estimator (faster by a factor of 10-12 relative to HE-reg) while the application of the Mailman algorithm reduces the runtime further by a factor of 2 (Table 1).

2.3.4 Accuracy of RHE-reg as a function of the number of random vectors

To explore the impact of the choice of the number of random vectors B on the accuracy of RHE-reg, we compared the heritability estimates of RHE-reg to those obtained from GCTA for the triglyceride (TG) phenotype as a function of B. We find good concordance between the estimates from RHE-reg and GCTA even for values of B as low as 10 suggesting that RHE-reg could be even faster in practice

with little loss in accuracy (Figure 2.3). In practice, the randomized estimator of trace is sufficient with random vectors of 10. In later discussion of this thesis, the default choice of random vector is 10 unless otherwise mentioned.

2.4 Application to NFBC data

We compared the statistical accuracy and runtime of BOLT-REML, GCTA, and RHE-reg on the Northern Finland Birth Cohort (NFBC) dataset. The NFBC dataset contains $315,529$ SNPs and $5,326$ individuals after applying standard filters (minor allele frequency > 0.05 and Hardy-Weinberg Equilibrium p-value < 0.01) [35]. We applied these methods to estimate the heritability of three phenotypes that were assayed in this dataset: triglycerides (TG), high-density lipoprotein (HDL) and body mass index (BMI).

We compared the runtime, point estimates of the heritability as well as standard errors for each of the three methods. We computed RHE-reg with $B = 100$ random vectors. As shown in the Table 2.1, the heritability estimates of RHE-reg are concordant with the other methods while being an order of magnitude faster to compute. We note that the NFBC dataset has a sample size $N \approx 5,000$ so that we expect RHE-reg to be more accurate on larger datasets.

2.5 Application to UK Biobank

We applied RHE-reg to the estimate the SNP heritability associated with SNPs genotypes on the UK Biobank Axiom array. The detailed quality control is described in Section A.2. After quality control, we obtained $291,273$ individuals and $494,207$

SNPs. We computed the heritability of 40 traits of choice (Section A.1). In Table 2.3, we report the estimated heritability and standard error. Notice that the reported heritability are on the observed scale for binary traits.

Trait	h_g^2	se
HbA1C	0.193	0.012
T1D	0.003	0.002
T2D	0.015	0.002
AgeFinishEducation	0.079	0.004
AgeMenarch	0.271	0.013
DurationOfWalk	0.054	0.003
GettingUpTime	0.099	0.004
MorningEveningPerson	0.061	0.004
SleepDuration	0.073	0.004
AlcoholIntake	0.114	0.005
EverSmoked	0.099	0.004
Former/CurrentSmoker	0.131	0.005
TownsendIndex	0.051	0.003
Angina	0.041	0.003
HeartAttack	0.029	0.002
Asthma	0.085	0.008
CrohnsDisease	0.0088	0.002
Eczema	0.062	0.003
RheumatoidArathritis	0.007	0.002
UlcerativeColitis	0.017	0.002
Bipolar	0.006	0.002
Depression	0.033	0.002

Height	0.721	0.017
BMI2010	0.329	0.016
BMI	0.329	0.016
Weight	0.348	0.013
Body Fat Percentage	0.309	0.006
Trunk Fat	0.291	0.009
Hip Circumference	0.283	0.014
Waist Circumference	0.272	0.011
Diastolic Blood Pressure	0.183	0.007
Systolic Blood Pressure	0.191	0.007
High Blood Pressure	0.140	0.006
Hypertension	0.173	0.007
Pulse Rate	0.195	0.012
High Cholesterol	0.105	0.011
Total Cholesterol	0.237	0.048
LDL	0.248	0.069
HDL	0.459	0.091
Triglycerides	0.304	0.049

Table 2.3: **Heritability and standard error of 40 traits in UK Biobank**.

2.6 Discussion

In this chapter, we proposed a scalable estimator of heritability using a randomized version of the Haseman-Elston regression (RHE-reg). The RHE-reg estimator is based on performing a small number of multiplications of the genotype matrix with

random vectors with mean zero and identity covariance. Using the properties of the genotype matrix, we can compute this estimator using the Mailman algorithm in $\mathcal{O}(\frac{NMB}{max(\log_3 N, \log_3 M)})$ time on a dataset containing N individuals, M SNPs and with a small number of B random vectors. We show that this estimator achieves similar accuracy as REML-based methods on both simulated and real data. RHE-reg can be effectively applied to whole-genome genotype data of hundreds of thousands of individuals for rapid variance components estimation.

The model in Equation 2.3 can be extended into a multi-component model described in detail in [31].

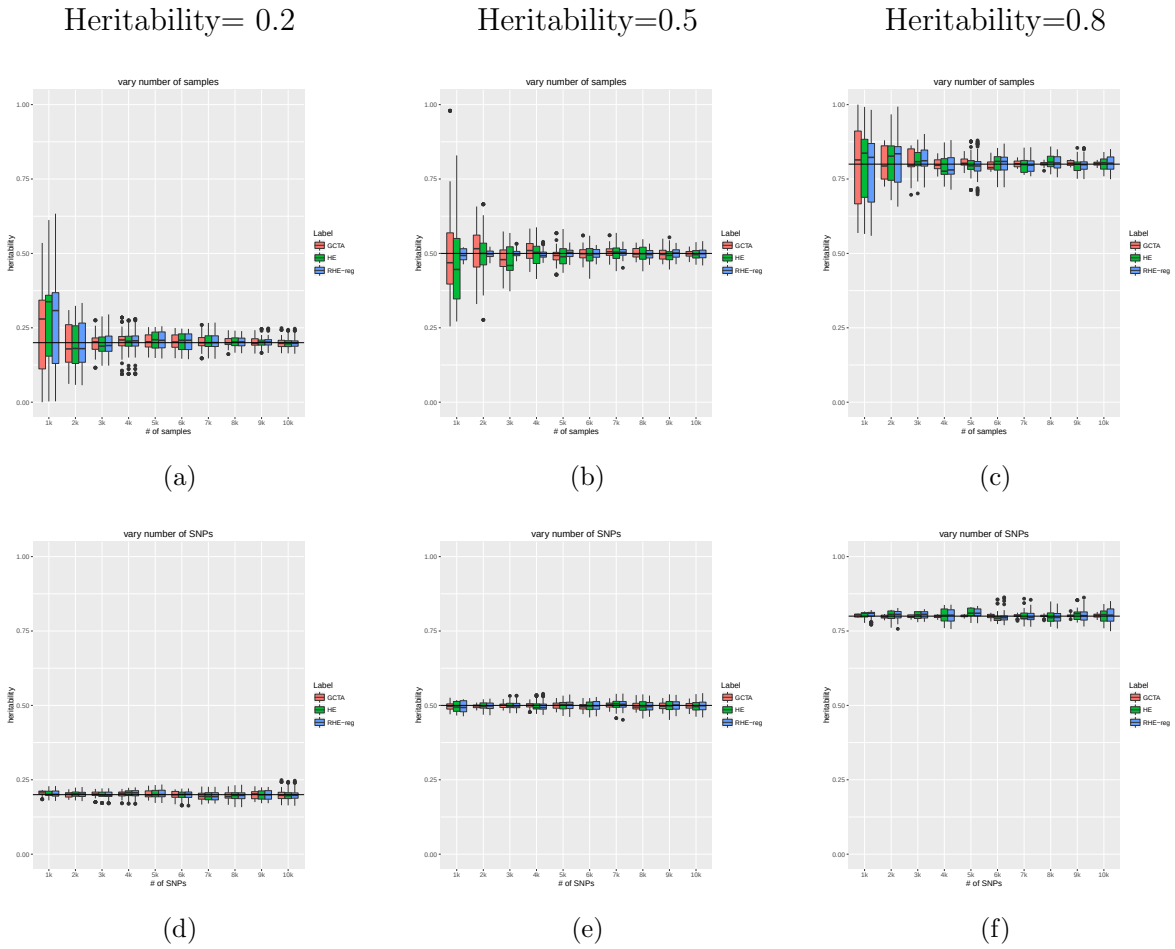

Figure 2.1: **RHE-reg accurately estimates heritability**: In the first series of figures, we fixed the number of SNPs to 10000 and varied the sample size. In Figures (a-c), we fixed the true heritability to 0.2, 0.5 and 0.8 respectively. In the second series of figure, we fixed the number of samples to 10000 and varies the number of SNPs. HE and RHE-reg are indistinguishable. Comparing to GCTA which is a REML method, MoM methods perform better when heritability is smaller.

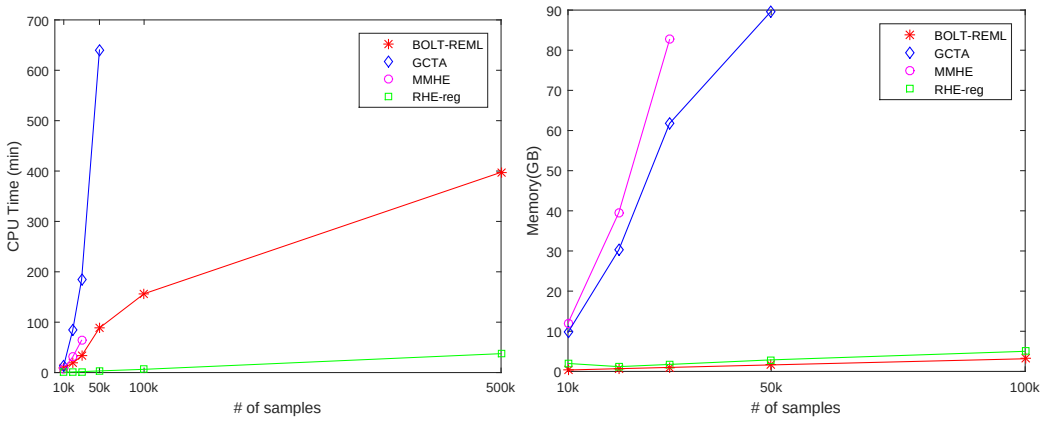

Figure 2.2: **RHE-reg is efficient**: In both figures, we fixed the number of SNPs to 100,000, and varied the number of samples and compare run time and memory usage. In the first figure, GCTA did not finish computation on 100K samples. For MMHE (an exact MoM method), the computation was stopped at a sample size of 50k due to memory constraints. BOLT-REML scales linearly while RHE-reg is significantly faster.

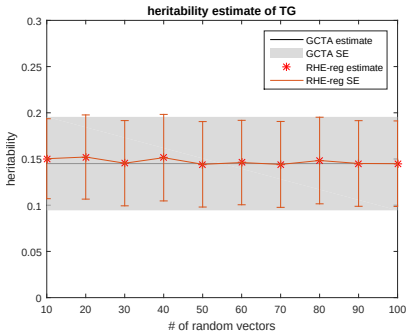

Figure 2.3: **Impact of the number of random vectors on accuracy of RHE-reg**: We ran RHE-reg with different number of random vectors B, and compared the point estimate and standard error to GCTA. The gray area indicates the standard error computed by GCTA. The RHE-reg estimates converge with increasing number of random vectors though even 10 random vectors are adequate for accurate estimation.

CHAPTER 3

Fast estimation of genetic correlation for biobank-scale data

In Chapter 2, we considered efficient methods for estimating SNP heritability. In this chapter, we will consider models of how genetic effects are shared across traits.

Genetic correlation is an important parameter that quantifies the genetic basis that is shared across two traits. Estimates of genetic correlation can reveal pleiotropy, uncover novel biological pathways underlying diseases, and improve the accuracy of genetic prediction [7].

While traditionally reliant on family studies, the availability of genome-wide genetic data has led to several approaches to estimate genetic correlation from these datasets [7]. An important class of methods for estimating genetic correlation relies on computing the restricted maximum likelihood within a bi-variate linear mixed model (LMM), termed genomic restricted maximum likelihood (GREML)[36, 25, 5, 37]. However, current GREML methods are computationally expensive to be applied to large-scale datasets such as the UK Biobank [38].

While GREML methods need individual-level data, several methods [10, 39, 40, 41, 42, 43], such as LD-score regression (LDSC) [10], have been proposed for estimating genetic correlation using GWAS summary statistics. While methods such

as LDSC often have substantially reduced computational requirements relative to GREML, LDSC estimates tend to have large standard errors which increase further when there is a mismatch between the samples used to estimate summary statistics and the reference datasets that are used to estimate LD scores [44]. High-definition likelihood (HDL) [43], a more recent summary-statistic based method, has been shown to be more precise relative to LDSC. HDL, however, requires computing a singular-value decomposition (SVD) of the LD matrix which increases its runtime. Further, recent studies [45, 46] have shown that the accuracy of genetic correlation estimates can deteriorate when there is a mismatch between reference and sample data. Thus, it is critical to develop methods for estimating genetic correlation that can work directly with large individual-level datasets.

In this chapter, we propose, SCORE (SCalable genetic cORrelation Estimator), a randomized Method-of-Moments (MoM) estimator of genetic correlations among traits using individual genotypes that can scale to the dataset sizes typical of the UK Biobank. While SCORE can estimate the heritability of traits as well as the genetic correlation between pairs of traits, we focus on the problem of estimating genetic correlation in this work. SCORE avoids the explicit computation of the genetic relationship matrix (GRM). Instead, we show that the genetic correlation can be computed using a *sketch* of the genotype matrix, *i.e.*, by multiplying the genotype matrix with a small number of random vectors.

In simulations, we show that SCORE yields accurate estimates of genetic correlation across a range of genetic architectures (with varying heritability, genetic correlation, and polygenicity). Relative to summary-statistic methods that can be applied to Biobank-scale data, SCORE obtains a reduction in the standard error of 44% relative to LDSC and 20% relative to HDL (averaged across all simulations).

Further, SCORE can estimate genetic correlation on $\approx$ 500K SNPs in $\approx$ 300K unrelated white British individuals in a few hours, orders of magnitude faster than methods that rely on individual data (GCTA-GREML and GCTA-HE). Analyzing 780 pairs of traits in $291,273$ unrelated white British individuals in the UK Biobank, the estimates of genetic correlation at $454,207$ common SNPs obtained by SCORE are largely concordant with those from LDSC (Pearson correlation $r = 0.95$). While 245 pairs of traits are identified to have significant genetic correlation by both methods (using a Bonferroni correction for the number of pairs of traits tested), the reduced standard error of estimates from SCORE leads to the discovery of the significant genetic correlations between additional 200 pairs of traits relative to LDSC. Finally, SCORE detects a significant positive correlation between serum liver enzyme levels (alanine (ALT) and aspartate aminotransferase (AST)) and coronary artery disease related traits (angina and heart attack) suggesting that coronary artery disease and liver dysfunction harbor a shared genetic component.

3.1 Statistical Models

We first define the generative model for a pair of phenotypes, the Bi-variate linear mixed model. Then we define the generative model over multiple phenotypes.

3.1.1 The Bi-variate Linear Mixed Model (LMM)

We describe our model in the general setting, where the traits are not observed on the same set of individuals. Assume we have N_1 individuals for trait 1 and N_2 individuals for trait 2 of which N individuals ($N \leq N_1, N \leq N_2$) contain measurements for both the traits. We have defined $\boldsymbol{X}_1, \boldsymbol{X}_2$ to be the $N_1 \times M$ and $N_2 \times M$ matrices

of standardized genotypes obtained by centering and scaling each column of the unstandardized genotype matrices $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ so that $\sum_n x_{t,n,m} = 0$ for all $m \in \{1, \ldots, M\}, t \in \{1, 2\}$. Let $\boldsymbol{y}_1, \boldsymbol{y}_2$ denote the two vectors of phenotypes with size N_1 and N_2 respectively. Additionally, we define an $N_1 \times N_2$ indicator matrix, $\boldsymbol{C}$ where $\boldsymbol{C}_{i,j} = 1$ when individual i among samples measured for the first phenotype and j in samples measured for the second phenotype refers to the same individual and 0 otherwise. We define $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ to be vectors of SNP effect sizes of length M.

We assume the following model relating a pair of traits $\boldsymbol{y_1}, \boldsymbol{y_2}$:

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\ \boldsymbol{y}_2 &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2 \end{aligned} \tag{3.1}$$

For the SNP effects, we assume $\mathbb{E}[\boldsymbol{\beta}_1] = 0, \mathbb{E}[\boldsymbol{\beta}_2] = 0$ and:

$$\begin{aligned} cov(\boldsymbol{\beta}_1, \boldsymbol{\beta}_1) &= \frac{1}{M}\sigma^2_{g1}\boldsymbol{I}_M \\ cov(\boldsymbol{\beta}_2, \boldsymbol{\beta}_2) &= \frac{1}{M}\sigma^2_{g2}\boldsymbol{I}_M \\ cov(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= \frac{1}{M}\gamma_g\boldsymbol{I}_M \end{aligned} \tag{3.2}$$

Here $\boldsymbol{I}_M$ is an $M \times M$ identity matrix, σ^2_{gt} denotes the genetic variance associated with trait t ($t \in \{1, 2\}$), and γ_g denotes the genetic covariance. For the environmental effects, we assume $\mathbb{E}[\boldsymbol{\epsilon}_1] = 0, \mathbb{E}[\boldsymbol{\epsilon}_2] = 0$ and:

$$\begin{aligned} cov(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_1) &= \sigma^2_{e1}\boldsymbol{I}_N \\ cov(\boldsymbol{\epsilon}_2, \boldsymbol{\epsilon}_2) &= \sigma^2_{e2}\boldsymbol{I}_N \\ cov(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2) &= \gamma_e\boldsymbol{C} \end{aligned} \tag{3.3}$$

The genetic correlation parameter ρ_g is defined as $\rho_g \equiv \frac{\gamma_g}{\sqrt{\sigma^2_{g1}}\sqrt{\sigma^2_{g2}}}$. Importantly, SCORE does not make additional assumptions on the distribution of the genetic effect sizes or the environmental noise.

3.1.2 Multivariate Linear Mixed Model

In a more general case, we define the joint model for multiple traits. We assume that we have K phenotypes, and thus let $\boldsymbol{\beta}$ be the $M \times K$ effect size matrix, and each phenotype has the generative model:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{\epsilon_i} \tag{3.4}$$

Where for each phenotype i, we have N_i samples. $N_{i,j}$ are the number of samples that contain measurements for both traits $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ ($N \leq N_i, N \leq N_j$). For each phenotype, we define $\boldsymbol{X}_i$ be the corresponding $N_i \times M$ matrix of standardized genotypes. $\boldsymbol{\beta_i}$ is the i^{th} column of $\boldsymbol{\beta}$, which is the vector of SNP effect sizes for phenotype i. $\boldsymbol{\epsilon_i}$ denotes trait-specific environmental noise that is independent of the genetic effect.

We assume the mean of $\boldsymbol{\beta}$ is 0, and assume $\boldsymbol{\beta}$ follows the matrix normal distribution:

$$\boldsymbol{\beta} \sim \mathcal{MN}(0, diag(\mathbf{1}_M), \boldsymbol{V}) \tag{3.5}$$

where $\boldsymbol{V}$ is a $K \times K$ matrix :

$$\boldsymbol{V}(i,j) = \begin{cases} \sigma^2_{g,i} & \text{if } i = j \\ \gamma_{g,ij} & \text{otherwise} \end{cases} \tag{3.6}$$

Where $\sigma^2_{g,i}$ is the genetic variance associated with phenotype i, and $\gamma_{g,ij}$ denotes the genetic covariance between phenotype i and j. Thus the genetic correlation between phenotype i and j is defined as: $\rho_{g,ij} = \frac{\gamma_{g,ij}}{\sqrt{\sigma^2_{g,i}}\sqrt{\sigma^2_{g,j}}}$.

Notice that $\boldsymbol{\beta}$ are row-wise independent and identically distributed.

3.2 SCORE: SCalable genetic cORrelation Estimator

3.2.1 Method of Moments for the Bi-variate LMM

SCORE uses a Method of Moments (MoM) estimator to estimate the parameters: $(\gamma_g, \gamma_e, \sigma^2_{g1}, \sigma^2_{g2}, \sigma^2_{e1}, \sigma^2_{e2})$.

Since the mean of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are zero, we focus on the covariance. The population covariance of the concatenated phenotypes $\boldsymbol{y} \equiv [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T]^T$ is now:

$$cov(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^T = \begin{bmatrix} \sigma^2_{g1}\boldsymbol{K}_1 & \gamma_g \boldsymbol{K}_A \\ \gamma_g \boldsymbol{K}_A^T & \sigma^2_{g2}\boldsymbol{K}_2 \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1} I_{N_1} & \gamma_e \boldsymbol{C} \\ \gamma_e \boldsymbol{C}^T & \sigma^2_{e2} I_{N_2} \end{bmatrix} \tag{3.7}$$

Here $\boldsymbol{K}_1 = \frac{\boldsymbol{X}_1 \boldsymbol{X}_1^T}{M}$ is the GRM for the samples observed for the first trait while $\boldsymbol{K}_2 = \frac{\boldsymbol{X}_2 \boldsymbol{X}_2^T}{M}$ is the GRM for the samples for the second trait and : $\boldsymbol{K}_A = \frac{\boldsymbol{X}_1 \boldsymbol{X}_2^T}{M}$ is the GRM for pairs of samples across traits.

The MoM estimator is obtained by minimizing the sum of squared differences between the population and empirical covariances:

$$\begin{aligned}(\widehat{\gamma_g}, \widehat{\gamma_e}, \widehat{\sigma^2_{g1}}, \widehat{\sigma^2_{g2}}, \widehat{\sigma^2_{e1}}, \widehat{\sigma^2_{e2}}) = argmin_{\gamma_g, \gamma_e, \sigma^2_{g1}, \sigma^2_{g2}, \sigma^2_{e1}, \sigma^2_{e2}} || \boldsymbol{y}\boldsymbol{y}^T - \\ \left(\begin{bmatrix} \sigma^2_{g1}\boldsymbol{K}_1 & \gamma_g \boldsymbol{K}_A \\ \gamma_g \boldsymbol{K}_A^T & \sigma^2_{g2}\boldsymbol{K}_2 \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1} \boldsymbol{I}_{N_1} & \gamma_e \boldsymbol{C} \\ \gamma_e \boldsymbol{C}^T & \sigma^2_{e2} \boldsymbol{I}_{N_2} \end{bmatrix} \right) ||^2_F \end{aligned} \tag{3.8}$$

The MoM estimator for the genetic covariance satisfies the normal equations:

$$\begin{bmatrix} tr(\boldsymbol{K}_A \boldsymbol{K}_A^{\mathrm{T}}) & tr(\boldsymbol{K}_C) \\ tr(\boldsymbol{K}_C) & N \end{bmatrix} \begin{bmatrix} \widehat{\gamma_g} \\ \widehat{\gamma_e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \boldsymbol{K}_A \boldsymbol{y}_2 \\ \boldsymbol{y}_1^T \boldsymbol{C} \boldsymbol{y}_2 \end{bmatrix} \tag{3.9}$$

where $\boldsymbol{K}_C = \frac{\boldsymbol{X}_1 \boldsymbol{X}_2^T \boldsymbol{C}^T}{M}$. Given the coefficients of the normal equations, we can solve analytically for $\widehat{\gamma_g}$, and $\widehat{\gamma_e}$.

Given MoM estimates of the variance components, the MoM estimate of the genetic correlation is given by the plug-in estimate:

$$\widehat{\rho_g} = \frac{\widehat{\gamma_g}}{\sqrt{\widehat{\sigma_{g1}^2}}\sqrt{\widehat{\sigma_{g2}^2}}} \tag{3.10}$$

3.2.2 Method of Moments for the multivariate LMM is equivalent to the bi-variate model applied to each pair of traits

As in section 3.1.2, we defined multivariate Model for multiple phenotypes. Recall we have $\boldsymbol{y}_i$ now be the i^{th} phenotype out of the total K phenotypes. Let $\boldsymbol{y} \equiv [\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_i^T, \ldots, \boldsymbol{y}_K^T]^T$, then the population covariance is now:

$$cov(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^T = \begin{bmatrix} \boldsymbol{V}(1,1)\boldsymbol{K}(1,1) & \ldots & \boldsymbol{V}(1,K)\boldsymbol{K}(1,K) \\ \ldots & \boldsymbol{V}(i,j)\boldsymbol{K}(i,j) & \ldots \\ \boldsymbol{V}(K,1)\boldsymbol{K}(K,1) & \ldots & \boldsymbol{V}(K,K)\boldsymbol{K}(K,K) \end{bmatrix} \tag{3.11}$$

$$+ \begin{bmatrix} \sigma_{e,1}^T \boldsymbol{I}_{N_1} & \ldots & \gamma_{e,1K}\boldsymbol{C}(1,K) \\ \ldots & \sigma_{e,i}^2 \boldsymbol{I}_{N_i} & \ldots \\ \gamma_{e,K1}\boldsymbol{C}(K,1) & \ldots & \sigma_{e,K}^2 \boldsymbol{I}_{N_K} \end{bmatrix} \tag{3.12}$$

Recall $\boldsymbol{V}$ is defined in section 3.1.2:

$$\boldsymbol{V}(i,j) = \begin{cases} \sigma_{g,i}^2 & \text{if } i = j \\ \gamma_{g,ij} & \text{otherwise} \end{cases} \tag{3.13}$$

Here $\boldsymbol{K}(i,j) = \frac{\boldsymbol{X_i X_j}^T}{M}$ is the genetic relatedness matrix (GRM) computed with all samples that have measurement on phenotype i and j. $\boldsymbol{C}(i,j)$ is an indicator matrix, where $\boldsymbol{C}(i,j)_{m,n} = 1$ if the m^{th} entry in $\boldsymbol{y}_i$ and n^{th} entry in $\boldsymbol{y}_j$ are measures on phenotype i and j for the same sample, and 0 otherwise. The MoM estimator is obtained by minimizing the sum of squared differences between the population and empirical covariance:

$$\{\widehat{\gamma_g}\} = argmin_{\{\gamma_g\}} ||\boldsymbol{y}\boldsymbol{y}^T - cov(\boldsymbol{y})||_F^2 \tag{3.14}$$

with $cov(\boldsymbol{y})$ defined in equation 3.12. Thus the MoM estimator for Multivariate LMM satisfies the normal equations:

$$\begin{bmatrix} tr(\boldsymbol{K}(1,2)^2) & \mathbf{0} & \dots & tr(\boldsymbol{K}(1,2)) & \mathbf{0} & \dots \\ \mathbf{0} & tr(\boldsymbol{K}(i,j)^2) & \dots & \mathbf{0} & tr(\boldsymbol{K}(i,j)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & tr(\boldsymbol{K}(K-1,K)^2) & \mathbf{0} & \dots & tr(\boldsymbol{K}(K-1,K)) \\ tr(\boldsymbol{K}(1,2)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(1,2)) & \dots & \mathbf{0} \\ \mathbf{0} & tr(\boldsymbol{K}(i,j)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(i,j)) & \mathbf{0} \\ \mathbf{0} & \dots & tr(\boldsymbol{K}(K-1,K)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(K-1,K)) \end{bmatrix} \begin{bmatrix} \widehat{\gamma_{g,12}} \\ \widehat{\gamma_{g,\boldsymbol{ij}}} \\ \gamma_{g,\widehat{(K-1)K}} \\ \widehat{\gamma_{e,12}} \\ \widehat{\gamma_{e,\boldsymbol{ij}}} \\ \gamma_{e,\widehat{(K-1)K}} \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{y}_1^T \boldsymbol{K}(1,2) \boldsymbol{y}_2 \\ \boldsymbol{y}_i^T \boldsymbol{K}(i,j) \boldsymbol{y}_j \\ \boldsymbol{y}_{K-1}^T \boldsymbol{K}(K-1,K) \boldsymbol{y}_K \\ \boldsymbol{y}_1^T \boldsymbol{C}(1,2) \boldsymbol{y}_2 \\ \boldsymbol{y}_i^T \boldsymbol{C}(i,j) \boldsymbol{y}_j \\ \boldsymbol{y}_{K-1}^T \boldsymbol{C}(K-1,K) \boldsymbol{y}_K \end{bmatrix} \tag{3.15}$$

where $\gamma_{g,ij}$ is a the genetic covariance between phenotype i and j. $tr(\boldsymbol{K}(i,j)^2)$ denotes $tr(\boldsymbol{K}(i,j)\boldsymbol{K}(i,j)^T)$. And $\boldsymbol{y}_i^T \boldsymbol{K}(i,j) \boldsymbol{y}_j = \frac{\boldsymbol{y}_i^T \boldsymbol{X} i \boldsymbol{X}_j^T \boldsymbol{y}_j}{M}$.

By observing the block-wise pattern of the normal equations, we can conclude the MoM estimator for the Multivariate LMM is equivalent to estimates obtained by applying the Bi-variate LMM to each pair of traits.

3.2.3 SCORE: SCalable genetic cORrelation Estimator

Naive computation of the MoM estimate of genetic covariance requires computing $tr(\boldsymbol{K}_A \boldsymbol{K}_A^{\mathrm{T}})$ which requires $\mathcal{O}(N_1 N_2 M)$ operations, where N_1, N_2 are the sample size

of each of the traits.

To overcome this computational bottleneck, we replace $tr(\boldsymbol{K}_A\boldsymbol{K}_A^{\mathrm{T}})$ with an unbiased randomized estimate: $tr(\widehat{\boldsymbol{K}_A\boldsymbol{K}_A^{\mathrm{T}}})$ [47].

Given B random vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$, $z_b \in \mathbb{R}^{N_2}, b \in 1 \ldots B$ drawn independently from a distribution with zero mean and identity covariance, our estimator is given by:

$$L_B = tr(\widehat{\boldsymbol{K}_A\boldsymbol{K}_A^{\mathrm{T}}}) = \frac{1}{B}\frac{1}{M^2}\sum_b ||\boldsymbol{X_1}\boldsymbol{X_2}^T\boldsymbol{z}_b||_2^2$$

The SCORE estimator $(\tilde{\gamma_g}, \tilde{\gamma_e})$ is obtained by solving Equation 3.9 by replacing $tr(\boldsymbol{K}_A\boldsymbol{K}_A^{\mathrm{T}})$ with L_B.

$$\begin{bmatrix} L_B & tr(\boldsymbol{K}_C) \\ tr(\boldsymbol{K}_C) & N \end{bmatrix} \begin{bmatrix} \tilde{\gamma_g} \\ \tilde{\gamma_e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T\boldsymbol{K}_A\boldsymbol{y}_2 \\ \boldsymbol{y}_1^T\boldsymbol{C}\boldsymbol{y}_2 \end{bmatrix}$$

Here $tr(\boldsymbol{K}_C)$ denotes the sum of the squared genotypes for individuals measured on both traits so that $tr(\boldsymbol{K}_C)$ can be computed in time:$\mathcal{O}(MN)$.

Computing L_B requires multiplying the genotype matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with B vectors resulting in a runtime of $\mathcal{O}(max(N_1, N_2)MB)$.

Leveraging the fact that each element of the genotype matrix takes values in the set $\{0, 1, 2\}$, L_B can be computed in time $\mathcal{O}(max(\frac{N_1}{max(log_3 N_1, log_3 M)}, \frac{N_2}{max(log_3 N_2, log_3 M)})MB)$ [30] (while the standardized genotypes are real-valued, SCORE computes the equivalent quantities by operating on the unstandardized genotype matrix to be able to leverage its discrete entries followed by subtracting the product of the mean of a SNP and random vectors and scaling by MAF). Combined with our previous efficient estimators of the genetic variance components [48, 31], we obtain an efficient estimator of ρ_g. SCORE can also handle fixed-effects covariates (Section C.1). Finally, we

obtain standard errors of the estimates of ρ_g using a block Jackknife [49] which can be computed with little additional computational overhead (Section C.2).

In the setting where the two traits are measured on the same set of individuals, we can estimate the ρ_g directly without the need for separately estimating γ_g, σ^2_{g1}, and σ^2_{g2}. This estimator does not rely on any randomized approximations and can be computed in time $\mathcal{O}\frac{NM}{max(\log_3 N, \log_3 M)}$. We term this modification $SCORE-OVERLAP$ in next section.

3.2.4 The scenario of completely overlapping samples

Here we describe the Bi-variate Linear Model in section 3.1.1 in the setting where the two traits are measured on the same set of individuals. Let $\boldsymbol{X}$ denote the genotype matrix for which both traits are observed. Denote the concatenated phenotype vector, $\boldsymbol{y} \equiv [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T]^T$, concatenated environmental effect vector $\boldsymbol{\epsilon} \equiv [\boldsymbol{\epsilon_1}^T, \boldsymbol{\epsilon_2}^T]^T$, and concatenated effect size vector$\boldsymbol{\beta} \equiv [\boldsymbol{\beta_1}^T, \boldsymbol{\beta_2}^T]^T$. In this setting, the modified generative model is:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{X} & 0 \\ 0 & \boldsymbol{X} \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.16}$$

In this model, the population covariance of the concatenated phenotype vector $\boldsymbol{y}$ is given by:

$$cov(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^T = \begin{bmatrix} \sigma^2_{g1}\boldsymbol{K} & \gamma_g \boldsymbol{K} \\ \gamma_g \boldsymbol{K}^T & \sigma^2_{g2}\boldsymbol{K} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1} I_N & \gamma_e \boldsymbol{I}_N \\ \gamma_e I_N & \sigma^2_{e2}\boldsymbol{I}_N \end{bmatrix} \tag{3.17}$$

Here $\boldsymbol{K} = \frac{\boldsymbol{X}\boldsymbol{X}^T}{M}$ is the genetic relatedness matrix (GRM). $\sigma^2_{gt}, \sigma^2_{et}$ denote the genetic and environmental variance components associated with trait t. Our approach to estimate both the variance components and the genetic correlation relies on a Method-

of-Moments (MoM) estimator obtained by equating the population covariance to the empirical covariance. The empirical covariance of the concatenated phenotype vector $\boldsymbol{y}$ is estimated by the sample covariance: $\boldsymbol{y}\boldsymbol{y}^T$. The MoM estimator is obtained by solving the following ordinary least squares problem:

$$(\widehat{\gamma_g}, \widehat{\gamma_e}, \widehat{\sigma^2_{g1}}, \widehat{\sigma^2_{g2}}, \widehat{\sigma^2_{e1}}, \widehat{\sigma^2_{e2}}) = argmin_{\gamma_g, \gamma_e, \sigma^2_{g1}, \sigma^2_{g2}, \sigma^2_{e1}, \sigma^2_{e2}} ||\boldsymbol{y}\boldsymbol{y}^T - \left(\begin{bmatrix} \sigma^2_{g1}\boldsymbol{K} & \gamma_g \boldsymbol{K} \\ \gamma_g \boldsymbol{K}^T & \sigma^2_{g2}\boldsymbol{K} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1}\boldsymbol{I}_N & \gamma_e \boldsymbol{I}_N \\ \gamma_e \boldsymbol{I}_N & \sigma^2_{e2}\boldsymbol{I}_N \end{bmatrix} \right)||^2_F \tag{3.18}$$

Setting the gradient of the objective function to zero gives us the normal equations. We observe that solving for the genetic and environmental covariance parameters (γ_g, γ_e) is decoupled from solving for the variance component parameters: $\sigma^2_{g1}, \sigma^2_{e1}, \sigma^2_{g2}, \sigma^2_{e2}$. Thus, MoM estimates of the covariance parameters can be obtained by solving the set of normal equations:

$$\begin{bmatrix} tr(\boldsymbol{K}^2) & tr(\boldsymbol{K}) \\ tr(\boldsymbol{K}) & N \end{bmatrix} \begin{bmatrix} \widehat{\gamma_g} \\ \widehat{\gamma_e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_2^T \boldsymbol{K} \boldsymbol{y}_1 \\ \boldsymbol{y}_2^T \boldsymbol{y}_1 \end{bmatrix} \tag{3.19}$$

The GRM $\boldsymbol{K}$ can be computed in time $\mathcal{O}(MN^2)$ and $\mathcal{O}(N^2)$ memory. Given the GRM, computing the coefficients for the normal equations requires $\mathcal{O}(N^2)$ time. Given each of the coefficients, we can solve analytically for $\widehat{\gamma_g}$, and $\widehat{\gamma_e}$:

$$\widehat{\gamma_g} = \frac{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_1^T \boldsymbol{y}_2}{tr[\boldsymbol{K}^2] - N}$$

Here we have used the property that $tr(\boldsymbol{K}) = N$ due to the use of a standardized genotype matrix.

Similarly, we can solve the following linear systems for the estimators of each of the genetic variance parameters:

$$\begin{bmatrix} tr(\boldsymbol{K}^2) & tr(\boldsymbol{K}) \\ tr(\boldsymbol{K}) & N \end{bmatrix} \begin{bmatrix} \widehat{\sigma^2_{g1}} & \widehat{\sigma^2_{g2}} \\ \widehat{\sigma^2_{e1}} & \widehat{\sigma^2_{e2}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_1 & \boldsymbol{y}_2^T \boldsymbol{K} \boldsymbol{y}_2 \\ \boldsymbol{y}_1^T \boldsymbol{y}_1 & \boldsymbol{y}_2^T \boldsymbol{y}_2 \end{bmatrix} \tag{3.20}$$

The estimators for σ_{g1}^2 and σ_{g2}^2 are given by $\widehat{\sigma_{g1}^2} = \frac{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_1 - \boldsymbol{y}_1^T \boldsymbol{y}_1}{tr[\boldsymbol{K}^2] - N}$ and $\widehat{\sigma_{g2}^2} = \frac{\boldsymbol{y}_2^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_2^T \boldsymbol{y}_2}{tr[\boldsymbol{K}^2] - N}$

Finally, we use estimates of the genetic variance parameters to obtain a plug-in estimate of the genetic correlation: $\widehat{\rho_g} = \frac{\widehat{\gamma_g}}{\sqrt{\widehat{\sigma_{g1}^2}}\sqrt{\widehat{\sigma_{g2}^2}}}$.

Substituting the expressions for the genetic covariance and variances and the GRM gives us the following estimator of genetic correlation:

$$\widehat{\rho_g} = \frac{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_1^T \boldsymbol{y}_2}{\sqrt{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_1 - \boldsymbol{y}_1^T \boldsymbol{y}_1}\sqrt{\boldsymbol{y}_2^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_2^T \boldsymbol{y}_2}} \tag{3.21}$$

Directly computing $\widehat{\rho_g}$ requires only computing $\boldsymbol{X}^T \boldsymbol{y}_1$, $\boldsymbol{X}^T \boldsymbol{y}_2$ and does not require computation of the GRM. Using the fact that the genotype matrix only contains entries in $\{0, 1, 2\}$, we can compute these quantities in time $\mathcal{O}(\frac{NM}{max(\log_3 N, \log_3 M)})$ [30]. Thus, when phenotypes are measured on the same set of samples, SCORE-OVERLAP can efficiently estimate ρ_g with no randomization.

3.3 Experiments

3.3.1 Accuracy

We performed simulations on a subset of $5,000$ unrelated white British individuals from the UK Biobank so that all methods compared could be run in a reasonable time. Our simulations used $305,630$ SNPs with minor allele frequency (MAF) above 1% (we chose these SNPs since these were also used for benchmarking the HDL [43] method and had reference eigenvectors available).

Given the genotypes, we simulated pairs of traits under varying genetic architectures. Our first set of architectures assumes an infinitesimal model (where all variants have a non-zero effect on both traits). We varied genetic correlation ρ_g

across $\{0, 0.2, 0.5, 0.8\}$ and the heritability of the pair of traits, (h_1^2, h_2^2), across values of $\{(0.1, 0.2), (0.2, 0.6), (0.5, 0.5), (0.6, 0.8)\}$ corresponding to the situation where both traits have low heritability, one trait has low while the other has moderate heritability, both traits have moderate heritability, and both have high heritability.

Our next set of non-infinitesimal architectures explores traits with medium polygenicity and low polygenicity respectively. For each SNP m, we specify a causal status, $\boldsymbol{c}_m$, which is a 2×1 vector with entries taking values in $\{0, 1\}$ according to whether SNP m has a non-zero effect on each of the two traits. For medium polygenicity, causal status at SNP m is drawn independently according to the following distribution: $P(\boldsymbol{c}_m = \begin{bmatrix} 1 \\ 1 \end{bmatrix}) = 0.1$, $P(\boldsymbol{c}_m = \begin{bmatrix} 0 \\ 1 \end{bmatrix}) = P(\boldsymbol{c}_m = \begin{bmatrix} 1 \\ 0 \end{bmatrix}) = 0.2$, and $P(\boldsymbol{c}_m = \begin{bmatrix} 0 \\ 0 \end{bmatrix}) = 0.5$.

The effect size $\boldsymbol{\beta}_m$ of SNP m on each trait is drawn from the following distribution:

$$\boldsymbol{\beta}_m|\boldsymbol{c}_m = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} \frac{\sigma_{g1}^2}{0.3M} & \gamma_g \\ \gamma_g & \frac{\sigma_{g2}^2}{0.3M} \end{bmatrix}, \boldsymbol{\beta}_m|\boldsymbol{c}_m = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} \frac{\sigma_{g1}^2}{0.3M} & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{\beta}_m|\boldsymbol{c}_m = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} 0 & 0 \\ 0 & \frac{\sigma_{g2}^2}{0.3M} \end{bmatrix}$$

For low polygenicity, we set the probability $P(\boldsymbol{c}_m = \begin{bmatrix} 1 \\ 1 \end{bmatrix}) = 0.01$, $P(\boldsymbol{c}_m = \begin{bmatrix} 0 \\ 1 \end{bmatrix}) = P(\boldsymbol{c}_m = \begin{bmatrix} 1 \\ 0 \end{bmatrix}) = 0.05$, and $P(\boldsymbol{c}_m = \begin{bmatrix} 0 \\ 0 \end{bmatrix}) = 0.89$.

The effect size $\boldsymbol{\beta}_m$ for genetic variant m on both traits are drawn from the fol-

lowing distribution:

$$\boldsymbol{\beta}_m|\boldsymbol{c}_m = \begin{bmatrix}1\\1\end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix}\frac{\sigma^2_{g1}}{0.06M} & \gamma_g\\ \gamma_g & \frac{\sigma^2_{g2}}{0.06M}\end{bmatrix}), \boldsymbol{\beta}_m|\boldsymbol{c}_m = \begin{bmatrix}1\\0\end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix}\frac{\sigma^2_{g1}}{0.06M} & 0\\ 0 & 0\end{bmatrix}), \boldsymbol{\beta}_m|\boldsymbol{c}_m = \begin{bmatrix}0\\1\end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix}0 & 0\\ 0 & \frac{\sigma^2_{g2}}{0.06M}\end{bmatrix})$$

We vary γ_g across $\{0, 0.2, 0.5, 0.8\}$. Under this model, the true total expected genome wide genetic correlation for medium polygenicity is $\{0, 0.06, 0.15, 0.24\}$, and $\{0, 0.0024, 0.03, 0.048\}$ for low polygenicity. Unless specified otherwise, we assume complete sample overlap, no environmental correlation, set the environmental variance so that the trait variance is 1, and simulate a total of 100 replicates for each architecture.

We performed simulations to compare the accuracy of SCORE to other estimators of genetic correlation under different genetic architectures. Specifically, we compared SCORE to methods that use individual data (bi-variate GREML [36], bi-variate Haseman-Elston regression) and methods that rely on summary statistics (LD-score regression (LDSC) [10] and HDL [43]). Bi-variate GREML (GCTA-GREML) and Haseman-Elston regression (GCTA-HE) are implemented in the GCTA software. LDSC is a widely used method to estimate genetic correlation when only summary statistics from GWAS on pairs of traits are available. HDL is a recent summary-statistics-based method that has been shown to obtain improved statistical efficiency relative to LDSC given additional information about LD. We ran all methods on the same set of SNPs to ensure a fair comparison (see Section A.4 for more details on data processing and methods).

We performed simulations to assess the estimation accuracy of each method using a subset of 5,000 unrelated white British individuals in the UK Biobank so that all

the methods could be run in a reasonable time. Unless otherwise specified, all our simulations used $305,630$ SNPs with MAF above 1%. We simulated pairs of traits under a total of 48 genetic architectures: varying heritability of the pair of traits (h_1^2, h_2^2), genetic correlation (ρ_g), and polygenicity (proportion of causal variants shared and unique to each trait).

The simulations assume that the two traits are measured on the same set of individuals so that both SCORE and SCORE-OVERLAP can be applied in this setting. Since SCORE is a randomized estimator, we first examined the choice of the number of random vectors (B) on the estimates of ρ_g. First, we confirmed that SCORE (with $B = 10$ and $B = 100$ random vectors) and SCORE-OVERLAP yield nearly identical results across the 48 architectures (Table 3.1). Second, we ran SCORE with different choices of $B = 10$ random vectors on a single replicate that was simulated under the infinitesimal architecture with trait heritability $(h_1^2, h_2^2) = (0.2, 0.6)$, and $\rho_g = 0.5$. We observe that the standard deviation of ρ_g estimates across choices of random vectors is about 18% of the total standard error (SE) indicating that the choice of $B = 10$ makes a modest contribution to variability in ρ_g estimates. These results lead us to use SCORE with $B = 10$ as our default.

Across the 48 architectures that we examined, the SE of SCORE ranges from 0.89 to 1.17 relative to the SE of GCTA-GREML with the SE of SCORE being 2.5% higher than that of GCTA-GREML on average (Figure 3.1). Interestingly, GCTA-HE tends to have a SE of 1.38 times that of SCORE on average (range 1.2 to 1.6). Compared to methods that rely on summary statistics, LDSC has 1.8 times the SE of SCORE on average (range 1.08 to 2.63) while the SE of HDL relative to SCORE is 1.24 (range 1.05 to 1.65) (Figure 3.1, Table 3.2). The reduction in the SE of SCORE relative to the summary statistic-based methods is equivalent to

a 3.24-fold increase in sample size over LDSC and a 1.56-fold increase in sample size over HDL on average. We find that the accuracy of SCORE relative to the other methods is consistent across infinitesimal (Figure 3.2) and non-infinitesimal architectures (Figure 3.3 for medium and Figure 3.4 for low polygenicity; the bias, SE, and MSE of each of the methods are listed in Tables 3.3, 3.4, 3.5). We additionally investigated the accuracy of each of the methods across a larger sample size of $10,000$ unrelated white British individuals chosen so that it was computationally feasible to run all methods including GCTA-GREML and GCTA-HE. Under a non-infinitesimal architecture with medium polygenicity, $\rho_g = 0.5$ and $(h_1^2, h_2^2) = (0.2, 0.6)$. In this larger sample size, we observe that SE of GCTA-GREML, GCTA-HE, and LDSC are $0.97, 1.54$, and 2.85 times of SCORE respectively, consistent with our results on a $N = 5,000$.

3.3.2 Robustness

We performed additional simulations to investigate the robustness of SCORE. First, we verified that the Jackknife standard error estimate used in SCORE is generally accurate while being conservative for low trait heritability (Table 3.6). Second, we verified the false positive rate of SCORE is controlled in simulations where ρ_g is zero. For each of 100 replicates in a given genetic architecture, we computed p-values for the two-tailed test of the null hypothesis that ρ_g is zero. Averaging across all architectures, we observe that the false positive rate (the fraction of simulations for which the p-value < 0.05) is 0.04 (Table 3.7).

3.3.3 The impact of sample overlap

We simulated traits under an infinitesimal architecture with $(h_1^2, h_2^2) = (0.2, 0.6)$ and $\rho_g = 0.5$. For each trait, we fixed the sample size to 5000 and varied the proportion of sample overlap across $\{0, 0.2, 0.5, 0.8, 1\}$ (ranging from no overlap to complete overlap). Specifically, for overlap proportion equal to 0, we have 5000 samples with observations on the first trait and a distinct set of 5000 samples with observations on the second trait. For overlap proportion equal to 1, we have 5000 samples with observations on both traits. We estimated genetic correlation with SCORE, LDSC, and GCTA-GREML.

The SE of SCORE relative to GCTA-GREML and LDSC remains stable as a function of sample overlap (Figure 3.5 and Table 3.8 for the bias, SE, and MSE of SCORE, GCTA-GREML, and LDSC as a function of sample overlap).

3.3.4 Accuracy for binary traits

Given $291,273$ unrelated white British individuals in the UK Biobank measured on $459,792$ genetic variants, we simulated pairs of traits under an infinitesimal architecture setting $(h_1^2, h_2^2) = (0.272, 0.12)$ and $\rho_g = -0.23$ while varying the environmental correlation across $\{0.04, -0.04, 0\}$. To simulate binary traits, we converted the second trait to a binary trait by thresholding the underlying continuous trait such that the prevalence varied across $\{0.01\%, 0.5\%, 1\%\}$.

We observe that the ρ_g estimates of SCORE are unbiased across the range of prevalence of the binary trait (Table 3.9). Further, the estimates of ρ_g obtained by SCORE tend to have relatively low SE provided the prevalence of the trait is greater than 0.5% (Table 3.9) so that we recommend applying SCORE to traits

whose prevalence is no less than 0.5%.

Finally, we validated the false positive rate of SCORE with different prevalence and observed that the false positive rate is not affected by the prevalence of binary trait (Table 3.7).

3.3.5 Computational Efficiency

We investigated the computational efficiency of SCORE relative to GCTA-GREML and GCTA-HE. The runtime and memory usage of summary statistic methods (LDSC and HDL) depends on the time needed to compute LD scores and summary statistics of each trait. In addition, HDL also requires the computation of the singular value decomposition (SVD) of LD matrices which is a computationally expensive step. Thus, we do not include runtimes for LDSC and HDL in these comparisons. We varied the number of individuals while the number of SNPs was fixed at $454,207$. Figure 3.6 shows that GCTA-GREML and GCTA-HE could not scale beyond sample sizes greater than 100,000 due to the requirement of computing and operating on a GRM (we extrapolate the runtime of GCTA-GREML and GCTA-HE to be about 340 days and 44 days on the set of $291,273$ unrelated white British individuals in the UK Biobank). On the other hand, SCORE ran in about 1.5 hours on the set of $291,273$ individuals using partial overlap mode with $B = 10$ random vectors while the SCORE-OVERLAP variant ran in about 1 hour on the same dataset.

3.4 Estimates of genome-wide genetic correlation in the UK Biobank

We applied SCORE to estimate ρ_g for pairs of phenotypes in the UK Biobank across $291,273$ unrelated white British individuals and $454,207$ SNPs (Material and Methods). We compared the ρ_g estimates obtained by LDSC versus SCORE for a subset of 28 traits in which LDSC produced valid estimates, *i.e.*, traits for which none of the ρ_g estimates were NA (Figure 3.7). While the point estimates of ρ_g from the two methods are highly concordant (Pearson correlation $r = 0.95$), the SE of LDSC is about 1.57 times that of SCORE on average which is equivalent to a 2.46-fold increase in sample size using SCORE (see Figures 3.8, 3.9). In total, 192 pairs of traits were detected to have a significant non-zero ρ_g by both SCORE and LDSC after Bonferroni correction for all pairs across the original set of forty phenotypes ($p < \frac{0.05}{780}$). Consistent with its reduced SE, SCORE found 58 pairs with significant ρ_g after Bonferroni correction that were not detected as significant by LDSC ($p < \frac{0.05}{780}$; stars in Figure 3.7). We conclude that SCORE obtains improved power to identify statistically significant genetic correlations relative to LDSC.

We obtain concordant results when analyzing all pairs in our initial set of forty traits. While the point estimates of SCORE and LDSC are highly correlated (Pearson correlation $r = 0.96$), the SE of LDSC is about 1.8 times that of SCORE on average, equivalent to a 3.24-fold increase in the sample size. In this setting, SCORE found 200 additional pairs of traits over LDSC (beyond the 245 pairs identified by both) while LDSC detected one pair as significant that SCORE did not detect as significant (Figure 3.7). To understand the impact of random vectors, we repeated our analysis with a different set of random vectors and observed that the Pearson correlation of

ρ_g estimates using the two sets is 0.999 (Figure 3.10).

We also analyzed all pairs in our initial set of forty traits with HDL using the set of $305,630$ SNPs for which reference eigenvectors are available [43] (Figure 3.11). The SE of HDL is about 2.53 times that of SCORE on average, which is equivalent to a 6.4-fold increase in the sample size (HDL failed to converge for 11% of the pairs where at least one of the traits is binary). Among these pairs, SCORE found 171 additional pairs of traits over HDL (beyond the 239 pairs identified by both) while HDL detected 14 pairs as significant that SCORE did not detect as significant. The summary of the ratio of SE of HDL and SCORE is shown in Figure 3.12.

To gain further insights into SCORE, we examined the SE of ρ_g estimates for pairs of traits according to whether the traits were both binary, both quantitative, or had one member of the pair being binary while the other was quantitative. The SE is largest when both traits are binary, intermediate when one of the traits is binary, and lowest when both traits are quantitative (average SE: 0.082, 0.035, and 0.02 respectively; Figure 3.13). We note that the SE increases when the prevalence of the binary trait decreases: the mean SE is 0.017 when the binary trait has prevalence $> 25\%$ while the mean SE is 0.047 for pairs in which the binary trait has prevalence $< 5\%$ (Figure 3.14).

We also applied SCORE to imputed genotypes in $291,273$ unrelated white British individuals and $4,824,392$ SNPs (MAF $> 1\%$). SCORE required about 19 hours to analyze a single pair of traits for imputed SNPs while requiring about 1.5 hours on array SNPs (scaling linearly with the number of variants). Since SCORE uses a streaming approach that does not require all SNPs to be stored in memory, it is memory efficient requiring about 2.3 GB to analyze imputed data. The estimates

of ρ_g are largely concordant across array and imputed SNPs (Pearson correlation of the ρ_g point estimates using two sets of SNPs is 0.973). We found 423 trait pairs that have significant non-zero ρ_g estimates (after Bonferroni correction) across both imputed and array genotypes while 19 pairs are significant only in the analysis of imputed genotypes while 22 pairs are significant in the analysis of array genotypes (Figure 3.15).

To further illustrate its utility, we applied SCORE to estimate genetic correlation between coronary artery disease related traits included in our set of forty traits (angina and heart attack) and serum biomarkers (alanine (ALT) and aspartate aminotransferase (AST)). Serum liver enzyme levels, including ALT and AST, are markers of liver health and hepatic dysfunction, and they have been shown to be associated with cardiovascular disease [50, 51, 52], though the strength and consistency has varied among the studies [50]. We observed significant positive ρ_g between ALT/AST and the two coronary artery-disease related trait (0.257 ± 0.04 and 0.169 ± 0.032 for angina with ALT and AST respectively; 0.239 ± 0.053 and 0.148 ± 0.04 for heart attack with ALT and AST respectively). Our finding of significant positive ρ_g suggests that hepatic dysfunction (higher serum levels of ALT and AST) and coronary artery disease have a shared genetic component.

3.5 Discussion

In this chapter, we defined model base on pairs of phenotypes and multiple traits jointly. We have described SCORE, a scalable and accurate estimator of genetic correlation. We observe that the estimates of genetic correlation obtained by SCORE obtain accuracy comparable to GREML [44] while being scalable to Biobank-scale

data. SCORE can estimate the genetic correlation across pairs of traits when applied to $\approx 500K$ common SNPs measured on $\approx 300K$ unrelated white British individuals in the UK Biobank within a few hours. In simulations, we showed that, compared to summary-statistic methods, SCORE obtains a reduction in the average standard error of 44% relative to LDSC and 20% relative to HDL, equivalent to a 3.24-fold and 1.56-fold increase in sample size. In application to 780 pairs of traits in the UK Biobank, SCORE discovered 200 pairs of traits with significant genetic correlation (after correcting for multiple testing) that were not discovered by LDSC. In application to 780 pairs, SCORE discovered 171 pairs of traits with significant genetic correlation (after correcting for multiple testing) that were not discovered by HDL while HDL discovered 14 significant pairs not discovered by SCORE. It is plausible that the results of HDL might be altered by the computation of eigenvectors from the analyzed genotypes although such an analysis can be computationally expensive

The statistical accuracy gain of SCORE relative to LDSC and HDL can be attributed to several factors. LDSC does not use all the available covariances among the summary statistics choosing to only model the variance. The LD information as summarized by the LD scores involves a number of approximations. Typically, LD scores are computed from an external reference panel. Even when in-sample LD is used (as we have here), computational considerations lead to the LD scores being computed from a subset of the samples and restricted to SNPs that fall within a fixed-length genomic window. While HDL models the covariance structure among the summary statistics thereby utilizing additional information relative to LDSC, HDL relies on approximate computations of LD scores like LDSC. To enable computational efficiency, HDL also uses a truncated singular value decomposition(SVD) of the LD score matrix that can potentially reduce accuracy.

We discuss several limitations of SCORE. Firstly, SCORE requires access to individual genotype and trait data. Summary-statistic methods such as LDSC and HDL have the advantage of being applicable in settings where access to individual-level data can be challenging. While summary-statistic methods also have the advantage of being relatively efficient, it is important to keep in mind that summary statistics are dependent on specific choices of marker sets and covariates. Applying these methods to different sets of covariates and marker sets requires regenerating the summary statistics (and auxiliary information such as LD score matrices). Second, the model underlying SCORE assumes a quantitative trait. We have shown empirically that SCORE provides accurate estimates of genetic correlation when applied to binary traits provided the traits are not too rare (prevalence $> 0.5\%$). It would be of interest to extend SCORE to the setting of binary traits along the lines of the PCGC method [42]. Finally, while SCORE estimates genome-wide genetic correlation, efficient methods that can partition genetic correlation across genomic annotations can provide additional insights into the shared genetic basis of traits.

Polygenicity	Method	Genetic correlation ρ_g	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE
	SCORE		$h_1^2 = 0.1, h_2^2 = 0.2$			$h_1^2 = 0.2, h_2^2 = 0.6$			$h_1^2 = 0.5, h_2^2 = 0.5$			$h_1^2 = 0.6, h_2^2 = 0.8$		
Infinitesimal	OVERLAP	0	-0.0826	0.1019	0.3083	-0.0302	0.0428	0.2048	0.0099	0.0145	0.1201	0.0083	0.0084	0.0912
	B=10	0	-0.0826	0.1018	0.3082	-0.0301	0.0428	0.2047	0.0099	0.0145	0.1201	0.0083	0.0084	0.0912
	B=100	0	-0.0826	0.1018	0.3082	-0.0301	0.0428	0.2047	0.0099	0.0145	0.1201	0.0083	0.0084	0.0912
	OVERLAP	0.2	-0.05	0.132	0.3598	-0.01	0.033	0.1814	0.0147	0.0159	0.1253	-0.0173	0.0089	0.0928
	B=10	0.2	-0.05	0.1319	0.3597	-0.01	0.033	0.1813	0.0147	0.0159	0.1253	-0.0173	0.0089	0.0928
	B=100	0.2	-0.05	0.1319	0.3597	-0.01	0.033	0.1813	0.0147	0.0159	0.1253	-0.0173	0.0089	0.0928
	OVERLAP	0.5	-0.1567	0.0841	0.244	-0.016	0.0266	0.1622	0.0115	0.0137	0.1164	0.0143	0.0047	0.0669
	B=10	0.5	-0.1568	0.0841	0.244	-0.0161	0.0266	0.1621	0.0114	0.0137	0.1164	0.0142	0.0047	0.0669
	B=100	0.5	-0.1568	0.0841	0.244	-0.0161	0.0266	0.1621	0.0116	0.0138	0.1169	0.0142	0.0047	0.0669
	OVERLAP	0.8	-0.3187	0.156	0.2334	-0.1106	0.0309	0.1365	-0.0206	0.0067	0.0789	-0.0016	0.0034	0.0583
	B=10	0.8	-0.3187	0.156	0.2333	-0.1107	0.0309	0.1365	-0.0207	0.0067	0.0789	-0.0017	0.0034	0.0583
	B=100	0.8	-0.3187	0.156	0.2333	-0.1107	0.0309	0.1365	-0.0207	0.0067	0.0789	-0.0017	0.0034	0.0583
Medium polygenicity	OVERLAP	0	0.0506	0.1227	0.3466	0.0023	0.0349	0.1869	0.0022	0.0198	0.1408	0.0052	0.0091	0.0952
	B=10	0	0.0506	0.1226	0.3465	0.0023	0.0349	0.1868	0.0022	0.0198	0.1408	0.0052	0.0091	0.0952
	B=100	0	0.0506	0.1226	0.3465	0.0023	0.0349	0.1868	0.0022	0.0198	0.1408	0.0052	0.0091	0.0952
	OVERLAP	0.2	0.0259	0.0965	0.3096	0.0116	0.0382	0.1952	-6e-04	0.0161	0.1267	-0.01	0.0065	0.0797
	B=10	0.2	0.0259	0.0965	0.3096	0.0115	0.0382	0.1952	-7e-04	0.0161	0.1267	-0.01	0.0065	0.0797
	B=100	0.2	0.0259	0.0965	0.3096	0.0115	0.0382	0.1952	-7e-04	0.0161	0.1267	-0.01	0.0065	0.0797
	OVERLAP	0.5	0.0317	0.0974	0.3104	0.0039	0.0287	0.1693	0.0197	0.0118	0.1066	0.0076	0.0083	0.0909
	B=10	0.5	0.0316	0.0973	0.3104	0.0039	0.0287	0.1693	0.0197	0.0117	0.1066	0.0076	0.0083	0.0909
	B=100	0.5	0.0316	0.0973	0.3104	0.0039	0.0287	0.1693	0.022	0.0114	0.1045	0.0076	0.0083	0.0909
	OVERLAP	0.8	-0.1181	0.1209	0.3271	0.0116	0.031	0.1758	0.0228	0.013	0.1119	0.031	0.0121	0.1054
	B=10	0.8	-0.1182	0.1209	0.327	0.0115	0.031	0.1758	0.0228	0.013	0.1118	0.0309	0.0121	0.1054
	B=100	0.8	-0.1182	0.1209	0.327	0.0115	0.031	0.1758	0.0228	0.013	0.1118	0.0309	0.0121	0.1054
Low polygenicity	OVERLAP	0	0.0694	0.136	0.3623	0.0115	0.0452	0.2123	-0.0223	0.0158	0.1236	-0.0091	0.0101	0.0999
	B=10	0	0.0694	0.136	0.3622	0.0115	0.0452	0.2123	-0.0223	0.0158	0.1236	-0.0091	0.0101	0.0999
	B=100	0	0.0694	0.136	0.3622	0.0115	0.0452	0.2123	-0.0223	0.0158	0.1236	-0.0091	0.0101	0.0999
	OVERLAP	0.2	0.0135	0.0594	0.2433	0.0366	0.0424	0.2025	0.018	0.0188	0.136	0.0193	0.0105	0.1004
	B=10	0.2	0.0135	0.0594	0.2433	0.0366	0.0424	0.2025	0.018	0.0188	0.136	0.0193	0.0105	0.1004
	B=100	0.2	0.0135	0.0594	0.2433	0.0366	0.0424	0.2025	0.018	0.0188	0.136	0.0193	0.0105	0.1004
	OVERLAP	0.5	0.0463	0.0992	0.3116	0.0478	0.0261	0.1544	0.0882	0.0215	0.1171	0.0334	0.0121	0.1048
	B=10	0.5	0.0463	0.0992	0.3116	0.0478	0.0261	0.1544	0.0882	0.0215	0.1171	0.0334	0.0121	0.1048
	B=100	0.5	0.0463	0.0992	0.3116	0.0478	0.0261	0.1544	0.0882	0.0215	0.1171	0.0334	0.0121	0.1048
	OVERLAP	0.8	0.0363	0.083	0.2857	0.0953	0.042	0.1815	0.0839	0.0253	0.1353	0.0869	0.0175	0.0998
	B=10	0.8	0.0363	0.083	0.2857	0.0953	0.042	0.1815	0.0839	0.0253	0.1353	0.0869	0.0175	0.0998
	B=100	0.8	0.0363	0.083	0.2857	0.0953	0.042	0.1815	0.0839	0.0253	0.1353	0.0869	0.0175	0.0998

Table 3.1: **Estimates of bias, mean square error, and standard error of SCORE for varying number of random vectors $B = 10$, $B = 100$ and SCORE-OVERLAP**.

Polygenicity	Method	Genetic correlation ρ_g	(h_1^2, h_2^2)			
			$(0.1, 0.2)$	$(0.2, 0.6)$	$(0.5, 0.5)$	$(0.6, 0.8)$
Infinitesimal	LDSC/SCORE	0	1.79	1.79	2.41	2.32
	HDL/SCORE	0	1.35	1.2	1.42	1.35
	LDSC/SCORE	0.2	1.26	1.64	2.63	2.32
	HDL /SCORE	0.2	1.05	1.48	1.47	1.5
	LDSC/SCORE	0.5	1.69	2	2.19	2.6
	HDL/SCORE	0.5	1.65	1.14	1.18	1.39
	LDSC/SCORE	0.8	1.4	1.62	1.95	2.16
	HDL/SCORE	0.8	1.26	1.23	1.41	1.28
Medium polygenicity	LDSC/SCORE	0	1.4	1.96	2.08	2.07
	HDL/SCORE	0	1.25	1.41	1.21	1.36
	LDSC/SCORE	0.2	1.74	2.09	2.15	2.28
	HDL/SCORE	0.2	1.44	1.54	1.38	1.4
	LDSC/SCORE	0.5	1.49	1.96	2.59	2.37
	HDL/SCORE	0.5	1.41	1.32	1.54	1.42
	LDSC/SCORE	0.8	1.6	2.01	2.42	2.2
	HDL/SCORE	0.8	1.17	1.45	1.61	1.23
Low polygenicity	LDSC/SCORE	0	1.08	1.92	2.57	2.17
	HDL/SCORE	0	1.11	1.18	1.27	1.15
	LDSC/SCORE	0.2	1.79	1.57	1.97	2.24
	HDL/SCORE	0.2	1.31	1.07	1.29	1.3
	LDSC/SCORE	0.5	1.18	2.25	2.49	1.9
	HDL/SCORE	0.5	1.08	1.44	1.26	1.29
	LDSC/SCORE	0.8	1.45	1.81	2.47	2.37
	HDL/SCORE	0.8	1.24	1.28	1.38	1.36

Table 3.2: **Ratio of SE of summary-statistic methods relative to SCORE ($N = 5,000$ individuals, $M = 305,630$ SNPs).**

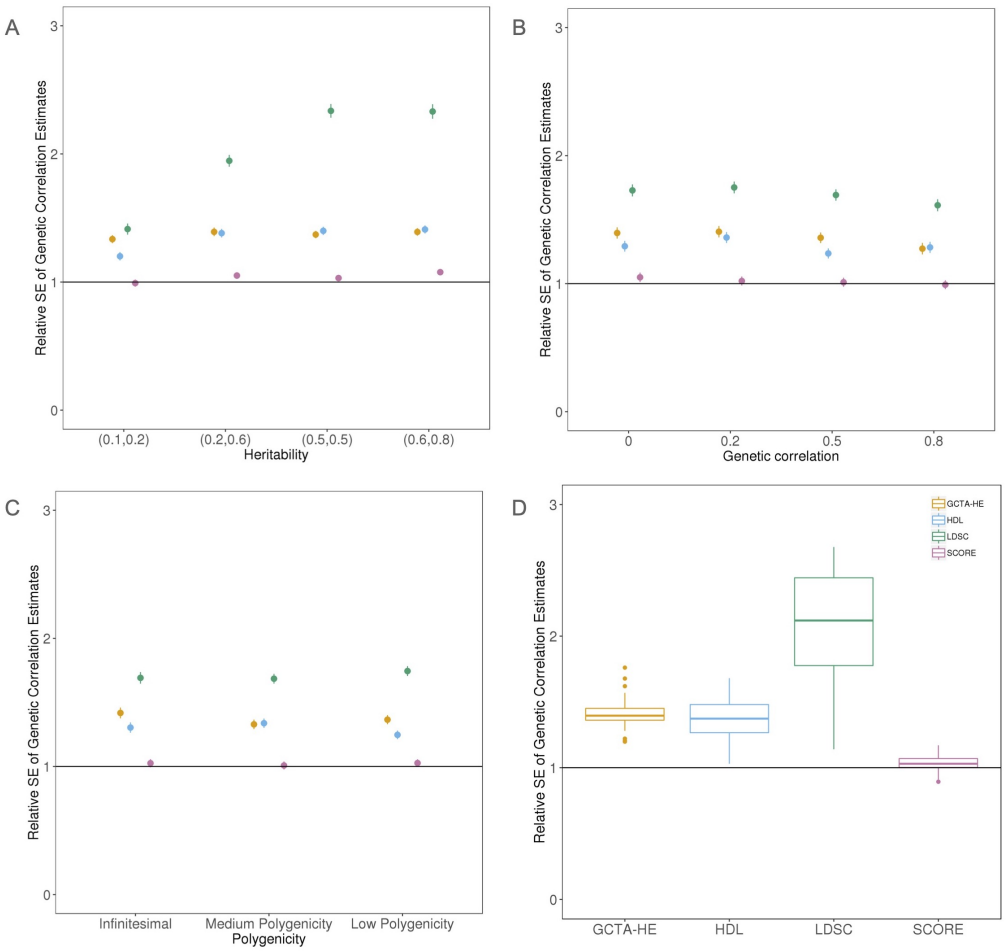

Figure 3.1: **Comparison of the estimates of genetic correlation from SCORE with GCTA-GREML, GCTA-HE, LDSC, and HDL** .

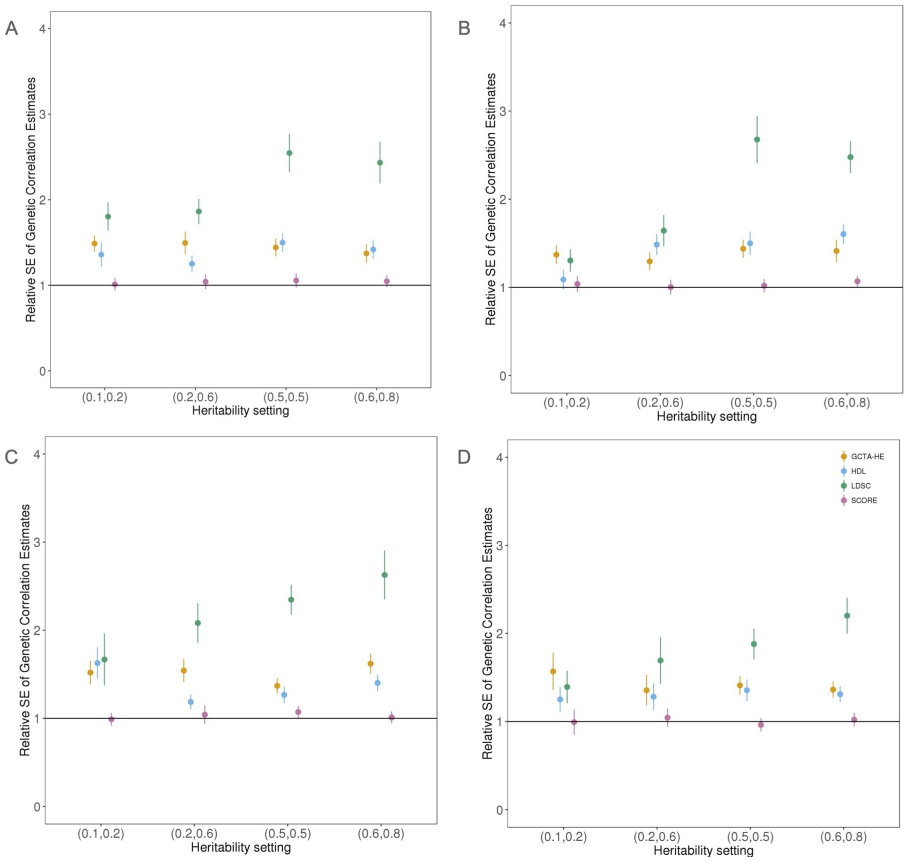

Figure 3.2: **Comparison of the estimates of genetic correlation from SCORE with GCTA-GREML, GCTA-HE, LDSC, and HDL in small-scale simulations ($N = 5,000$ unrelated individuals, $M = 305,630$ SNPs) under infinitesimal genetic architectures**. We simulated pairs of phenotypes under 16 different infinitesimal genetic architectures. Panel A, B, C, D correspond to a different value of the genetic correlation chosen from the set: $\{0, 0.2, 0.5, 0.8\}$. Within each panel, we varied the SNP heritability for the pair of traits across $\{(0.1, 0.2), (0.2, 0.6), (0.5, 0.5), (0.6, 0.8)\}$ (see Simulations to assess accuracy section of Materials and Methods). We plot the standard error (SE) of each method relative to GCTA-GREML. We estimate the standard error of the relative SE using Jackknife (error bars denote 1 standard error).

Method	Genetic	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE
	correlation ρ_g	$h_1^2 = 0.1, h_2^2 = 0.2$			$h_1^2 = 0.2, h_2^2 = 0.6$			$h_1^2 = 0.5, h_2^2 = 0.5$			$h_1^2 = 0.6, h_2^2 = 0.8$		
GCTA-GREML	0	-0.0818	0.0998	0.305	-0.029	0.0395	0.197	0.0096	0.013	0.114	0.0014	0.0076	0.0871
GCTA-HE	0	-0.104	0.217	0.454	0.0366	0.0878	0.294	0.0026	0.0268	0.164	0.0225	0.0148	0.119
HDL	0	-0.0265	0.172	0.414	-0.0095	0.0606	0.246	0.0185	0.0294	0.17	0.007	0.0153	0.123
LDSC	0	0.0519	0.3055	0.5503	-0.007	0.1343	0.3665	-0.0014	0.0837	0.2894	0.0083	0.045	0.2119
SCORE	0	-0.0826	0.102	0.308	-0.0302	0.0428	0.205	0.0099	0.0145	0.12	0.0083	0.0084	0.0912
GCTA-GREML	0.2	-0.0571	0.123	0.346	-0.0103	0.0328	0.181	0.0074	0.0152	0.123	-0.0114	0.0077	0.0868
GCTA-HE	0.2	-0.188	0.261	0.475	-0.0234	0.0554	0.234	0.029	0.0321	0.177	-0.0263	0.0157	0.123
HDL	0.2	-0.094	0.151	0.377	0.0254	0.0729	0.269	0.0236	0.0346	0.184	0.0027	0.0194	0.139
LDSC	0.2	-0.0946	0.2134	0.4522	0.0134	0.0885	0.2973	0.0177	0.1088	0.3293	0.0051	0.0463	0.215
SCORE	0.2	-0.05	0.132	0.36	-0.01	0.033	0.181	0.0147	0.0159	0.125	-0.0173	0.0089	0.0928
GCTA-GREML	0.5	-0.158	0.0858	0.247	-0.012	0.0244	0.156	0.0159	0.0121	0.109	0.0092	0.0045	0.0662
GCTA-HE	0.5	-0.142	0.161	0.375	-0.0394	0.0593	0.24	0.0181	0.0224	0.149	0.0111	0.0116	0.107
HDL	0.5	-0.162	0.188	0.401	-0.0195	0.0345	0.185	0.0047	0.0189	0.138	0.0207	0.009	0.0928
LDSC	0.5	-0.3298	0.278	0.4114	-0.0955	0.1143	0.3243	0	0.0649	0.2547	0.0308	0.0312	0.1739
SCORE	0.5	-0.1567	0.0841	0.244	-0.016	0.0266	0.1622	0.0115	0.0137	0.1164	0.0143	0.0047	0.0669
GCTA-GREML	0.8	-0.317	0.156	0.235	-0.106	0.0283	0.131	-0.0232	0.0073	0.082	-0.0004	0.0033	0.0571
GCTA-HE	0.8	-0.389	0.287	0.369	-0.115	0.0446	0.177	-0.023	0.0139	0.116	0.0093	0.0061	0.0778
HDL	0.8	-0.391	0.239	0.294	-0.154	0.0518	0.168	-0.0421	0.0141	0.111	-0.0114	0.0057	0.0748
LDSC	0.8	-0.4841	0.3415	0.3273	-0.1851	0.0833	0.2215	-0.0727	0.029	0.1541	-0.019	0.0162	0.1257
SCORE	0.8	-0.319	0.156	0.233	-0.111	0.0309	0.137	-0.0206	0.0067	0.0789	-0.0016	0.0034	0.0583

Table 3.3: **Bias, mean square error and standard error of genetic correlation estimation methods in simulations corresponding to Figure 3.2 ($N = 5,000$ individuals, $M = 305,630$ SNPs)**.

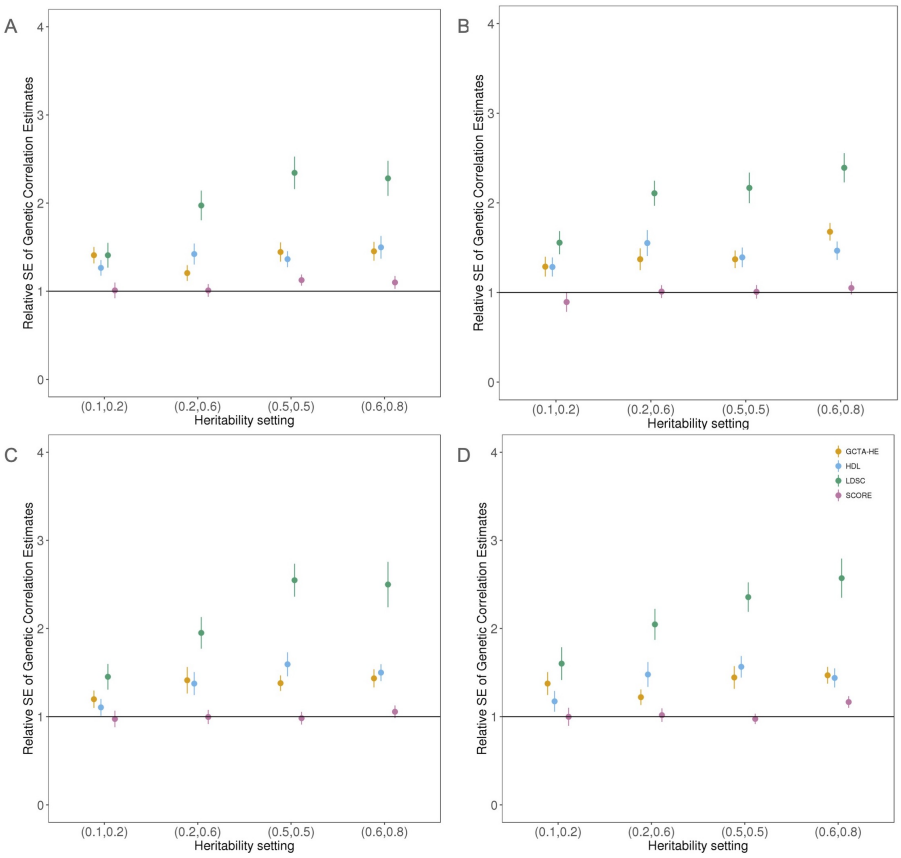

Figure 3.3: **Comparison of the estimates of genetic correlation from SCORE with GCTA-GREML, GCTA-HE, LDSC, and HDL in small-scale simulations** ($N = 5,000$ **unrelated individuals,** $M = 305,630$ **SNPs) under non-infinitesimal architectures with medium polygenicity**. We simulated pairs of phenotypes under 16 different non-infinitesimal genetic architectures. The probability of a variant being causal for both traits is 0.20, and the probability of a variant being causal for exactly one of the traits is 0.10. Panels (A, B, C, D) correspond to a different value of the genetic correlation at SNPs causal for both traits: $\{0, 0.2, 0.5, 0.8\}$. The causal variants are distributed uniformly across the genome. Within each panel, we varied the per-SNP heritability of variants causal for both traits to be proportional to $\{(0.1, 0.2), (0.2, 0.6), (0.5, 0.5), (0.6, 0.8)\}$ (see Simulations to assess accuracy section of Materials and Methods). We plot the SE of each method relative to GCTA-GREML. We ran LDSC with in-sample LD and HDL with eigenvectors that preserve 90% variance. We estimate the standard error of the relative SE using Jackknife (error bars denote 1 standard error).

Method	Genetic	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE
	correlation ρ_g	$h_1^2 = 0.1, h_2^2 = 0.2$			$h_1^2 = 0.2, h_2^2 = 0.6$			$h_1^2 = 0.5, h_2^2 = 0.5$			$h_1^2 = 0.6, h_2^2 = 0.8$		
GCTA-GREML	0	0.0414	0.12	0.344	0.003	0.0343	0.185	-0.0025	0.0157	0.125	-0.0011	0.0075	0.0866
GCTA-HE	0	0.0964	0.243	0.484	0.0163	0.0502	0.223	0.0103	0.0328	0.181	-0.0051	0.0159	0.126
HDL	0	0.0626	0.193	0.434	-0.0225	0.0699	0.263	0.0103	0.0292	0.171	0.0155	0.0171	0.13
LDSC	0	0.0608	0.2375	0.4835	-0.0273	0.1344	0.3656	0.0179	0.0862	0.2931	0.0301	0.0399	0.1975
SCORE	0	0.0506	0.123	0.347	0.0023	0.0349	0.187	0.0022	0.0198	0.141	0.0052	0.0091	0.0952
GCTA-GREML	0.2	-0.0034	0.12	0.347	0.0143	0.0376	0.193	-0.0055	0.0159	0.126	-0.007	0.0058	0.0759
GCTA-HE	0.2	0.0162	0.2	0.447	0.0182	0.0706	0.265	-0.0073	0.0298	0.172	-0.0247	0.0168	0.127
HDL	0.2	-0.0066	0.198	0.445	0.0155	0.0902	0.3	0.005	0.0307	0.175	-0.0121	0.0125	0.111
LDSC	0.2	-0.0846	0.2979	0.5392	0.0209	0.1664	0.4073	-0.0089	0.0744	0.2726	0.0052	0.033	0.1815
SCORE	0.2	0.0259	0.0965	0.31	0.0116	0.0382	0.195	-0.0006	0.0161	0.127	-0.01	0.0065	0.0797
GCTA-GREML	0.5	0.0292	0.102	0.319	0.0081	0.0289	0.17	0.0182	0.0121	0.108	0.0068	0.0075	0.0861
GCTA-HE	0.5	-0.0213	0.146	0.382	0.0061	0.0576	0.24	0.0138	0.0226	0.15	0.0155	0.0155	0.123
HDL	0.5	0.0137	0.124	0.352	-0.0255	0.0552	0.234	0.0223	0.0304	0.173	0.0055	0.0167	0.129
LDSC	0.5	-0.1026	0.2247	0.4628	-0.0634	0.1138	0.3313	-0.0142	0.0766	0.2764	-0.0202	0.0467	0.2151
SCORE	0.5	0.0317	0.0974	0.31	0.0039	0.0287	0.169	0.0197	0.0118	0.107	0.0076	0.0083	0.0909
GCTA-GREML	0.8	-0.112	0.12	0.328	0.0146	0.0301	0.173	0.0164	0.0134	0.115	0.036	0.0095	0.0904
GCTA-HE	0.8	-0.0941	0.212	0.451	0.0307	0.0454	0.211	0.0151	0.0277	0.166	0.0288	0.0185	0.133
HDL	0.8	-0.155	0.172	0.385	0.0355	0.0665	0.255	0.0362	0.0336	0.18	0.0453	0.019	0.13
LDSC	0.8	-0.2322	0.3293	0.5248	0.0022	0.125	0.3536	0.0144	0.0733	0.2704	0.0134	0.0542	0.2324
SCORE	0.8	-0.118	0.121	0.327	0.0116	0.031	0.176	0.0228	0.013	0.112	0.031	0.0121	0.105

Table 3.4: **Bias, mean square error and standard error of genetic correlation estimation methods in simulations corresponding to Figure 3.3 ($N = 5,000$ individuals, $M = 305,630$ SNPs).**

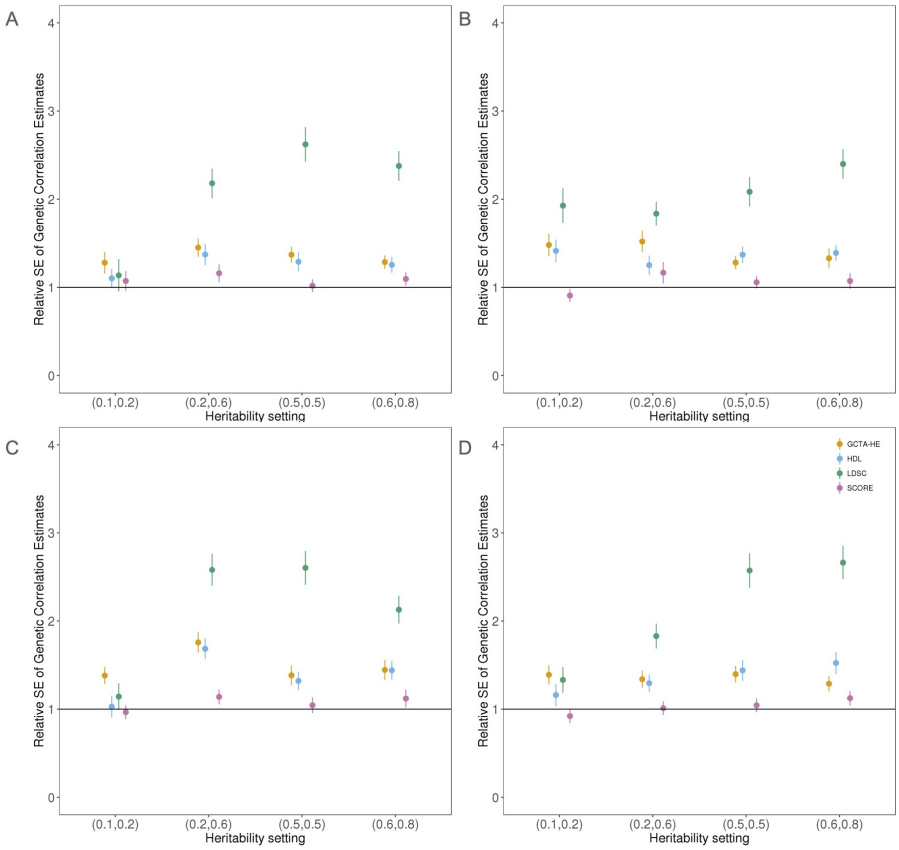

Figure 3.4: **Comparison of the estimates of genetic correlation from SCORE with GCTA-GREML, GCTA-HE, LDSC, and HDL in small-scale simulations ($N = 5,000$ unrelated individuals, $M = 305,630$ SNPs) under non-infinitesimal architectures with low polygenicity**. We simulated pairs of phenotypes under 16 different non-infinitesimal genetic architectures. The probability of a variant being causal for both traits is 0.01, and the probability of a variant being causal for exactly one of the trait is 0.05. Panels (A, B, C, D) correspond to a different value of the genetic correlation at SNPs causal for both traits: $\{0, 0.2, 0.5, 0.8\}$. The causal variants are distributed uniformly across the genome. Within each panel, we varied the per-SNP heritability of variants causal for both traits to be proportional to $\{(0.1, 0.2), (0.2, 0.6), (0.5, 0.5), (0.6, 0.8)\}$ (see Simulations to assess accuracy section of Materials and Methods). We plot the SE of each method relative to GCTA-GREML. We ran LDSC with in-sample LD and HDL with eigenvectors that preserve 90% variance. We estimate the standard error of the relative SE using Jackknife (error bars denote 1 standard error).

Method	Genetic	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE
	correlation ρ_g	$h_1^2 = 0.1, h_2^2 = 0.2$			$h_1^2 = 0.2, h_2^2 = 0.6$			$h_1^2 = 0.5, h_2^2 = 0.5$			$h_1^2 = 0.6, h_2^2 = 0.8$		
GCTA-GREML	0	0.0396	0.1156	0.3377	0.0188	0.0339	0.183	-0.0248	0.0153	0.1213	-0.0084	0.0084	0.0912
GCTA-HE	0	-0.0011	0.1867	0.4321	0.0314	0.0715	0.2656	-0.02	0.028	0.1662	-0.0041	0.0138	0.1174
HDL	0	0.0286	0.1395	0.3724	0.0212	0.0635	0.2511	-0.007	0.0245	0.1565	-0.0058	0.0132	0.1145
LDSC	0	-0.0844	0.1544	0.3838	0.0206	0.1595	0.3989	0.0035	0.1012	0.3181	0.0082	0.0471	0.2168
SCORE	0	0.0694	0.136	0.3623	0.0115	0.0452	0.2123	-0.0223	0.0158	0.1236	-0.0091	0.0101	0.0999
GCTA-GREML	0.2	-0.0198	0.0722	0.268	0.0296	0.031	0.1736	0.014	0.0167	0.1285	0.0115	0.0089	0.0936
GCTA-HE	0.2	0.0051	0.1576	0.3969	0.0224	0.0701	0.2639	0.012	0.0273	0.1647	0.0315	0.0165	0.1245
HDL	0.2	0.0812	0.1502	0.379	0.0744	0.0527	0.2173	0.0054	0.031	0.1761	0.0143	0.0171	0.1302
LDSC	0.2	0.0072	0.2668	0.5165	0.0589	0.105	0.3187	-0.0129	0.0719	0.2678	0.0119	0.0506	0.2246
SCORE	0.2	0.0135	0.0594	0.2433	0.0366	0.0424	0.2025	0.018	0.0188	0.136	0.0193	0.0105	0.1004
GCTA-GREML	0.5	0.0353	0.1053	0.3226	0.0472	0.0206	0.1356	0.0893	0.0206	0.1122	0.0379	0.0102	0.0936
GCTA-HE	0.5	0.0019	0.1984	0.4454	0.0411	0.0584	0.2382	0.1056	0.0352	0.1552	0.0527	0.021	0.1352
HDL	0.5	0.0171	0.1102	0.3315	0.0397	0.0537	0.2284	0.1039	0.0327	0.1481	0.0362	0.0195	0.1348
LDSC	0.5	0.1266	0.1518	0.3684	0.0636	0.1265	0.3499	0.0994	0.0952	0.2921	0.0514	0.0423	0.1992
SCORE	0.5	0.0463	0.0992	0.3116	0.0478	0.0261	0.1544	0.0882	0.0215	0.1171	0.0334	0.0121	0.1048
GCTA-GREML	0.8	0.0725	0.1016	0.3103	0.0939	0.0411	0.1796	0.0814	0.0235	0.1297	0.0876	0.0156	0.0888
GCTA-HE	0.8	-0.0097	0.1859	0.431	0.0844	0.0649	0.2405	0.0806	0.0393	0.1811	0.0765	0.0189	0.1144
HDL	0.8	0.1205	0.144	0.3598	0.0922	0.0625	0.2324	0.0993	0.0447	0.1867	0.0888	0.0262	0.1353
LDSC	0.8	0.0414	0.1724	0.4132	0.1331	0.1256	0.3284	0.0853	0.1186	0.3337	0.0741	0.0614	0.2365
SCORE	0.8	0.0363	0.083	0.2857	0.0953	0.042	0.1815	0.0839	0.0253	0.1353	0.0869	0.0175	0.0998

Table 3.5: **Bias, mean square error and standard error of genetic correlation estimation methods in simulations corresponding to Figure 3.4 ($N = 5,000$ individuals, $M = 305,630$ SNPs)**.

Polygenicity	(h_1^2, h_2^2)	ρ_g	$\hat{SE}$	SE
Infinitesimal	0.1, 0.2	0	0.45	0.42
	0.2, 0.6	0	0.2	0.2
	0.5, 0.5	0	0.12	0.12
	0.6, 0.8	0	0.09	0.09
	0.1, 0.2	0.2	0.4	0.38
	0.2, 0.6	0.2	0.19	0.19
	0.5, 0.5	0.2	0.12	0.12
	0.6, 0.8	0.2	0.09	0.09
	0.1, 0.2	0.5	0.41	0.3
	0.2, 0.6	0.5	0.18	0.17
	0.5, 0.5	0.5	0.11	0.11
	0.6, 0.8	0.5	0.07	0.07
	0.1, 0.2	0.8	0.41	0.34
	0.2, 0.6	0.8	0.18	0.15
	0.5, 0.5	0.8	0.09	0.09
	0.6, 0.8	0.8	0.05	0.05
Medium polygenicity	0.1, 0.2	0	0.45	0.39
	0.2, 0.6	0	0.19	0.19
	0.5, 0.5	0	0.12	0.14
	0.6, 0.8	0	0.09	0.09
	0.1, 0.2	0.2	0.43	0.39
	0.2, 0.6	0.2	0.18	0.19
	0.5, 0.5	0.2	0.12	0.12
	0.6, 0.8	0.2	0.09	0.08
	0.1, 0.2	0.5	0.43	0.36
	0.2, 0.6	0.5	0.18	0.18
	0.5, 0.5	0.5	0.12	0.11
	0.6, 0.8	0.5	0.09	0.09
	0.1, 0.2	0.8	0.44	0.36
	0.2, 0.6	0.8	0.20	0.18
	0.5, 0.5	0.8	0.12	0.11
	0.6, 0.8	0.8	0.1	0.11
Low polygenicity	0.1, 0.2	0	0.39	0.38
	0.2, 0.6	0	0.21	0.2
	0.5, 0.5	0	0.12	0.12
	0.6, 0.8	0	0.1	0.09
	0.1, 0.2	0.2	0.35	0.29
	0.2, 0.6	0.2	0.19	0.2
	0.5, 0.5	0.2	0.12	0.14
	0.6, 0.8	0.2	0.19	0.2
	0.1, 0.2	0.5	0.42	0.33
	0.2, 0.6	0.5	0.18	0.16
	0.5, 0.5	0.5	0.13	0.12
	0.6, 0.8	0.5	0.09	0.1
	0.1, 0.2	0.8	0.35	0.3
	0.2, 0.6	0.8	0.19	0.18
	0.5, 0.5	0.8	0.12	0.12
	0.6, 0.8	0.8	0.1	0.1

Table 3.6: **Assessment of Jackknife estimates of standard error ($N = 5,000$ samples and $305,630$ SNPs, block size $= 4,000$ SNPs)**. We report the average of estimates of standard error across 100 replicates.

Small-scale simulations		
Polygenicity	(h_1^2, h_2^2)	FPR
Infinitesimal	0.1, 0.2	0.014
	0.2 ,0.6	0.037
	0.5, 0.5	0.01
	0.6, 0.8	0.07
Medium polygenicity	0.1, 0.2	0
	0.2, 0.6	0.026
	0.5, 0.5	0.071
	0.6, 0.8	0.02
Low polygenicity	0.1, 0.2	0.065
	0.2, 0.6	0.054
	0.5, 0.5	0.05
	0.6, 0.8	0.06
Large-scale simulations		
Prevalence	(h_1^2, h_2^2)	FPR
Continuous	0.272, 0.12	0.04
50%	0.272, 0.12	0.08
10%	0.272, 0.12	0.07
1%	0.272, 0.12	0.08
0.5%	0.272, 0.12	0.02
0.01%	0.272, 0.12	0

Table 3.7: **The false positive rate of SCORE is controlled**. We evaluated the false positive rate of SCORE in simulations where ρ_g is zero. We considered small-scale ($N = 5,000$ individuals and $M = 305,630$ SNPs) and large-scale simulations ($N = 291,273$ individuals and $M = 305,630$ SNPs). We also considered simulations with binary traits with varying prevalence. Standard error estimates were obtained using a Block Jackknife with a block size of 4000 SNPs. For each genetic architecture, we performed 100 replicates and reported the FPR as the rate with which SCORE rejects the null hypothesis of $\rho_g = 0$.

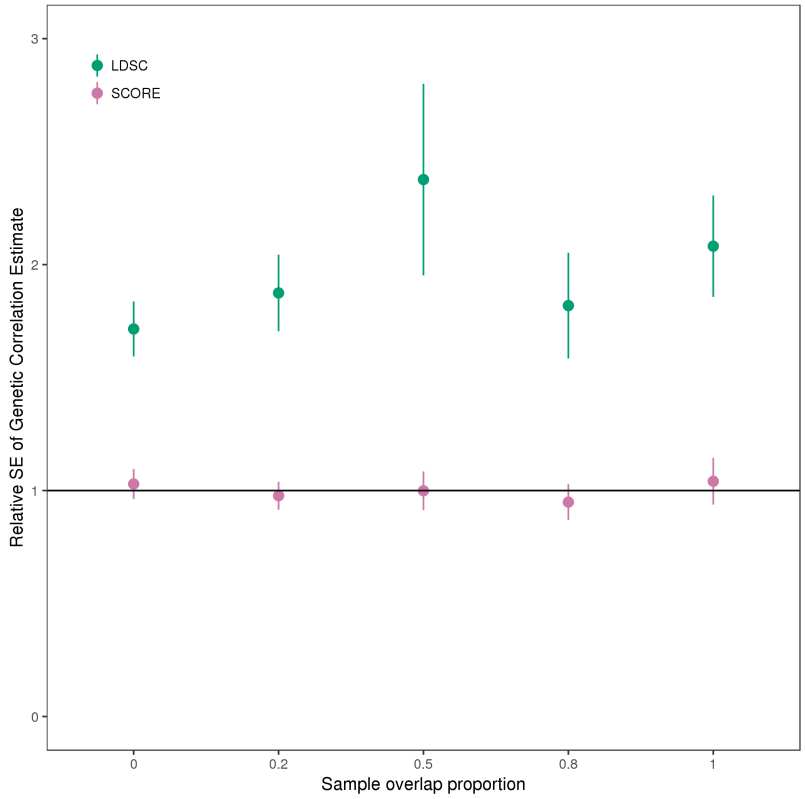

Figure 3.5: **Comparison of the estimates of genetic correlation from SCORE with GCTA-GREML and LDSC as a function of sample overlap ($M = 305,630$ SNPs).** We vary the proportion of sample overlap across $\{0, 0.2, 0.5, 0.8, 0.1\}$. For sample overlap proportion of 0, we have a total of $10,000$ samples where each sample only has observation on one of the traits. For overlap proportion of 1, we have a total $5,000$ samples with each sample having observations on both traits (see Simulations to assess the impact of sample overlap in Materials and Methods). We report the SE of SCORE and LDSC relative to GCTA-GREML. We ran LDSC with in-sample LD. We estimate the standard error of the relative SE using jackknife.

Method	Overlap Proportion	Bias	MSE	SE
LDSC	0	-0.1672	0.1405	0.3355
	0.2	0.0044	0.0921	0.3035
	0.5	0.1708	0.1022	0.2702
	0.8	0.0808	0.0815	0.2737
	1	-0.0955	0.1143	0.3243
SCORE	0	-0.0037	0.0405	0.2013
	0.2	0.2242	0.0753	0.1582
	0.5	0.2843	0.0916	0.1036
	0.8	0.2722	0.087	0.1428
	1	-0.016	0.0266	0.1622
GCTA-GREML	0	-0.0063	0.0383	0.1956
	0.2	0.2635	0.0957	0.1619
	0.5	0.3301	0.1219	0.1137
	0.8	0.2371	0.0788	0.1505
	1	-0.012	0.0244	0.1558

Table 3.8: **Accuracy of SCORE, LDSC, and GCTA-GREML as a function of varying sample overlap corresponding to Figure 3.5**.

h_1^2	h_2^2	ρ_g	ρ_e	Prevalence	$\overline{\rho_g}$	SE	p-value
0.272	0.12	0	0	Continuous trait	0.001	0.018	0.53
0.272	0.12	-0.23	0	Continuous trait	-0.238	0.093	0.39
0.272	0.12	-0.23	0	50%	-0.238	0.097	0.41
0.272	0.12	-0.23	0	25%	-0.239	0.101	0.43
0.272	0.12	-0.23	0	10%	-0.238	0.103	0.44
0.272	0.12	-0.23	0	1%	-0.234	0.124	0.75
0.272	0.12	-0.23	0	0.5%	-0.248	0.134	0.18
0.272	0.12	-0.23	0	0.01%	-0.215	0.205	0.44
0.272	0.12	-0.23	-0.04	Continuous trait	-0.211	0.107	0.08
0.272	0.12	-0.23	-0.04	0.01%	-0.243	0.342	0.70
0.272	0.12	-0.23	0.04	Continuous trait	-0.221	0.089	0.52
0.272	0.12	-0.23	0.04	0.01%	-0.235	0.352	0.89

Table 3.9: **Estimates of ρ_g as a function of the prevalence of binary traits ($N = 291,273$ individuals and $305,630$ SNPs).** We report the average of the point estimates of ρ_g, the SE and p-value of a test of the null hypothesis that the estimates of ρ_g are unbiased. We compute p-values of a test of no bias from the Z-score defined as $\frac{\overline{\rho_g}}{SE/10}$.

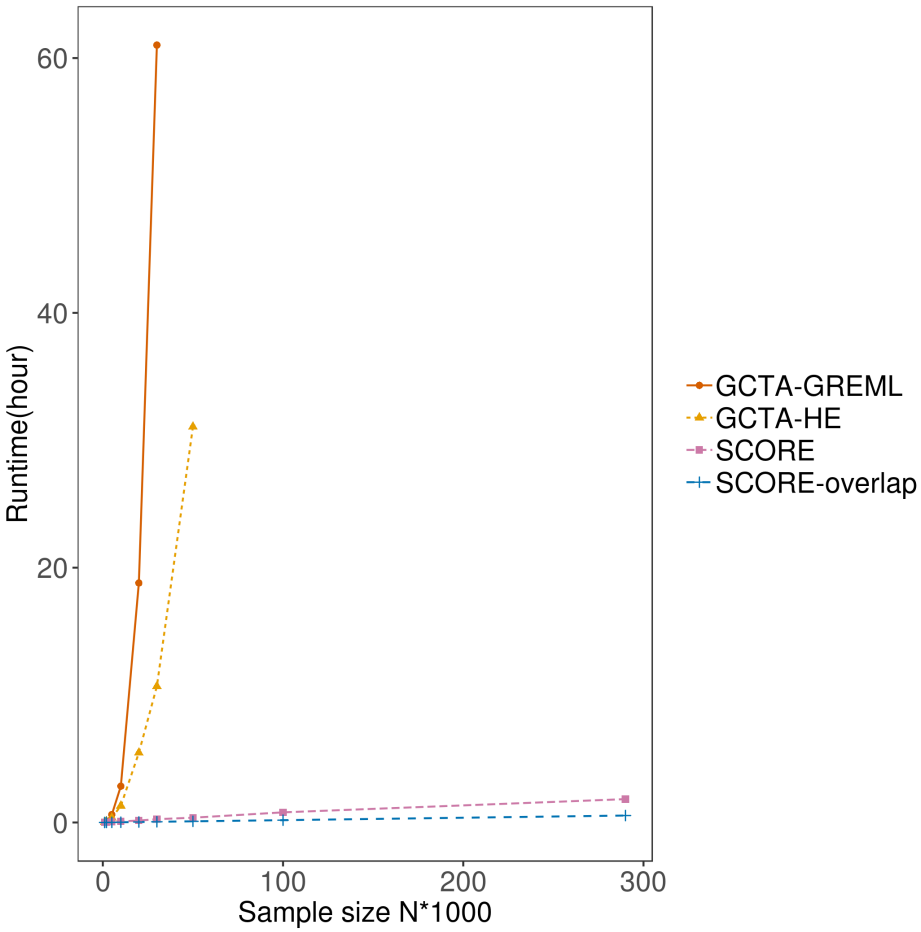

Figure 3.6: **Comparison of the runtime of SCORE with GCTA-GREML and GCTA-HE as a function of the number of sample.**

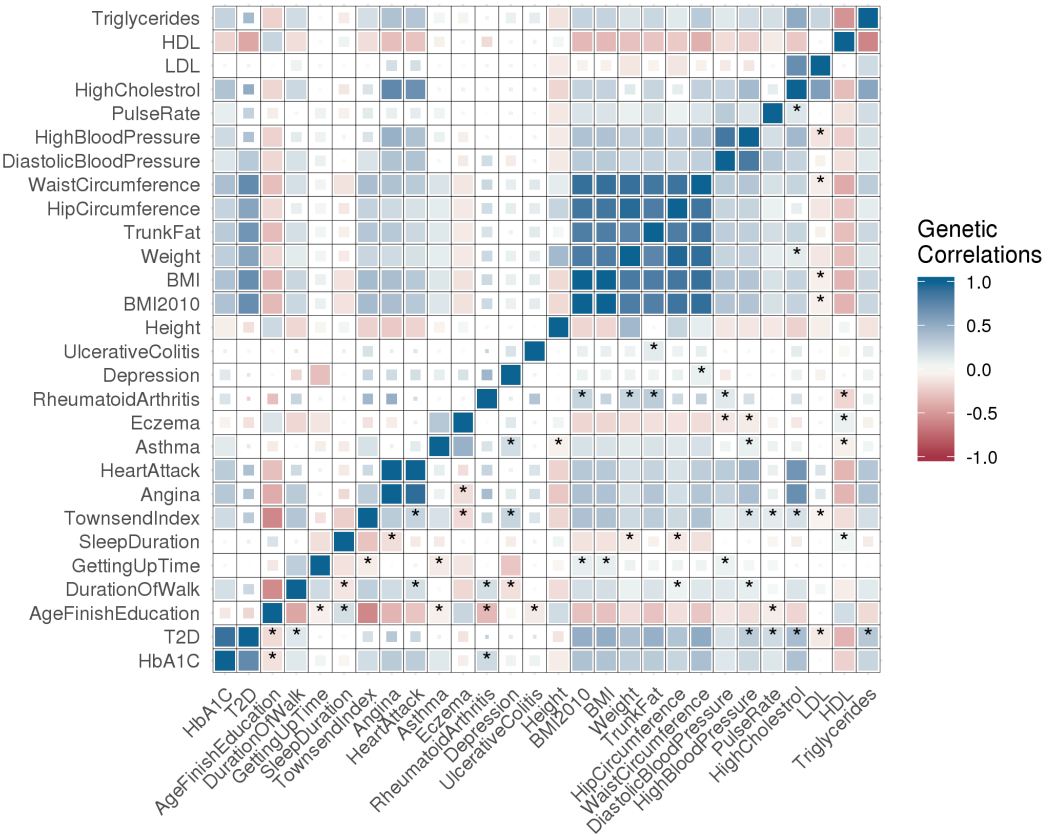

Figure 3.7: **Comparing SCORE with LDSC in 28 traits the UK Biobank.**

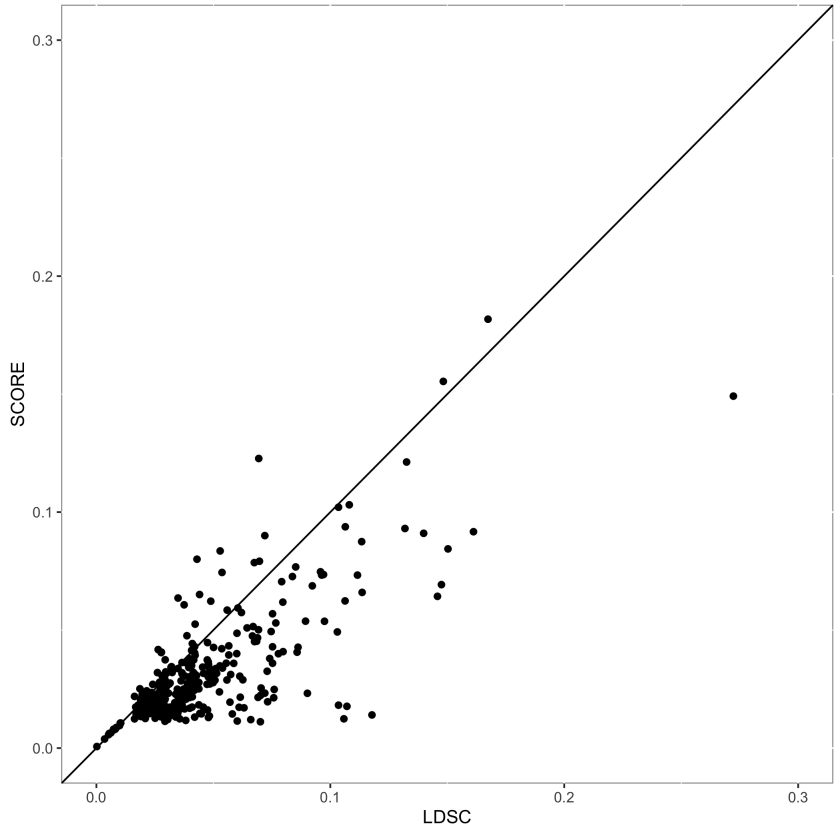

Figure 3.8: **Standard error estimates of genetic correlation between 28 UK biobank phenotypes with LDSC and SCORE corresponding to Figure 3.7.**

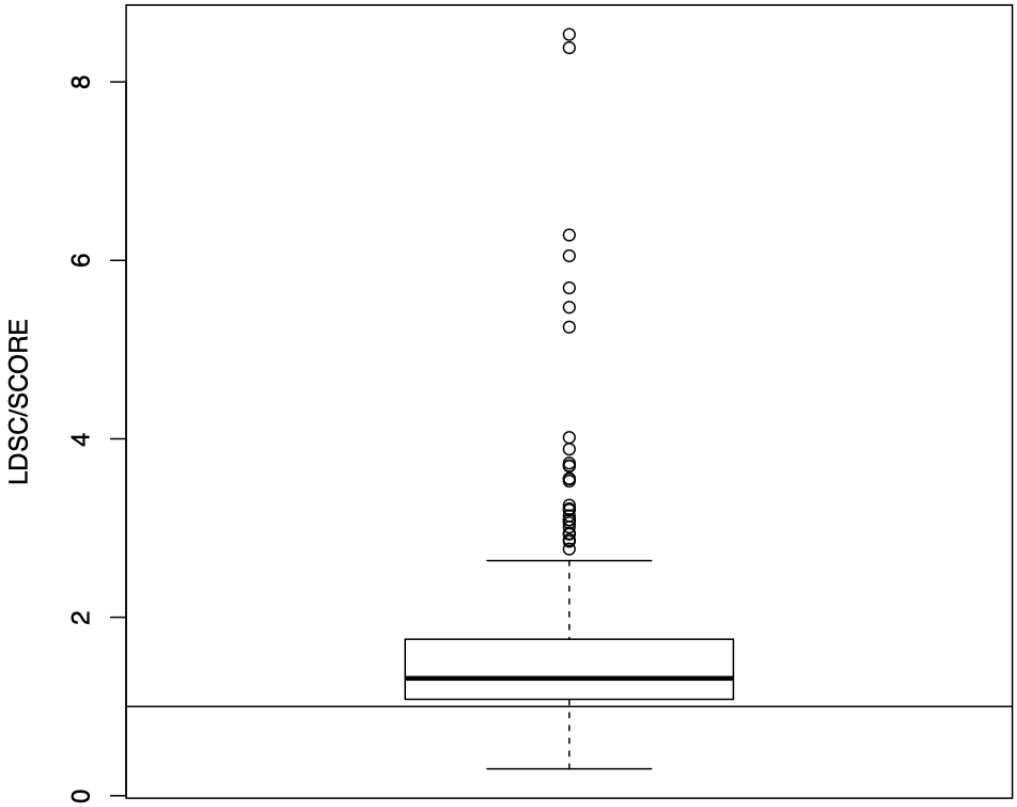

Figure 3.9: **Ratio of standard error estimates of genetic correlation between 28 UK biobank phenotypes with LDSC and SCORE corresponding to Figure 3.7.**

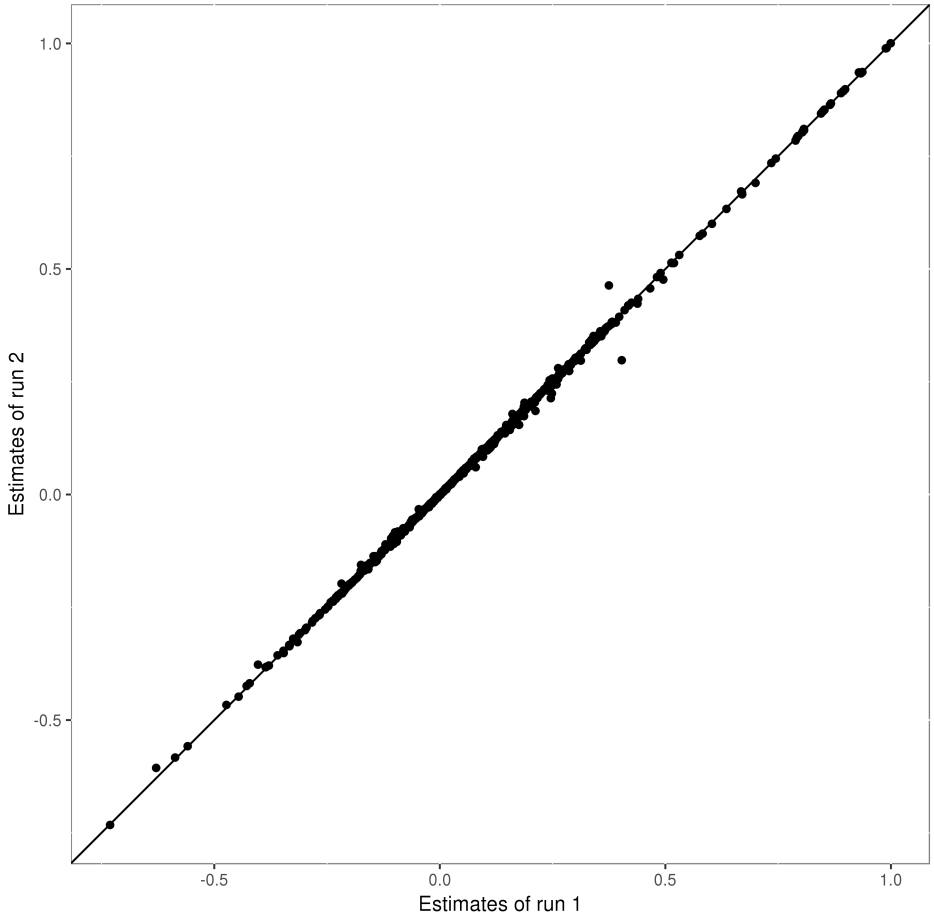

Figure 3.10: **Estimates of genetic correlation in the UK Biobank with different random vectors**.

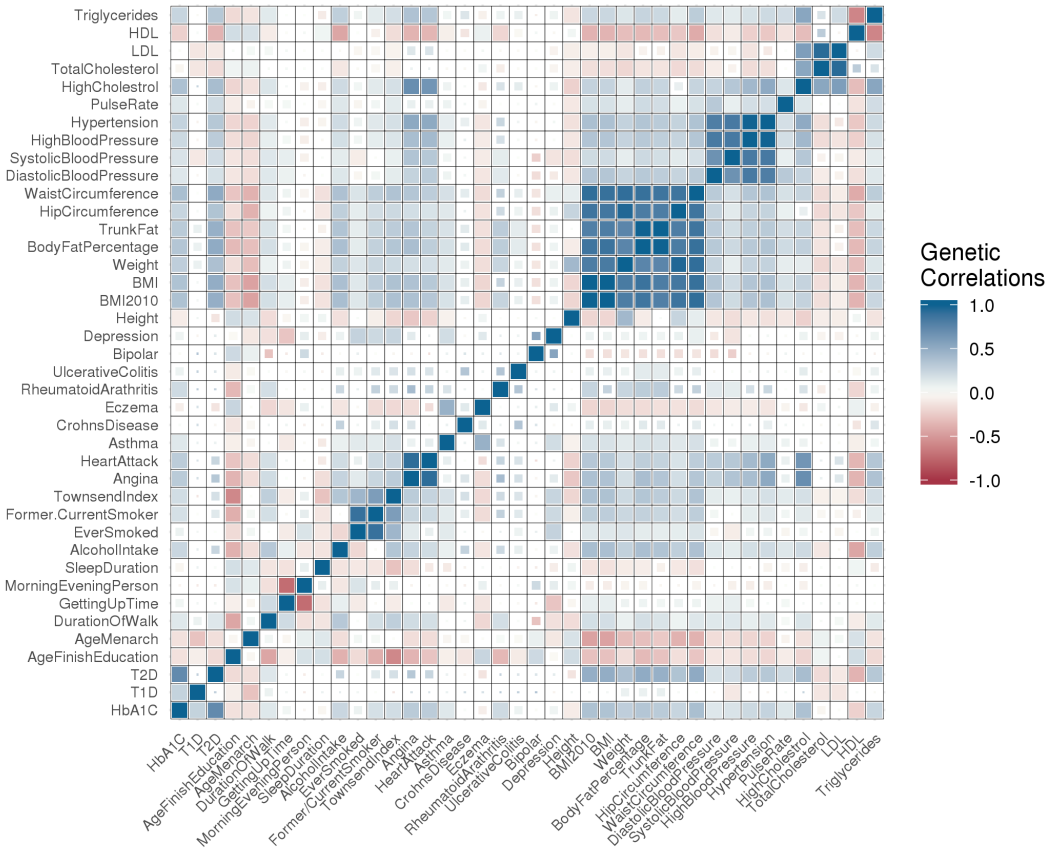

Figure 3.11: **Genetic correlation estimates in the UK Biobank on array SNPs for 40 traits in Table A.1.**

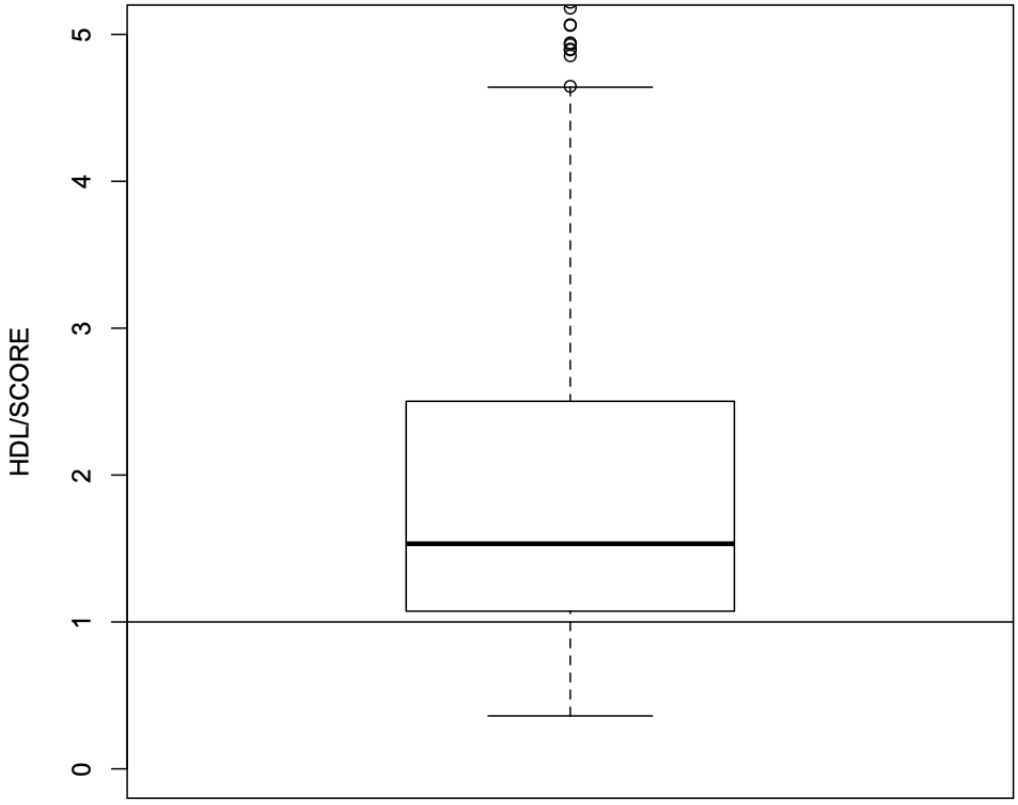

Figure 3.12: **Ratio of standard error estimates of genetic correlation between 40 UK biobank phenotypes with HDL and SCORE corresponding to Table A.1.**

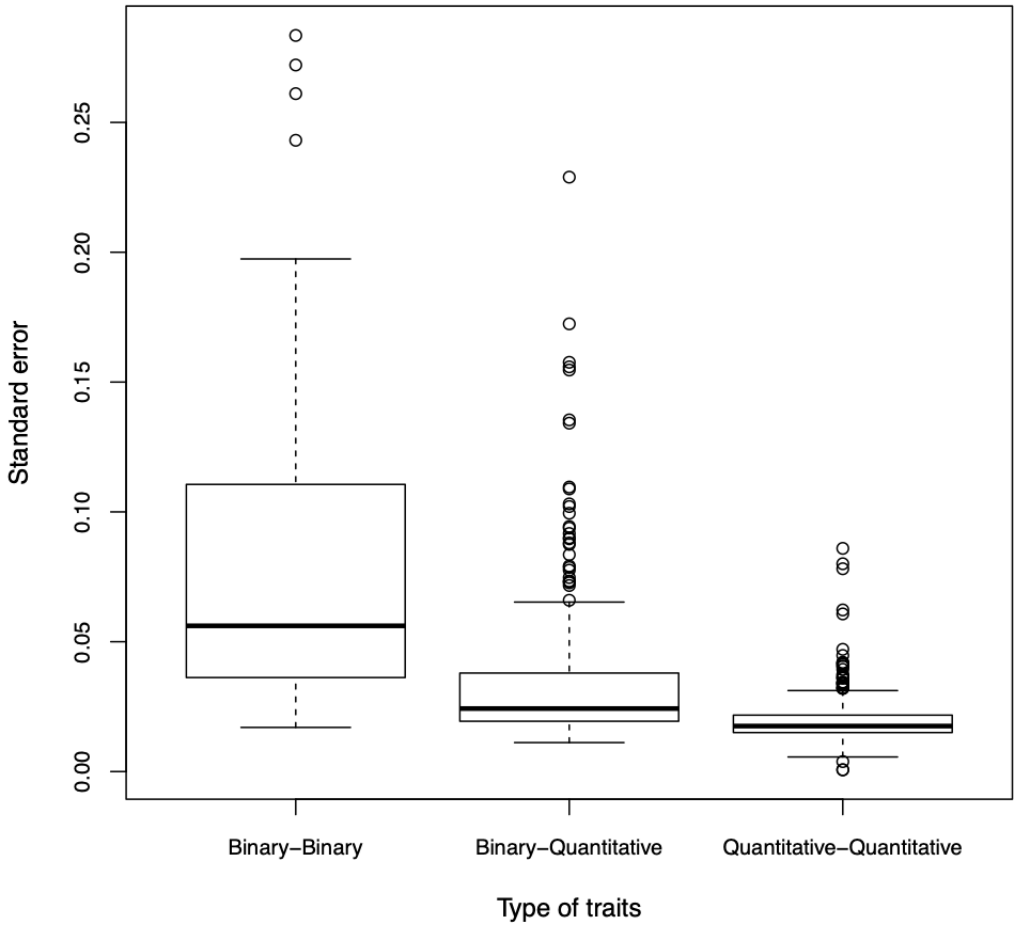

Figure 3.13: **Standard error of genetic correlation estimates from SCORE stratified by the type of phenotype pairs**.

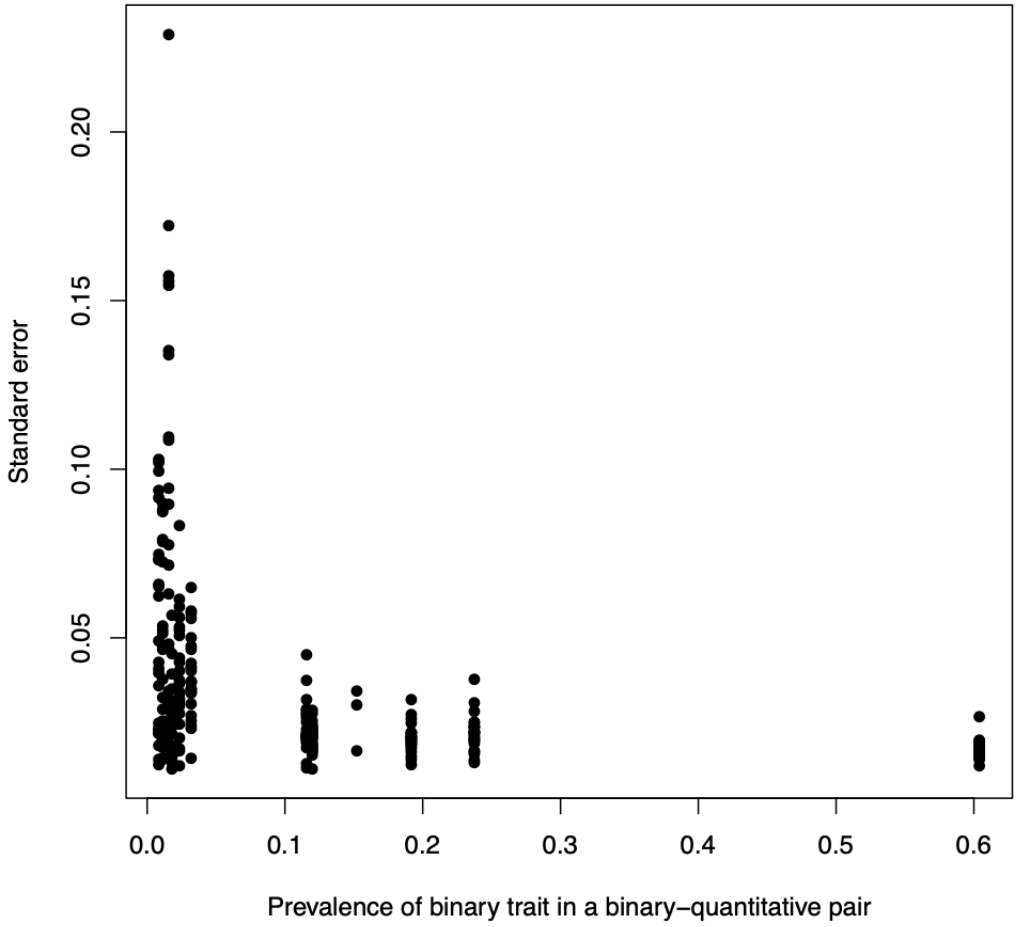

Figure 3.14: **Standard error of genetic correlation estimates from SCORE as a function of the prevalence of the binary phenotype when applied to a pair of phenotypes where one of traits in the pair is binary.**

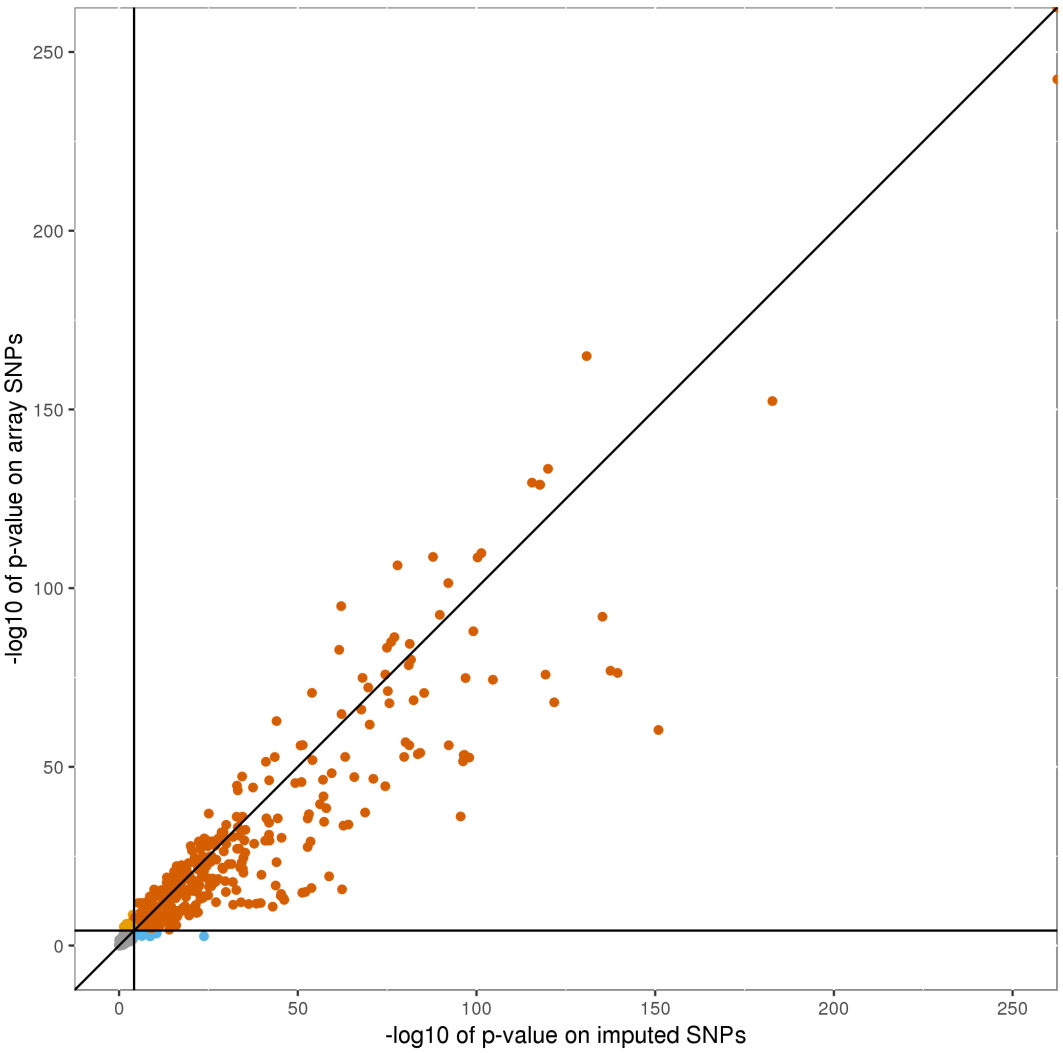

Figure 3.15: **Comparison of the p-values of ρ_g estimates obtained by SCORE in the UK Biobank on imputed versus array SNPs.**

CHAPTER 4

Efficiently partitioning genetic correlation to specific regions of the genome

In Chapter 3, we investigated the shared genetic architecture among traits across the genome through the lens of genome-wide genetic correlation. Given that two traits show substantial genome-wide genetic correlation, a key question of interest is whether the correlation is enriched within certain groups (e.g. a set of genes) or distributed relatively evenly across the genome [11, 12, 10]. In addition, a pair of traits can have significant positive and negative genetic correlations in different regions that lead to a genome-wide genetic correlation of zero [40].

To investigate this question, we start by extending the multivariate model in Chapter 3 into a multi-component model and propose a scalable estimator, SMORE. Having a scalable estimator, we applied SMORE to traits in the UK Biobank to estimate genetic correlation in sets of genes that are identified as expressed in specific tissues by [53]. We investigate the genetic correlation of a few diseases to other complex traits within these gene sets and compared these estimates to genome-wide genetic correlations. Our findings could elucidate the shared biological pathways, highlight disease-relevant tissues, and improve our understanding of the etiology of complex traits and diseases.

4.1 Statistical Models and Estimators

4.1.1 Multivariate Multi-component Linear Mixed Model

Assume we have K phenotypes and each phenotype is associated with the generative model:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{\epsilon_i} \tag{4.1}$$

Here for each phenotype i, we have N_i samples. $N_{i,j}$ are the number of samples that contain measurements for both traits $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ ($N \leq N_i, N \leq N_j$). For each phenotype, we define $\boldsymbol{X}_i$ as the corresponding $N_i \times M$ matrix of standardized genotypes. $\boldsymbol{\beta_i}$ is the vector of SNP effect sizes while $\boldsymbol{\epsilon_i}$ denotes trait-specific environmental noise that is independent of the genetic effect. We assume that SNPs can be assigned to one or more than one of P annotation groups, where in each group p, we have M_p SNPs. Let $\mathbf{1}_p$ be the $M \times 1$ indicator vector indicating if the SNPs belong to group p.

Additionally, we let $\boldsymbol{\beta}$ be the $M \times K$ matrix with columns being the genetic effects and assume that $\mathbb{E}[\boldsymbol{\beta}] = 0$. We assume that $\boldsymbol{\beta}$ is a sum of matrix normal:

$$\boldsymbol{\beta} = \sum_{p=1}^{P} \mathcal{MN}(0, diag(\mathbf{1}_p), \boldsymbol{V}_p) \tag{4.2}$$

where $\boldsymbol{V}_p$ is a $K \times K$ matrix for the p^{th} annotation group:

$$\boldsymbol{V}_p(i,j) = \begin{cases} \sigma^2_{gp,i} & \text{if } i = j \\ \gamma_{gp,ij} & \text{otherwise} \end{cases} \tag{4.3}$$

Here $\sigma_{gp,i}$ denotes the genetic variance component for the p^{th} annotation group for phenotype i, and $\gamma_{gp,ij}$ denotes the genetic covariance between phenotype i and j in annotation p. Importantly, we do not impose any constraints on the grouping of SNPs, *e.g.* the genetic variant could belong to more than one annotation group and could be randomly distributed along the genome or a continuous region. However, we assume that variances and covariance from multiple annotation groups are independent and additive. The genetic correlation parameter for group p between phenotype i and j is defined as $\rho_{gp,ij} = \frac{\gamma_{gp,ij}}{\sqrt{\sigma^2_{gp,i}}\sqrt{\sigma^2_{gp,j}}}$.

Additionally, the trait-specific environmental noise in each individual is assumed to have zero mean and variance of $\sigma^2_{e,i}$ for the i^{th} trait. If an individual has measurements on both trait i, j, the environmental covariance is denoted as $\gamma_{e,ij}$

4.1.2 Method of Moments(MoM) for multi-component multivariate model

Our proposed method SMORE uses a scalable method-of-moments (MoM) estimator for the genetic correlations, $\{\widehat{\gamma_{gp,i}}\}$. SMORE works by finding values of the model parameters, *i.e.*, the group-specific variance components and genetic covariances, such that minimizing the distance between the population and sample moments.

Here we first describe the MoM estimator in the multi-component multivariate model. Since the mean of $\boldsymbol{y_i}, \forall i \in \{1, \ldots, K\}$ are zero, we focus on the covariance. The population covariance of the concatenated phenotypes $\boldsymbol{y} \equiv [\boldsymbol{y_1}^T, \ldots, \boldsymbol{y_i}^T, \ldots, \boldsymbol{y_K}^T]^T$

is now:

$$cov(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^T = \sum_p \begin{bmatrix} \boldsymbol{V}_p(1,1)\boldsymbol{K}_{\boldsymbol{p}}(1,1) & \dots & \boldsymbol{V}_p(1,K)\boldsymbol{K}_{\boldsymbol{p}}(1,K) \\ \dots & \boldsymbol{V}_p(i,j)\boldsymbol{K}_p(i,j) & \dots \\ \boldsymbol{V}_p(K,1)\boldsymbol{K}_{\boldsymbol{p}}(K,1) & \dots & \boldsymbol{V}_p(K,K)\boldsymbol{K}_{\boldsymbol{p}}(K,K) \end{bmatrix}$$
$$+ \begin{bmatrix} \sigma^2_{e,1}\boldsymbol{I}_{N_1} & \dots & \gamma_{e,1K}\boldsymbol{C}(1,K) \\ \dots & \sigma^2_{e,i}\boldsymbol{I}_{N_i} & \dots \\ \gamma_{e,K1}\boldsymbol{C}(K,1) & \dots & \sigma^2_{e,K}\boldsymbol{I}_{N_K} \end{bmatrix} \tag{4.4}$$

Here $\boldsymbol{K}_{\boldsymbol{p}}(i,j) = \frac{\boldsymbol{X_{i,p}}\boldsymbol{X_{j,p}}^T}{M_p}$ is the genetic relatedness matrix (GRM) computed with the SNPs in functional annotation group p. $\boldsymbol{C}(i,j)$ is an indicator matrix, where $\boldsymbol{C}(i,j)_{m,n} = 1$ if the m^{th} entry in $\boldsymbol{y}_i$ and n^{th} entry in $\boldsymbol{y}_j$ are measures on phenotype i and j for the same sample, and 0 otherwise. The MoM estimator is obtained by minimizing the sum of squared differences between the population and empirical covariance:

$$\{\widehat{\gamma_{gp}}\} = argmin_{\{\gamma_{gp}\}}||\boldsymbol{y}\boldsymbol{y}^T - cov(\boldsymbol{y})||^2_F \tag{4.5}$$

with $cov(\boldsymbol{y})$ defined in equation 4.4. The MoM estimator for the genetic covariances $\{\boldsymbol{\gamma}_{gp,ij}\}$ satisfies the normal equations:

$$\begin{bmatrix} tr(\boldsymbol{K}(1,2)^2) & \mathbf{0} & \dots & tr(\boldsymbol{K}(1,2)) & \mathbf{0} & \dots \\ \mathbf{0} & tr(\boldsymbol{K}(i,j)^2) & \dots & \mathbf{0} & tr(\boldsymbol{K}(i,j)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & tr(\boldsymbol{K}(K-1,K)^2) & \mathbf{0} & \dots & tr(\boldsymbol{K}(K-1,K)) \\ tr(\boldsymbol{K}(1,2)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(1,2)) & \dots & \mathbf{0} \\ \mathbf{0} & tr(\boldsymbol{K}(i,j)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(i,j)) & \mathbf{0} \\ \mathbf{0} & \dots & tr(\boldsymbol{K}(K-1,K)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(K-1,K)) \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\gamma}_{gP,12}} \\ \widehat{\boldsymbol{\gamma}_{gP,ij}} \\ \boldsymbol{\gamma}_{gP,\widehat{(K-1)K}} \\ \widehat{\gamma_{e,12}} \\ \widehat{\gamma_{e,ij}} \\ \gamma_{e,\widehat{(K-1)K}} \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{y}_1^T\boldsymbol{K}_P(1,2)\boldsymbol{y}_2 \\ \boldsymbol{y}_i^T\boldsymbol{K}_P(i,j)\boldsymbol{y}_j \\ \boldsymbol{y}_{K-1}^T\boldsymbol{K}_P(K-1,K)\boldsymbol{y}_K \\ \boldsymbol{y}_1^T\boldsymbol{C}(1,2)\boldsymbol{y}_2 \\ \boldsymbol{y}_i^T\boldsymbol{C}(i,j)\boldsymbol{y}_j \\ \boldsymbol{y}_{K-1}^T\boldsymbol{C}(K-1,K)\boldsymbol{y}_K \end{bmatrix} \tag{4.6}$$

where $\boldsymbol{\gamma}_{gP,ij}$ is a vector for the genetic covariance for all P annotation groups between phenotype i and j. Here we define a $P \times P$ matrix $tr(\boldsymbol{K}(i,j)^2)$ where

$tr(\boldsymbol{K}(i,j)^2)_{p,q} = tr(\boldsymbol{K}_p(i,j)\boldsymbol{K}_q(i,j))$. $\boldsymbol{y}_i^T\boldsymbol{K}_P(i,j)\boldsymbol{y}_j$ is a $P \times 1$ vector, where the p^{th} entry is $\boldsymbol{y}_i^T\boldsymbol{K}_p(i,j)\boldsymbol{y}_j = \frac{\boldsymbol{y}_i^T\boldsymbol{X}_{i,p}\boldsymbol{X}_{j,p}^T\boldsymbol{y}_j}{M_p}$, corresponding to annotation group p. Given the coefficients of the normal equations, we can solve analytically for $\widehat{\{\gamma_{gp,ij}\}}, \widehat{\gamma_{e,ij}}$.

Given the MoM estimates of the variance components, the MoM estimate for the genetic correlation for annotation p between phenotype i and j , $i \neq j$ is given by the plug-in estimate:

$$\widehat{\rho_{gp,ij}} = \frac{\widehat{\gamma_{gp,ij}}}{\sqrt{\sigma^2_{gp,i}}\sqrt{\sigma^2_{gp,j}}} \tag{4.7}$$

Due to the block-wise property of the normal equations in equation 4.6, this MoM estimator is equivalent to solving the Bi-variate model between each pair of phenotypes independently.

4.1.3 MoM estimator for Bivariate multi-component model

In this section, we consider only 2 phenotypes, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, and show that the MoM estimator is equivalent to Equation 4.6. Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be the vectors of effect sizes for phenotype 1 and 2. Let $\mathbf{1}_p$ be the $M \times 1$ indicator vector for annotation group p, where $\mathbf{1}_{p,m} = 1$ if m^{th} SNP belongs to group p, and 0 otherwise.

We assume that the effect size has mean of 0, and the effect sizes have the covariance:

$$\begin{aligned} cov(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) &= \sum_{p=1}^{P} \sigma^2_{gp,1} diag(\mathbf{1}_p) \\ cov(\boldsymbol{\beta_2}, \boldsymbol{\beta_2}) &= \sum_{p=1}^{P} \sigma^2_{gp,2} diag(\mathbf{1}_p) \\ cov(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) &= \sum_{p=1}^{P} \gamma^2_{gp,12} diag(\mathbf{1}_p) \end{aligned} \tag{4.8}$$

Here $\sigma^2_{gp,1}$ and $\sigma^2_{gp,2}$ denote the genetic variance associated with trait 1 and 2 for group p respectively. And $\gamma_{gp,12}$ denotes the genetic covariance associated with group p between trait 1 and 2. The trait-specific environmental noise in each individual is assumed to have zero mean and variance σ^2_{et}, $t \in \{1, 2\}$ for trait t.

We assume that phenotypes $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ have the following generative model:

$$\begin{aligned} \boldsymbol{y_1} &= \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_1} \\ \boldsymbol{y_2} &= \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon_2} \end{aligned} \tag{4.9}$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the standardized genotypes for phenotype 1 and 2. Thus the estimator for $\sigma gp, 1^2, \sigma^2_{gp,2}$ The population covariance of $\boldsymbol{y}$ is now:

$$cov(\boldsymbol{y}) = \begin{bmatrix} \sum_p \sigma^2_{p,1} \frac{\boldsymbol{X}_{1,p}\boldsymbol{X}^T_{1,p}}{M_p} & \sum_p \gamma_{gp,12} \frac{\boldsymbol{X}_{1,p}\boldsymbol{X}^T_{2,p}}{M_p} \\ \sum_p \gamma_{gp,12} \frac{\boldsymbol{X}_{2,p}\boldsymbol{X}^T_{1,p}}{M_p} & \sum_p \sigma^2_{p,2} \frac{\boldsymbol{X}_{2,p}\boldsymbol{X}^T_{2,p}}{M_p} \end{bmatrix} \tag{4.10}$$

Now we let $\boldsymbol{y}$ be the concatenated phenotype, $\boldsymbol{y} \equiv [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T]$. We derive the estimator by minimizing:

$$\{\widehat{\gamma_{p,12}, \gamma_{e,12}}\} = argmin_{\gamma_{p,12}, \gamma_{e,12}} ||\boldsymbol{y}\boldsymbol{y}^T - cov(\boldsymbol{y})||^T_F \tag{4.11}$$

Thus the MoM estimator for the genetic covariances for $\{\gamma_{gp,ij}\}$ satisfies the following normal equations:

$$\begin{bmatrix} tr(\boldsymbol{K}(1,2)^2 & tr(\boldsymbol{K}(1,2)) \\ tr(\boldsymbol{K}(2,1)) & tr(\boldsymbol{C}(1,2)) \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\gamma_{gp,12}}} \\ \gamma_{e,12} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \boldsymbol{K}_p(1,2) \boldsymbol{y}_2 \\ \boldsymbol{y}_1^T \boldsymbol{C}(1,2) \boldsymbol{y}_2 \end{bmatrix} \tag{4.12}$$

Here as we defined in the previous section, $tr(\boldsymbol{K}(1,2)^2)$ is a $P \times P$ matrix, where $tr(\boldsymbol{K}(1,2)^2)_{p,q} = tr(\boldsymbol{K}_p(1,2)\boldsymbol{K}_q(1,2)) = tr(\frac{\boldsymbol{X}_{1,p}\boldsymbol{X}^T_{2,p}\boldsymbol{X}_{1,q}\boldsymbol{X}^T_{2,q}}{M_p M_q})$. $\boldsymbol{y}_1^T \boldsymbol{K}_P(1,2)\boldsymbol{y}_2$ is a $P \times 1$ vector, where the p^{th} entry is corresponding to the p^{th} annotation group $\boldsymbol{y}^T \boldsymbol{K}_p(1,2)\boldsymbol{y}_2$. $\boldsymbol{C}(1,2)$ is an indicator matrix, where $\boldsymbol{C}(1,2)_{m,n} = 1$ if the m^{th} entry

in $\boldsymbol{y}_1$ and n^{th} entry in $\boldsymbol{y}_2$ are measurements on the same sample, and 0 otherwise. In other words, $tr(\boldsymbol{C}(1,2))$ equal to the total number of samples that have measurements on both phenotype 1 and 2. This estimator is equivalent to solving all pairs of phenotypes jointly as described in the previous section.

4.1.4 Non-overlapping and overlapping grouping

Assume we have in total P annotation groups. Recall that for each group p, we have an $M \times 1$ indicator vector $\mathbf{1}_p$, $\mathbf{1}_{p,m} = 1$ if m^{th} SNP belongs to group p, and 0 otherwise. If every genetic variant belongs to no more than one annotation group, in other words, if $\sum_p \mathbf{1}_{p,m} \leq 1$, we refer this annotation group has non-overlapping group.

However, if a genetic variant belongs to more than one annotation group, then the annotation groups share some some genetic variants, and we refer this type of annotation as overlapping grouping. For this type of annotation, we need to make an additional assumption that the variance components and covariances are additive among groups. Specifically, the effect size of m^{th} SNP $\boldsymbol{\beta}_m^T$ follows the following multivariate normal distribution:

$$\boldsymbol{\beta_m} \sim \mathcal{N}(0, \sum_p \mathbf{1}_{p,m} \boldsymbol{V}_p) \tag{4.13}$$

where $\boldsymbol{V}_p$ is a $K \times K$ matrix for total K phenotypes defined as following:

$$\boldsymbol{V}_p(i,j) = \begin{cases} \sigma^2_{gp,i} & \text{if } i = j \\ \gamma_{gp,ij} & \text{otherwise} \end{cases} \tag{4.14}$$

Here $\sigma_{gp,i}$ denotes the genetic variance component for the p^{th} annotation group for phenotype i, and $\gamma_{gp,ij}$ denotes the genetic covariance between phenotype i and

j in annotation p. Thus the total variances explained by genetic variance in the annotations for phenotype i is $\sum_p \sigma^2_{gp,i}$, and the total covariance between phenotype i and j explained by the genetic variants in all annotation groups are $\sum_p \gamma_{gp,ij}$

4.1.5 SMORE: Scalable Multivariate multi-component genetic cORrelation Estimator

Naive computation of the MoM estimate of genetic covariance requires computing $tr(\boldsymbol{K}_p(i,j)\boldsymbol{K}_q(i,j))$ for all $p, q \in [1, \dots P]$ and for all pairs $i, j \in [1, \dots K]$. Each entry requires $\mathcal{O}(N_i N_j max\{M_p, M_q\})$ operations, where N_i, N_j are the sample size of each of the traits, and M_p, M_q are the number of genetic variants in an annotation group.

To overcome this computational bottleneck, we replace each $tr(\boldsymbol{K}_p(i,j)\boldsymbol{K}_q(i,j))$ with an unbiased randomized estimate: $tr(\boldsymbol{K}_p\widehat{(i,j)\boldsymbol{K}}_q(i,j))$ [47].

Given B random vectors, $\boldsymbol{z}_1, \dots, \boldsymbol{z}_B$, $z_b \in \mathbb{R}^{N_2}, b \in 1 \dots B$ drawn independently from a distribution with zero mean and identity covariance, our estimator is given by:

$$L_{ij,pq} = tr(\boldsymbol{K}_p\widehat{(i,j)\boldsymbol{K}}_q(i,j)) = \frac{1}{B}\frac{1}{M^2}\sum_b ||\boldsymbol{X_{i,p}}\boldsymbol{X_{j,q}}^T\boldsymbol{z}_b||_2^2$$

The SMORE estimator $(\tilde{\gamma}_g, \tilde{\gamma}_e)$ is obtained by solving Equation 4.6 by replacing each matrix $tr(\boldsymbol{K}(i,j)^2)$ with matrix $\boldsymbol{L}_{ij}$, where $\boldsymbol{L}_{ij}$ is a $P \times P$ symmetric matrix

with the p^{th} row q^{th} column entry being $L_{ij,pq}$.

$$\begin{bmatrix} \boldsymbol{L}_{12}^2 & \mathbf{0} & \dots & tr(\boldsymbol{K}(1,2)) & \mathbf{0} & \dots \\ \mathbf{0} & \boldsymbol{L}_{ij}^2 & \dots & \mathbf{0} & tr(\boldsymbol{K}(i,j)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{L}_{K-1,K}^T & \mathbf{0} & \dots & tr(\boldsymbol{K}(K-1,K)) \\ tr(\boldsymbol{K}(1,2)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(1,2)) & \dots & \mathbf{0} \\ \mathbf{0} & tr(\boldsymbol{K}(i,j)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(i,j)) & \mathbf{0} \\ \mathbf{0} & \dots & tr(\boldsymbol{K}(K-1,K)) & \mathbf{0} & \dots & tr(\boldsymbol{C}(K-1,K)) \end{bmatrix} \begin{bmatrix} \widehat{\gamma_{gp,12}} \\ \widehat{\gamma_{gp,ij}} \\ \widehat{\gamma_{gp,(K-1)K}} \\ \widehat{\gamma_{e,12}} \\ \widehat{\gamma_{e,ij}} \\ \widehat{\gamma_{e,(K-1)K}} \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{y}_1^T \boldsymbol{K}_P(1,2) \boldsymbol{y}_2 \\ \boldsymbol{y}_i^T \boldsymbol{K}_P(i,j) \boldsymbol{y}_j \\ \boldsymbol{y}_{K-1}^T \boldsymbol{K}_P(K-1,K) \boldsymbol{y}_K \\ \boldsymbol{y}_1^T \boldsymbol{C}_{1,2} \boldsymbol{y}_2 \\ \boldsymbol{y}_i^T \boldsymbol{C}_{i,j} \boldsymbol{y}_j \\ \boldsymbol{y}_{K-1}^T \boldsymbol{C}_{K-1,K} \boldsymbol{y}_K \end{bmatrix} \tag{4.15}$$

Computing $L_{ij,pq}$ requires multiplying the genotype matrices $\boldsymbol{X}_{i,p}$ and $\boldsymbol{X}_{j,q}$ with B vectors resulting in a runtime of $\mathcal{O}(max(N_i, N_j)max(M_p, M_q B)$. Regardless of annotation groups are overlap or not, we only need to compute the product of genetic variant to random vector once. Thus the total compute time for $L_{ij,pq} \forall p, q \in [1 \dots P]$ can be computed in time $\mathcal{O}(max(N_i, N_j)MB)$, where M is the total number of genetic variants.

Leveraging the fact that each element of the genotype matrix takes values in the set $\{0, 1, 2\}$, runtime could be further reduce to $\mathcal{O}(max(\frac{N_i}{max(log_3 N_i, log_3 M)}, \frac{N_j}{max(log_3 N_j, log_3 M)})$ MB) [30] (while the standardized genotypes are real-valued, SMORE computes the equivalent quantities by operating on the unstandardized genotype matrix to be able to leverage its discrete entries followed by subtracting the product of the mean of a SNP and random vectors and scaling by MAF). Combined with our previous efficient estimators of the genetic variance components [48, 31], we obtain an efficient estimator of $\{\rho_{gp}\}$.

Further, SMORE uses a streaming algorithm that has scalable memory requirements and uses an efficient block Jackknife with a block size to estimate standard errors with little additional computational overhead. SMORE also provided multi-

threading option, where the computation for each group could be executed in parallel to maximize the computation efficiency.

4.2 Related Work

In order to assess the accuracy of SMORE, we compare the estimates to GREML [36, 25]. To apply GREML, we construct the GRM (genetic relationship matrix) with the genetic variants in each annotation separately. GREML estimates the genetic correlation by solving the bi-variate model with restricted maximum likelihood method [36, 25] in each annotation group separately. We use the estimate of GREML as a baseline and show that SMORE achieves similar accuracy as GREML with statistical efficiency in the following simulations.

A number of recent methods have been proposed for computing the local genetic correlation [39, 54] though these methods do not support arbitrary partitions and are not equivalent to the model we defined, thus not in the scope of comparison.

4.3 Experiments

4.3.1 Accuracy and robustness

We performed simulations on a subset of $10,000$ unrelated white British individuals from the UK Biobank so that all methods compared could be run in a reasonable time. Our simulations used $454,207$ SNPs with minor allele frequency (MAF) above 1% (see Section A.2 for details).

We first consider the case in which annotation groups are not overlapping. Given

the genotypes, we simulated pairs of traits under varying genetic architectures : constant genetic correlation, strong regional genetic correlation, opposite regional genetic correlation, and regional genetic correlation without genome-wide genetic correlation[11]. In each scenario, we assign genetic variants into one of four groups with same probability. We assign the annotation groups in 2 ways: 1. randomly assigned and 2. continuously assigned. With the randomly assigned annotation, each annotation is distributed along the genome, where as continuously assigning cause the annotations in continuous region, with each region being 1/4 of the genome. The true genetic correlations are: $\{0.2, 0.2, 0.2, 0.2\}, \{0.05, 0.05, 0.5, 0.2\}, \{-0.2, 0.2, 0.6, 0.2\}$, $\{-0.2, 0, 0.2, 0\}$. The heritability for two traits are $0.27, 0.12$ uniformly distributed along genome.

We also evaluated the case where annotation groups overlap. We assume two groups with true genetic correlation being $\{-0.1, 0.4\}$. For a given SNP, the probability of belonging to both group is 1/3, and 1/3 probability belonging only to one of the group and 1/3 belonging only to the other one. The heritability of two traits are fixed to $0.27, 0.12$.

In each architecture, we simulated 100 replicates, estimate genetic correlations with SMORE and GREML, and report the bias, mean squared error (MSE) and the standard error for each method in average of the annotation groups in Table 4.1. We observe that SMORE performs as well as GREML with no bias and similar MSE. SMORE is as statistical efficient as likelihood based method GREML.

We also estimate genetic correlations by applying SMORE-SEP to one annotation group at a time. We refer to this approach as SMORE-SEP. We compare the results to the results obtained by estimating the annotation groups jointly. In Table 4.2, we

Table 4.1: **Estimates of bias, mean square error (MSE) and standard error (SE) of genetic correlation estimation methods in simulations.**

architecture	Software	Bias	MSE	SE
constant	SMORE	-0.0061	0.1592	0.1593
	GREML	-0.0063	0.1514	0.1514
zero	SMORE	-0.0001	0.1585	0.1586
	GREML	-0.0001	0.1552	0.1552
enrich	SMORE	-0.0025	0.1552	0.1555
	GREML	-0.0027	0.1523	0.1523
opposite	SMORE	-0.0174	0.153	0.1548
	GREML	-0.0182	0.1514	0.1511
overlap	SMORE	-0.0217	0.1244	0.125
	GREML	-0.021	0.1214	0.121

compare the bias, the mean squared error (MSE) and standard error of estimating the genetic correlation separately and jointly. We observe that estimating genetic correlations separately yields almost identical accuracies to joint estimation.

4.3.2 Power analysis

In this section, we simulate a pair of phenotypes with the full UK Biobank with $291,273$ and $454,207$ SNPs. We randoms assign the genetic variants into two groups, varying the size of the first group being $\{1,000, 2,000, 5,000, 10,000, 20,000, 50,000\}$ with a fixed genetic correlation of -0.2. The rest of variants are assigned to the other group with genetic correlation 0.2. We aim to assess the sufficient group size to identify significant genetic correlation.

For a more realistic setting, we consider two more scenarios, where we randomly

Table 4.2: **Accuracy of genetic correlation estimates when annotations are considered separately and jointly**

architecture	Software	Bias	MSE	SE
constant	SMORE	-0.0061	0.1592	0.1593
	SMORE-SEP	-0.0061	0.1592	0.1593
zero	SMORE	-0.0001	0.1585	0.1586
	SMORE-SEP	-4e-04	0.1584	0.1585
enrich	SMORE	-0.0025	0.1552	0.1555
	SMORE-SEP	-0.0191	0.234	0.4836
opposite	SMORE	-0.0174	0.153	0.1548
	SMORE-SEP	0.0149	0.153	0.1548

picked 2000 genes and 500 genes being the first group, and the rest being the second annotation group. The first annotation group has the true genetic correlation of 0.5, and the second annotation group has the true genetic correlation of -0.2.

For each scenario, we simulate 100 replicates to report the standard error of the genetic correlation estimates on the first annotation group. With the simulations that two annotation groups have a true genetic correlation in opposite directions, we aim to find out the annotation size that has the power to identify the genetic correlation in practice. In Table 4.3, we observe that with a true genetic correlation of 0.2, the annotation with $\geq 10,000$ SNPs has a p-value < 0.05. In the simulations with randomly chosen genes as annotations, we observed that 500 genes ($\sim 7,000$ SNPs) is sufficient to reject the null hypothesis.

In reality, the regions and annotations we are interested in might have a higher true genetic correlation > 0.5, thus fewer SNPs are sufficient.

Table 4.3: **Power analysis for SMORE**

Annotation size (SNPs)	Standard error	p-value
1,000	0.283	0.479
2,000	0.190	0.293
5,000	0.106	0.059
10,000	0.072	0.005
20,000	0.045	8e-06
50,000	0.034	4e-09
Annotation size (genes)	Standard error	p-value
2000	0.036	3e-44
500	0.076	6e-10

4.3.3 False positive rate

In this section, we simulated a pair of phenotypes with the full UK Biobank with $291,273$ and $454,207$ SNPs. Again, we generate annotation groups by randomly assigning the genetic variants into two groups, varying the size of the first group being $\{1,000, 2,000, 5,000, 10,000, 20,000, 50,000\}$, and the rest of the genetic variants as the second group. We fixed the true genetic correlation of the first group to 0 and that of the second group to 0.5. For each annotation, we simulated 100 replicates. With the first annotation group having the true genetic correlation of 0, the goal of this experiment is to assess the probability of rejecting the null hypothesis of no genetic correlation in the first annotation group.

In Table 4.4, we report the false positive rate for the first annotation group. We observe that the false positive rate is controlled ($<= 0.05$), irrespective of the size of

annotation group.

Table 4.4: **SMORE has a controlled false positive rate**

Annotation size	false positive rate
1,000	0.02
2,000	0.06
5,000	0.05
10,000	0.05
20,000	0.03
50,000	0.05

4.4 Functional Annotations

4.4.1 Tissue Specific Annotations

The specifically expressed gene (SEG) annotations are generated following previous work [53]. Given a matrix of normalized gene expression values across genes, [53] computed $t-$statistics for specific expression in the focal tissue for each gene. There are in total 53 annotations, one for each focal tissue. For each focal tissue, we pick the top 2000 genes as the annotation. Unlike in [53], we do not add windows around the genes. In Figure 4.1, we plot the percentage of overlapping genes between all pairs of annotations. Outside of brain tissues, most annotation pairs have an overlap of less than 0.25 (500 genes) while we observe a high overlap in genes that are specifically expressed in brain-related tissues.

4.5 Analysis of the UK Biobank

We applied SMORE to estimate ρ_g for pairs of phenotypes in the UK Biobank across $291,273$ unrelated white British individuals, and $454,207$ SNPs . We computed the genetic correlation within SEG annotations between a focal trait and the rest of the traits from Table A.1 for each of the three diseases chosen as a focal trait.

4.5.1 Focal trait: Depression

In Figure 4.2, we plot a heatmap of the genetic correlation in SEG annotations between depression and the remaining traits from Table A.1. The full squares denote significant genetic correlation after Bonferroni correction for the total number of tests (39×53 pairs). We find several traits that have significant genetic correlation with depression: asthma, ever smoked, former/current smoker, easiness getting up, and Townsend index. Between depression and smoking-related traits (ever smoked, former/current smoker), the signals are found in genes specifically expressed in the lung. Between depression and the traits with significant genetic correlations, the signals are found in prostate and breast mammary tissue, which could suggest a gender effect that has not been completely removed by including gender as a covariate in the analysis.

4.5.2 Focal trait: Autoimmune diseases

In this section, we take autoimmune diseases (asthma and eczema) as focal traits. In Figure 4.3, we plot the heatmap of the genetic correlation in SEG annotations between asthma and the rest 39 from Table A.1. Similarly, we plot the heatmap for

eczema for in Figure 4.4.

We observe that, as asthma and eczema are both autoimmune diseases with a strong genome-wide genetic correlation of $0.475(pval = 2.55e-41)$ 3.11, there are significant positive genetic correlations in 22 out of total 53 tissues. Notably, there are significant genetic correlations in all the gene sets that are specifically expressed in brain-related tissues.

Despite the strong positive genetic correlation between Asthma and Eczema, the genetic correlation pattern of each of the diseases versus other 38 traits sometimes show opposite patterns. For instance, while asthma has positive genetic correlations with anthropometric traits (*e.g.*, Trunk fat and Body fat percentage) while eczema has overall negative genetic correlations with significant signals too.

We also observe signals of genetic correlations in gene sets that are not reflected in genome-wide analyses. Although there is no significant genome-wide genetic correlation between eczema and hypertension (-0.088, p=$6.2e-04$) 3.11, we found significant genetic correlation within the genes specifically expressed in artery aorta tissue. This is an example of the utility of regional genetic correlation that is not apparent in genome-wide genetic correlation analyses.

4.5.3 Focal trait: Type 2 Diabetes

In Figure 4.5, we plot a heatmap of the genetic correlation in tissue specifically expressed gene sets between asthma and the remaining 39 traits from Table A.1.

For type 2 diabetes, we observe strong genetic correlation signals in the tissues that have been reported relevant to the disease: skeleton muscle, liver, kidney, and intestines. We also observe strong genetic correlations to HbA1C in all brain-related

tissues, and heart-left-ventricle tissue. We observe some gene sets that are specifically expressed in digestive-related tissues (*e.g.*, colon transverse, minor salivary gland tissues) that harbor significant genetic correlation with anthropometric traits (eg. weight, trunk fat, body fat percentage).

4.6 Discussion

We have introduced a multivariate multi-component LMM to estimate genetic correlation in specific genomic annotations and proposed an efficient estimator SMORE. In simulations, we have shown that SMORE is unbiased and nearly as statistically efficient as a maximum likelihood estimator while being highly scalable. SMORE has a controlled false positive rate across these simulations and enough power to detect signals of genetic correlation provided the annotation of interest has an adequate number of variants.

In the application of SMORE to traits in the UK biobank, we identified a significant genetic correlation between eczema and hypertension in genes that are expressed in specific tissues even when the traits do not have significant genome-wide genetic correlation. There are many other ways to define annotations of interest including based on population genomic principles (minor allele frequency range) or other functional genomic data. SMORE offers a powerful tool to perform arbitrary queries to identify pleiotropy within genomic regions and annotations.

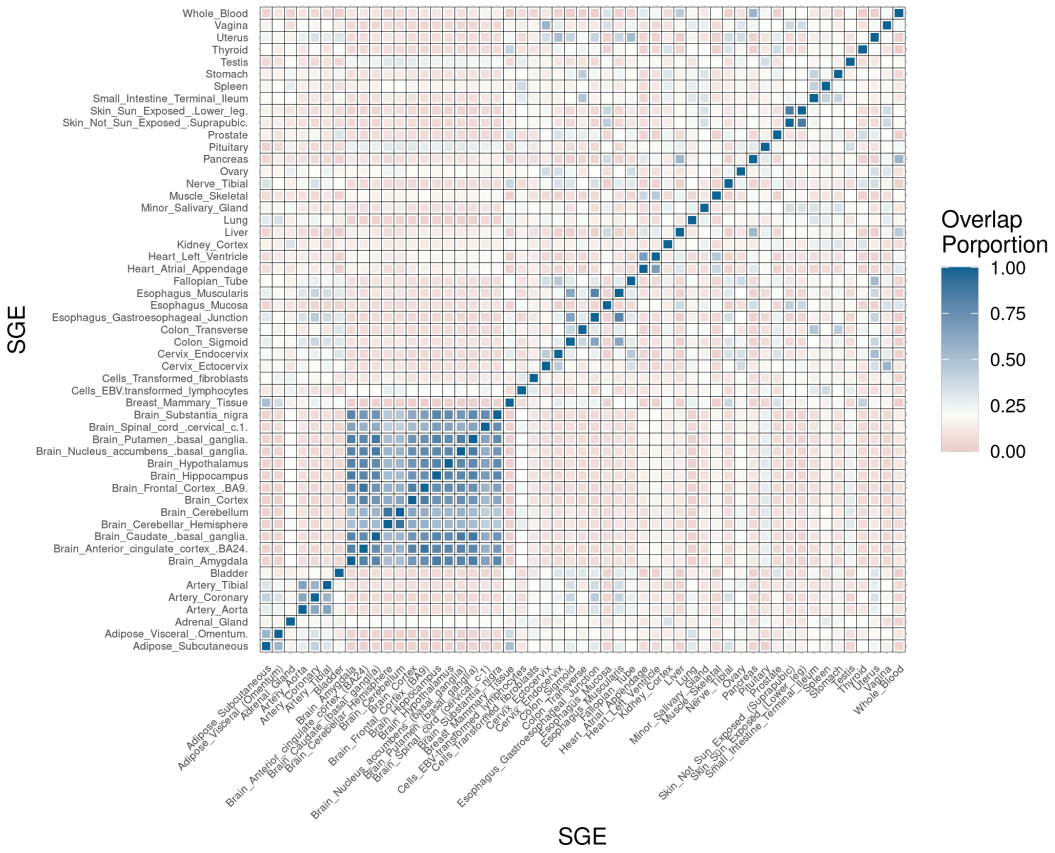

Figure 4.1: **Gene overlap between specific gene expression annotations across tissues.**

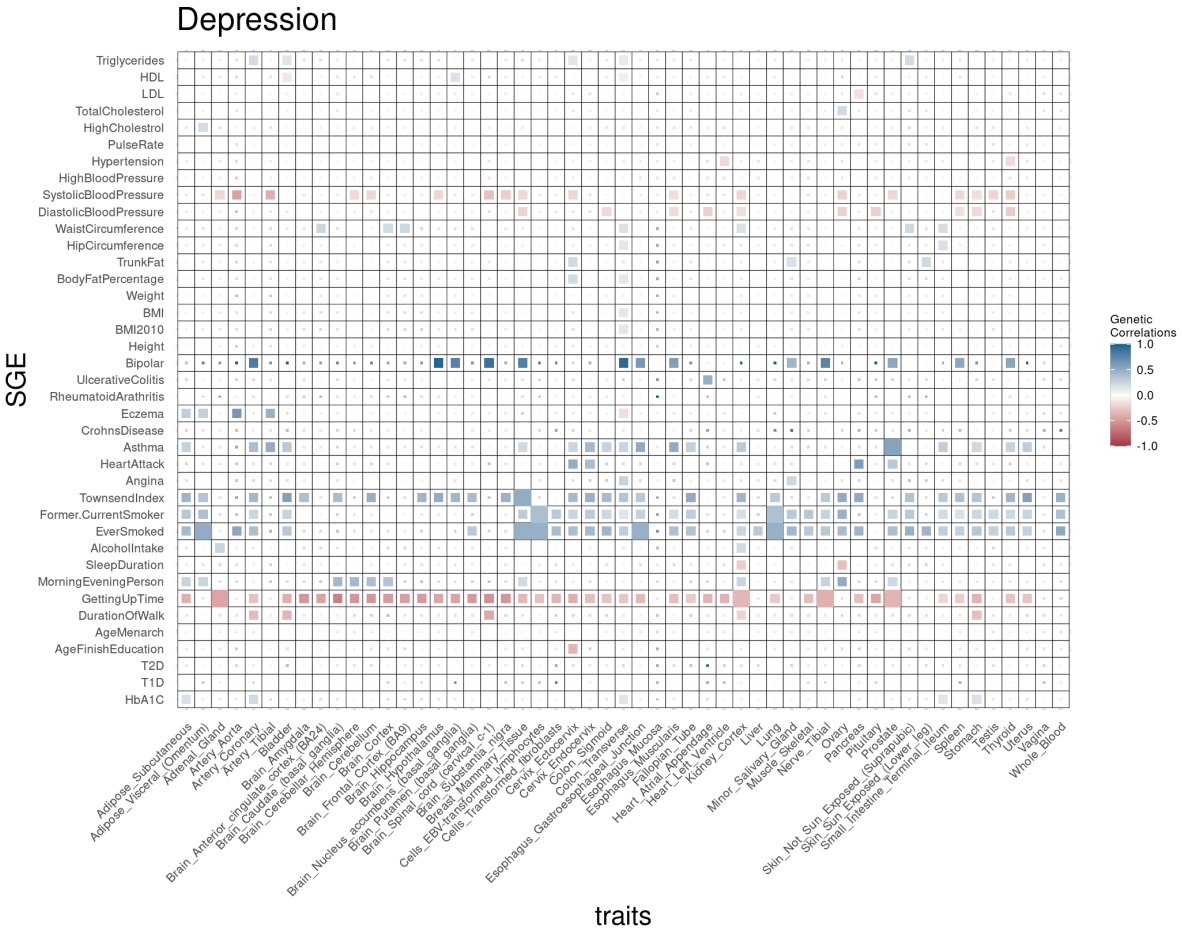

Figure 4.2: **Genetic correlation between Depression and other traits in tissue specifically expressed genes**.

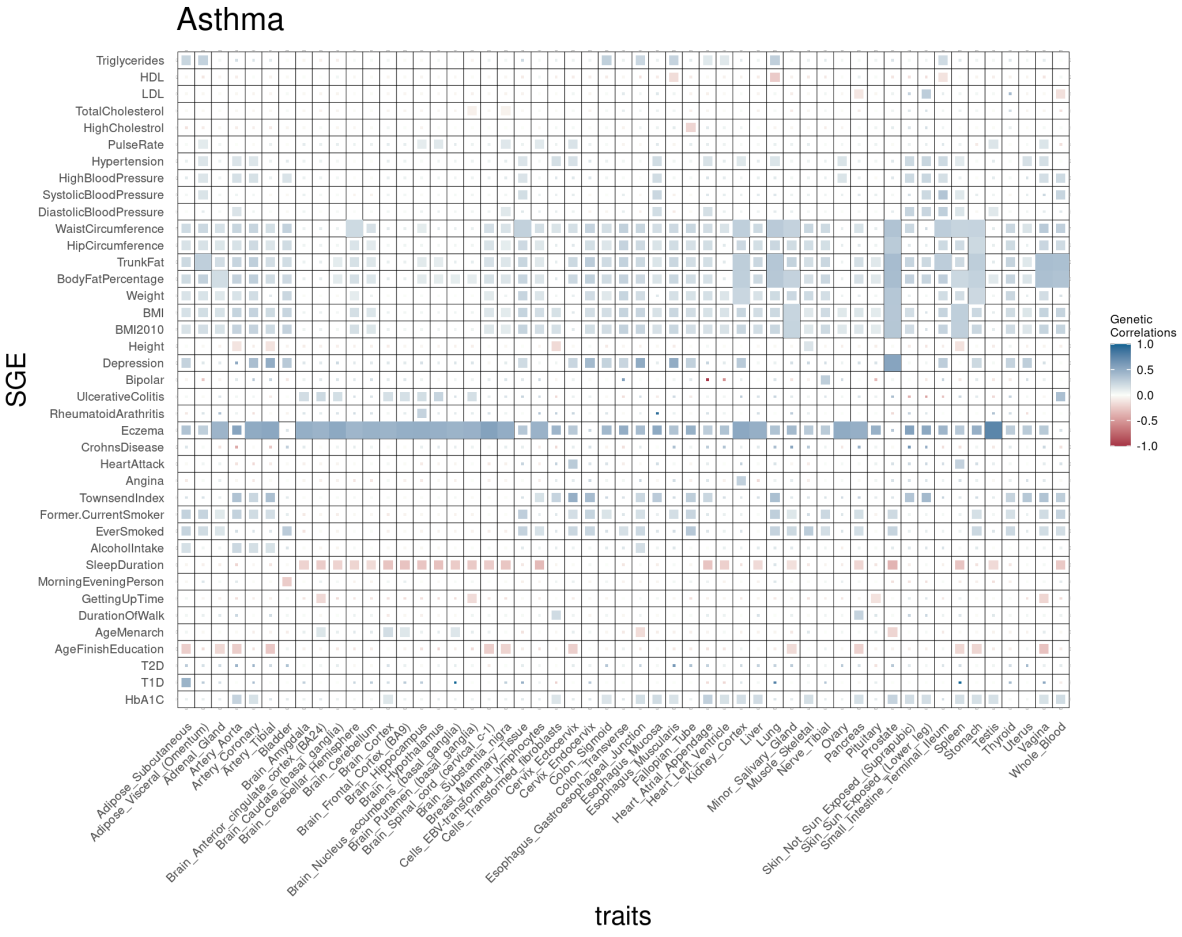

Figure 4.3: **Genetic correlation between Asthma and other traits in tissue specifically expressed genes**

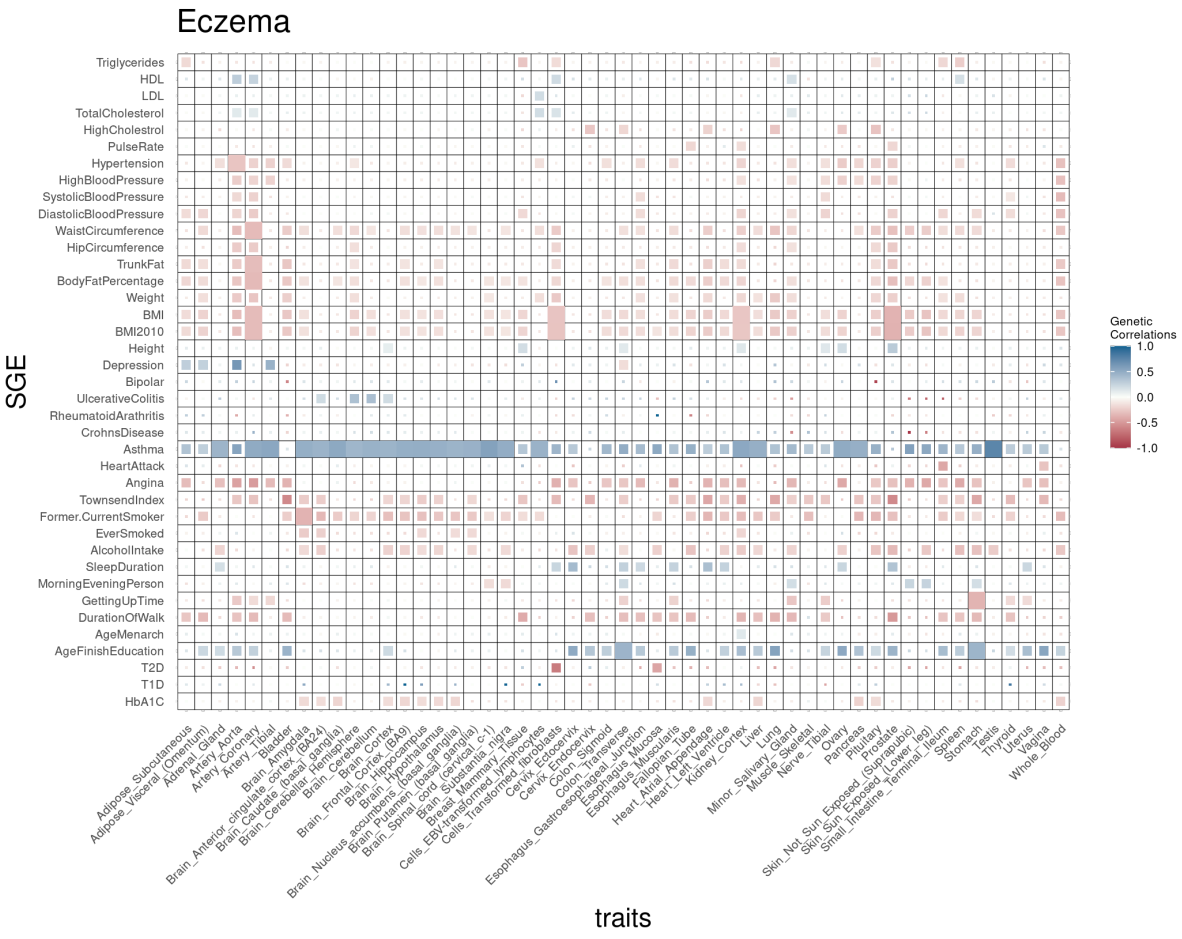

Figure 4.4: **Genetic correlation between Eczema and other traits in tissue specifically expressed genes**

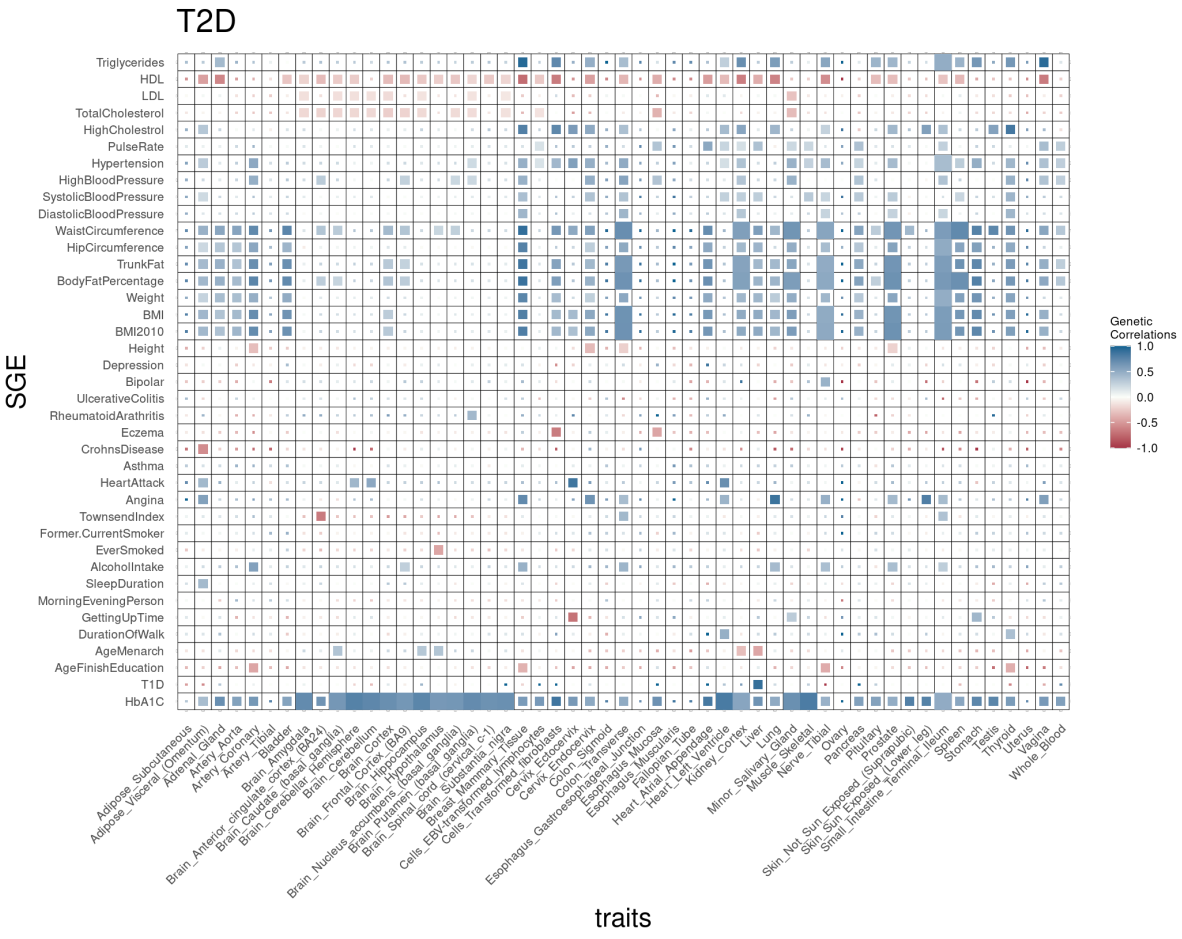

Figure 4.5: **Genetic correlation in tissue specifically expressed gene set for Type 2 Diabetes and other traits**.

CHAPTER 5

Conclusions

5.1 Contributions

In this thesis, we started with a basic statistical model – a linear mixed model (LMM) – that can be used to estimate the SNP heritability of a single trait. We then extended the LMM into bi-variate and multivariate settings that allow for more than one trait to be considered allowing for estimates of genetic correlation. We then considered models that are more flexible – relaxing the assumption that the genetic architecture is uniform along the genome.

For each of these statistical models, we proposed scalable randomized method-of-moments (MoM) estimators. These estimators are derived by equating the first two sample moments to population moments. Importantly, these estimators can be computed efficiently on genotype datasets containing hundreds of thousands of individuals and millions of genetic variants. This computational efficiency stems from the property that, instead of computing the exact genetic relationship matrices (GRM) (as in MoM estimators), our estimators use random vectors to approximate the trace of the GRM and the square of the GRM. This approximation leads to the runtime of these estimators having a linear scaling with the sample size and the number of genetic variants. The computation time is further reduced to sub-linear in

the sample size and number of genetic variants by utilizing the property of genotype matrix, that the entries are discrete and take values from $\{0, 1, 2\}$.

Having developed a set of scalable estimators, we analyze the genotypes and traits in the UK Biobank to find novel signals. The tools we have developed are extremely flexible and allow users to query arbitrary sets of traits and genomic regions in large biobank datasets.

5.2 Future Directions

Our estimators require access to individual genotypes. This requirement makes it challenging to apply to biobanks that do not make their genotypes available. Extending these estimators to the setting where only summary statistics are available is an important direction for future work.

By modeling multiple traits within a region or annotation, we can also attempt to answer the question: within this region or annotation of interest, to what extent is the genetic signal of a trait explained by that of several predictor phenotypes? This leads us to derive the notion of group-specific conditional genetic correlation.

Given the annotation group p, the correlation between two phenotypes, $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, conditional on a set of other phenotypes $\boldsymbol{y}_Z, i \notin Z, j \notin Z$, is denoted by $\rho_{gp,ij|Z}$. This could be computed from $\boldsymbol{V}_p$, as defined in equation 4.4, the effect size $\boldsymbol{\beta}$ is row-wise independent and identically distributed. The partial covariance $\gamma_{gp,ij|Z} = \boldsymbol{V}_p(i,j) - \boldsymbol{V}_{p,iZ}\boldsymbol{V}_{p,ZZ}^{-1}\boldsymbol{V}_{p,Zj}$, where $\boldsymbol{V}_{p,iZ}$ is a submatrix from $\boldsymbol{V}_p$ between phenotype i and phenotypes in Z. $\boldsymbol{V}_{p,ZZ}$ is the submatrix of $\boldsymbol{V}_p$ among phenotypes in Z only. And the partial variance $\sigma^2_{gp,i|Z} = \sigma_{gp,i} - \boldsymbol{V}_{p,iZ}\boldsymbol{V}_{p,ZZ}^{-1}\boldsymbol{V}_{p,Zi}$. So we could compute the

partial correlation specific for annotation group p as $\rho_{gp,ij|Z} = \frac{\gamma_{gp,ij|Z}}{\sqrt{\sigma^2_{gp,i|Z}}\sqrt{\sigma^2_{gp,j|Z}}}$. We can thus estimate the group-specific conditional genetic correlation using the estimates from SMORE.

APPENDIX A

Details on the UK Biobank dataset

A.1 Phenotypes in the UK Biobank

In Table A.1, we list the phenotypes in the UK Biobank that we analyze in this thesis. The phenotypes could be classified into 9 groups: glucose metabolism and diabetes, socioeconomic and general medical information, environmental factors, coronary artery disease related, autoimmune disorder, psychiatric disorders, anthropocentric, blood pressure and circulatory, and lipid metabolism.

These traits were chosen to be representative of different phenotypic categories. Further, these traits have low missingness ($< 30\%$), and high prevalence for binary traits ($> 0.5\%$). The typical approaches for dealing with missing data consist of either omitting the missing sample in the analysis or imputing the missing entry. Each of these approaches can lead to reduced power to detect the genetic signals or can bias estimates. Thus, we focus on phenotypes with low missingness in this thesis. In these 40 phenotypes, there are 14 binary traits, 3 categorical traits, and 23 continuous traits. The binary (e.g. disease status) and categorical traits (e.g. alcohol intake) are treated as continuous in our analyses. While not strictly justified, our empirical results show that treating these traits as continuous do not introduce any substantial bias.

Category	Trait	Field ID
Lipid metabolism traits	Triglycerides	30870
Lipid metabolism traits	HDL	30760
Lipid metabolism traits	LDL	30780
Lipid metabolism traits	Total Cholesterol	30690
Lipid metabolism traits	High Cholesterol	1473
Blood pressure and circulatory traits	Pulse Rate	102
Blood pressure and circulatory traits	Hypertension	1065
Blood pressure and circulatory traits	High Blood Pressure	6150
Blood pressure and circulatory traits	Systolic Blood Pressure	4080
Blood pressure and circulatory traits	Diastolic Blood Pressure	4079
Anthropometric traits	Waist Circumference	48
Anthropometric traits	Hip Circumference	49
Anthropometric traits	Trunk Fat	23127
Anthropometric traits	Body Fat Percentage	23099
Anthropometric traits	Weight	23098
Anthropometric traits	BMI	21002
Anthropometric traits	BMI2010	23104
Anthropometric traits	Height	50
Psychiatric disorders	Depression	20002, 41270, 20544

Psychiatric disorders	Bipolar	20002, 41270, 20544
Autoimmune disorders	Ulcerative Colitis	20002, 41270
Autoimmune disorders	Rheumatoid Arthritis	20002, 41270
Autoimmune disorders	Eczema	1452, 6152
Autoimmune disorders	Crohn's Disease	1462
Autoimmune disorders	Asthma	1111, 6152
Coronary artery disease related traits	Heart Attack	1075, 6150
Coronary artery disease related traits	Angina	1074, 6150
Environmental factor traits	Townsend Index	189
Environmental factor traits	Former/Current Smoker	20116
Environmental factor traits	Ever Smoked	20160
Environmental factor traits	Alcohol Intake	1558
Socioeconomic and general medical information traits	Sleep Duration	1169
Socioeconomic and general medical information traits	Morning Evening Person	1180
Socioeconomic and general medical information traits	Easiness of Getting up	1170
Socioeconomic and general medical information traits	Duration Of Walk	874
Socioeconomic and general medical information traits	Age Menarch	2714

Socioeconomic and general medical information traits	Age Finish Education	845
Glucose metabolism and diabetes traits	T2D	20002, 2976, 41270
Glucose metabolism and diabetes traits	T1D	1222
Glucose metabolism and diabetes traits	HbA1C	30750

Table A.1: **UK Biobank traits analyzed in this work**

A.2 Quality control for genotypes

We restricted most of our analysis to SNPs genotyped on the UK Biobank Axiom array, filtering out markers that had a high missingness rate ($> 1\%$) and low minor allele frequency ($< 1\%$), and we exclude the major histocompatibility complex (MHC) region. Moreover, SNPs that fail the Hardy-Weinberg Equilibrium (HWE) test at significance threshold 10^{-7} were removed. We also filter the samples that have a genetic kinship with any other sample (samples having any relatives in the dataset using the field 22021: Genetic kinship to other participants) and restricted the study to samples with self-reported British white ancestry (field 21000 with coding 1001). After quality control, we obtained $291,273$ individuals and $454,207$ SNPs.

We performed similar quality control on the imputed genotypes in the UK Biobank: filtering out markers with high missingness rate ($> 1\%$), low MAF ($< 1\%$), with HWE p-value $< 1 \times 10^{-7}$, and fall within the MHC region.

After quality control, we obtained $4,824,392$ SNPs.

A.3 Covariates

In this thesis, unless otherwise mentioned, all analyses conducted on UK Biobank are corrected for the following covariates: age, gender, principal components $1-10$, assessment center, and genotype measurement batch. We treat covariates as fixed effects in our methods. For running summary statistics-based method, e.g. LDSC, HDL, the covariates are included while generating the summary statistics.

A.4 Data processing

LD scores were computed from $305,630$ SNPs chosen for the simulations. The LD scores were computed from a random subset of $50,000$ individuals in the UK Biobank (the individuals used in our simulations were a subset of the $50,000$ individuals used to compute LD score). For analysis of UK Biobank data, LD scores were computed on $459,792$ SNPs common SNPs (MAF $> 1\%$) present on the UK Biobank Axiom array. LD scores were computed using flags $--l2$ and $--ld-wind-kb2000.0$.

Summary statistics input to LDSC were generated using PLINK. We used linear regression to generate summary statistics for continuous traits and categorical traits and logistic regression for binary traits. In computing summary statistics for traits in the UK Biobank, we include the following covariates: age, gender, principal components 1-10, assessment center, and genotype measurement batch. We used the same covariates as input to SCORE.

We ran LDSC under default settings with an unconstrained intercept.

APPENDIX B

Appendix to: A scalable estimator of SNP heritability for Biobank-scale data

B.1 Randomized Estimator of trace of a Matrix

For a $N \times N$ matrix, $\boldsymbol{A}$, a randomized estimator of $\text{tr}\,[\boldsymbol{A}]$ is $\widehat{\text{tr}\,[\boldsymbol{A}]} \equiv \frac{1}{B}\sum_b \boldsymbol{z_b}^T \boldsymbol{A} \boldsymbol{z_b}$, where $\boldsymbol{z_b}$ are i.i.d. random vectors with each entry drawn from a standard normal distribution. To see this:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}] &= \mathbb{E}[tr(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z})] \quad \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} \texttt{ is a scalar} \\
&= \mathbb{E}[\text{tr}\left[\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{A}\right]] \quad \texttt{cyclic property of the trace} \\
&= \text{tr}\left[\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{A}]\right] \quad \texttt{trace and expectation are linear} \\
&= \text{tr}\left[\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{A}\right] \; \boldsymbol{A} \texttt{ is fixed} \\
&= \text{tr}\,[\boldsymbol{A}] \quad \texttt{using the distributional assumptions on } \boldsymbol{z}
\end{aligned}$$

B.2 Bias of the RHE-reg Estimator

Our estimator of $\text{tr}\left[\boldsymbol{K}^2\right]$ is $L_B \equiv \widehat{tr[K^2]} = \frac{1}{B}\sum_B \boldsymbol{z}_b^T\boldsymbol{K}\boldsymbol{K}\boldsymbol{z}_b$. The RHE-reg estimators for (σ_g^2, σ_e^2) are given by: $\begin{bmatrix}\widetilde{\sigma}_g^2 \\ \widetilde{\sigma}_e^2\end{bmatrix} = \boldsymbol{A}^{-1}\begin{bmatrix}\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y}\end{bmatrix}$ where $\boldsymbol{A} = \begin{bmatrix}L_B & N \\ N & N\end{bmatrix}$.

We first compute the expectation of this estimator :

$$\mathbb{E}\begin{bmatrix}\widetilde{\sigma}_g^2\\ \widetilde{\sigma}_e^2\end{bmatrix} = \mathbb{E}[\boldsymbol{A}^{-1}\begin{bmatrix}\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}\\ \boldsymbol{y}^T\boldsymbol{y}\end{bmatrix}]$$

$$= \mathbb{E}[\boldsymbol{A}^{-1}]\mathbb{E}\begin{bmatrix}\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}\\ \boldsymbol{y}^T\boldsymbol{y}\end{bmatrix}] \quad \texttt{since random vectors } \boldsymbol{z}_b \texttt{ and } \boldsymbol{y} \texttt{ are independent}$$

We know that $\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] = cov(\boldsymbol{y}) = \sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}$. We can compute $\mathbb{E}[\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}]$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}] &= \mathbb{E}[\mathrm{tr}\left[\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}\right]] \quad \boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \texttt{ is a scalar}\\
&= \mathbb{E}[\mathrm{tr}\left[\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{K}\right]] \quad \texttt{cyclic property of the trace}\\
&= \mathrm{tr}\left[\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{K}]\right] \quad \texttt{expectation and trace are linear}\\
&= \mathrm{tr}\left[\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T]\boldsymbol{K}\right] \quad \texttt{as } \boldsymbol{K} \texttt{ is constant}\\
&= \mathrm{tr}\left[\sigma_g^2\boldsymbol{K}^2 + \sigma_e^2\boldsymbol{K}\right]\\
&= \sigma_g^2\mathrm{tr}\left[\boldsymbol{K}^2\right] + N\sigma_e^2 \quad \texttt{using } \mathrm{tr}\,[K] = N
\end{aligned}$$

And for $\mathbb{E}[\boldsymbol{y}^T\boldsymbol{y}]$, we have;

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}^T\boldsymbol{y}] &= \mathbb{E}[\mathrm{tr}\left[\boldsymbol{y}^T\boldsymbol{y}\right]] \quad \boldsymbol{y}^T\boldsymbol{y} \texttt{ is a scalar}\\
&= \mathbb{E}(\mathrm{tr}\left[\boldsymbol{y}\boldsymbol{y}^T\right]] \quad \texttt{cyclic property of the trace}\\
&= \mathrm{tr}\left[\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T]\right] \quad \texttt{expectation and trace are linear}\\
&= \mathrm{tr}\left[\boldsymbol{K}\right]\sigma_g^2 + N\sigma_e^2\\
&= N\sigma_g^2 + N\sigma_e^2
\end{aligned}$$

Defining $b \equiv \mathbb{E}[\frac{1}{L_B - N}]$ and computing $\boldsymbol{A}^{-1} = \begin{bmatrix}\frac{1}{L_B-N} & \frac{-1}{L_B-N}\\ \frac{-1}{L_B-N} & \frac{L_B}{N(L_B-N)}\end{bmatrix}$, we have

$$\mathbb{E}\begin{bmatrix}\widetilde{\sigma_g^2}\\ \widetilde{\sigma_e^2}\end{bmatrix} = \mathbb{E}[\boldsymbol{A}^{-1}]\mathbb{E}[\begin{bmatrix}\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}\\ \boldsymbol{y}^T\boldsymbol{y}\end{bmatrix}]$$
$$= \begin{bmatrix} b & -b \\ -b & \frac{1}{N}+b\end{bmatrix}\begin{bmatrix}\text{tr}\left[\boldsymbol{K}^2\right]+N\sigma_e^2\\ N\sigma_g^2+N\sigma_e^2\end{bmatrix}$$
$$= \begin{bmatrix} b(\text{tr}\left[\boldsymbol{K}^2\right]-N)\sigma_g^2 \\ b(N-\text{tr}\left[\boldsymbol{K}^2\right])\sigma_g^2+\sigma_g^2+\sigma_e^2\end{bmatrix}$$

We approximate $b = \mathbb{E}[\frac{1}{L_B-N}]$ using Taylor expansion. As we have : $f(y) \approx f(x) + f'(x)(y-x) + \frac{1}{2}f''(x)(y-x)^2$. Let $X \equiv L_B - N$, and thus $\mu_x = \mathbb{E}[L_B - N] = \text{tr}\left[\boldsymbol{K}^2\right] - N$. We have $f(x) = \frac{1}{x}, f'(x) = -\frac{1}{x^2}, f''(x) = \frac{2}{x^3}$.

Thus :

$$\begin{aligned} b = \mathbb{E}[f(X)] &\approx \mathbb{E}[f(\mu_x) + f'(\mu_x)(X-\mu_x) + \frac{1}{2}f''(\mu_x)(X-\mu_x)^2] \\ &= f(\mu_x) + \frac{1}{-\mu_x^2}\mathbb{E}[X-\mu_x] + \frac{1}{2}\frac{2}{\mu_x^3}\mathbb{E}[(X-\mu_x)^2 \\ &= \frac{1}{\mu_x} + \frac{1}{\mu_x}\frac{\sigma_x^2}{\mu_x^2}\end{aligned}$$

where $\sigma_x^2 = var(X)$.

Thus $\mathbb{E}[\frac{\mu_x}{x}] = 1 + \frac{\sigma_x^2}{\mu_x^2}$. Thus $\mathbb{E}[\widetilde{\sigma_g^2}] = \sigma_g^2 + \frac{\sigma_x^2}{\mu_x^2}\sigma_g^2$, $\mathbb{E}[\sigma_e^2] = \sigma_e^2 - \frac{\sigma_x^2}{\mu_x^2}\sigma_g^2$, $\mathbb{E}[\widetilde{\sigma_g^2} + \widetilde{\sigma_e^2}] = \sigma_g^2 + \sigma_e^2$.

For σ_x^2, we have:

$$\begin{aligned}
\sigma_x^2 &= \mathbb{E}[(L_B - \text{tr}\left[\boldsymbol{K}^2\right])^2] \\
&= var(L_B) \\
&= var(\frac{1}{B}\sum_B \boldsymbol{z}_b^T \boldsymbol{K}^2 \boldsymbol{z}_b) \quad \boldsymbol{z}_b \texttt{ are independent} \\
&= \frac{1}{B^2}\sum_B var(\boldsymbol{z}_b^T \boldsymbol{K}^2 \boldsymbol{z}_b) \quad \boldsymbol{z}_b \texttt{ are identically distributed} \\
&= \frac{1}{B}\sum_{i,j} \boldsymbol{K}_i^T \boldsymbol{K}_j \boldsymbol{z}_i \boldsymbol{z}_j \quad \texttt{elements of } \boldsymbol{z} \texttt{ are independent} \\
&= \frac{1}{B}\sum_i \boldsymbol{K}_i^2 = \frac{1}{B}\text{tr}\left[\boldsymbol{K}^2\right]
\end{aligned}$$

Here $\boldsymbol{K}_i$ is the i^{th} column of K.

Thus, substituting μ_x and σ_x^2, we get $\mathbb{E}[\widetilde{\sigma_g^2}] = \sigma_g^2 + \frac{1}{B}\frac{\text{tr}\left[\boldsymbol{K}^2\right]}{(\text{tr}\left[\boldsymbol{K}^2\right]-N)^2}\sigma_g^2 = \sigma_g^2 + \frac{1}{B}\frac{1}{\text{tr}\left[\boldsymbol{K}^2\right]-2N+\frac{N^2}{\text{tr}\left[\boldsymbol{K}^2\right]}}\sigma_g^2$. The bias of the estimator decreases with larger number of random vectors B.

B.3 Standard Error Estimate for the RHE-reg estimator

We define $var(\boldsymbol{y}) \equiv \Sigma = \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I}$. As we know $\widetilde{\sigma_g^2} = \frac{\boldsymbol{y}^T(\boldsymbol{K}-\boldsymbol{I})\boldsymbol{y}}{L_B - N}$. Let $\widetilde{\sigma_g^2} \equiv \frac{A}{B}$ where $A \equiv y^T(K-I)y$ and $B \equiv L_B - N$. Define $\mu_A \equiv \mathbb{E}[A]$, $\mu_B \equiv \mathbb{E}[B]$, $\sigma_A^2 \equiv var(A)$ and $\sigma_B^2 \equiv var(B)$. From Lemma 2, we have

$$\begin{aligned}
var(\widetilde{\sigma_g^2}) &= var(\frac{A}{B}) \\
&= \frac{1}{(\mu_B)^2}\sigma_A^2 - 2\frac{\mu_A}{(\mu_B)^3}cov(A,B) + \frac{(\mu_A)^2}{(\mu_B)^4}\sigma_B^2 \\
&= \frac{1}{(\mu_B)^2}\sigma_A^2 + \frac{(\mu_A)^2}{(\mu_B)^4}\sigma_B^2
\end{aligned}$$

as A, B are independent. By using Lemma 1, we have:

$$\mu_A = \mathbb{E}[\boldsymbol{y}^T(\boldsymbol{K}-\boldsymbol{I})\boldsymbol{y}] = (\text{tr}\left[\boldsymbol{K}^2\right] - N)\sigma_g^2$$
$$\sigma_A^2 = var(\boldsymbol{y}^T(\boldsymbol{K}-\boldsymbol{I})\boldsymbol{y}) = 2\text{tr}\left[\boldsymbol{\Sigma}(\boldsymbol{K}-\boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{K}-\boldsymbol{I})\right]$$
$$\mu_B = \text{tr}\left[\boldsymbol{K}^2\right] - N$$
$$\sigma_B^2 = \frac{\text{tr}\left[\boldsymbol{K}^2\right]}{B}$$

Thus we have:

$$SE(\widetilde{\sigma_g^2}) = \frac{1}{\text{tr}\left[\boldsymbol{K}^2\right] - N}\sqrt{2\text{tr}\left[\boldsymbol{\Sigma}(\boldsymbol{K}-\boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{K}-\boldsymbol{I})\right] + \frac{1}{B}(\sigma_g^2)^2\text{tr}\left[\boldsymbol{K}^2\right]}$$

In order to estimate the standard error of $\widetilde{\sigma_g^2}$, we use the plug-in estimator:

$$\widehat{SE(\widetilde{\sigma_g^2})} = \frac{1}{L_B - N}\sqrt{2\text{tr}\left[\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{K}-\boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{K}-\boldsymbol{I})\right] + \frac{1}{B}(\widetilde{\sigma_g^2})^2 L_B} \tag{B.1}$$

Each term in this estimator could be efficiently computed in $O(\frac{NMB}{max(\log_3 N, \log_3 M)})$.

Useful identities

Lemma 1: For a random vector $\boldsymbol{z}$ that is distributed according to a multivariate normal distribution: $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{C})$ and for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

$$cov(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}, \boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z}) = 2\text{tr}\left[\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{B}\right]$$

Thus

$$\begin{aligned}\mathbb{E}[(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z})(\boldsymbol{z}^T B\boldsymbol{z})] &= 2\text{tr}\left[\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{B}\right] + \mathbb{E}[(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z})]\mathbb{E}[(\boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z})] \\ &= 2\text{tr}\left[\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{B}\right] + \text{tr}\left[\boldsymbol{A}\boldsymbol{C}\right]\text{tr}\left[\boldsymbol{B}\boldsymbol{C}\right]\end{aligned}$$

Lemma 2: For two random variables, A and B, where B is either discrete or has support [0,∞), and $\mathbb{E}[A] = \mu_A$, $\mathbb{E}[B] = \mu_B$.

$$var(\frac{A}{B}) \approx \frac{1}{(\mu_B)^2} var(A) + 2\frac{-\mu_A}{(\mu_B)^3} cov(A,B) + \frac{(\mu_A)^2}{(\mu_B)^4} var(B)$$

APPENDIX C

Appendix to: Fast estimation of genetic correlation for biobank-scale data

C.1 Modeling fixed-effect covariates

Let $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ denote the corresponding covariate matrices for each trait. To include covariate, the generative model in Equation 3.1 is modified to:

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{W}_1\boldsymbol{\alpha}_1 + \boldsymbol{X}_1\boldsymbol{\beta_1} + \boldsymbol{\epsilon}_1 \\ \boldsymbol{y}_2 &= \boldsymbol{W}_2\boldsymbol{\alpha}_2 + \boldsymbol{X}_2\boldsymbol{\beta_2} + \boldsymbol{\epsilon}_2 \end{aligned} \tag{C.1}$$

Here $\boldsymbol{W}_1$ is a $N_1 \times C_1$ matrix of covariates while $\boldsymbol{\alpha}_1$ denotes the fixed effect effect. Similarly, $\boldsymbol{W}_2$ is a $N_1 \times C_2$ matrix of covariates while $\boldsymbol{\alpha}_2$ is a fix effect effect of C_2-vector. We multiply each of the equations in Equation C.1 by the projection matrices $\boldsymbol{V}_1 = \boldsymbol{I}_{N1} - \boldsymbol{W}_1(\boldsymbol{W}_1^T\boldsymbol{W}_1)^{-1}\boldsymbol{W}_1^T$ and $\boldsymbol{V}_2 = \boldsymbol{I}_{N2} - \boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2^T$:

$$\begin{aligned} \boldsymbol{V}_1\boldsymbol{y}_1 &= \boldsymbol{V}_1\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{V}_1\boldsymbol{\epsilon}_1 \\ \boldsymbol{V}_2\boldsymbol{y}_2 &= \boldsymbol{V}_2\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{V}_2\boldsymbol{\epsilon}_2 \end{aligned} \tag{C.2}$$

Similar to Equation 3.8, the MoM estimator is obtained by minimizing the sum

of squared differences between the population and empirical covariance as :

$$(\widehat{\gamma_g}, \widehat{\gamma_e}, \widehat{\sigma^2_{g1}}, \widehat{\sigma^2_{g2}}, \widehat{\sigma^2_{e1}}, \widehat{\sigma^2_{e2}}) =$$

$$argmin_{\gamma_g,\gamma_e,\sigma^2_{g1},\sigma^2_{g2},\sigma^2_{e1},\sigma^2_{e2}} ||\widetilde{\boldsymbol{y}}\widetilde{\boldsymbol{y}}^T - (\begin{bmatrix} \sigma^2_{g1}\widetilde{\boldsymbol{K}_1} & \gamma_g\widetilde{\boldsymbol{K}_A} \\ \gamma_g\widetilde{\boldsymbol{K}_A}^T & \sigma^2_{g2}\widetilde{\boldsymbol{K}_2} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1}\boldsymbol{V}_1 & \gamma_e\boldsymbol{V}_1\boldsymbol{C}\boldsymbol{V}_2 \\ \gamma_e\boldsymbol{V}_2\boldsymbol{C}^T\boldsymbol{V}_1 & \sigma^2_{e2}\boldsymbol{V}_2 \end{bmatrix})||^2_F$$

where $\widetilde{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{V}_1\boldsymbol{y}_1 \\ \boldsymbol{V}_2\boldsymbol{y}_2 \end{bmatrix}$, $\widetilde{\boldsymbol{K}_1} = \frac{\boldsymbol{V}_1\boldsymbol{X}_1\boldsymbol{X}_1^T\boldsymbol{V}_1}{M}$ and $\widetilde{\boldsymbol{K}_2} = \frac{\boldsymbol{V}_2\boldsymbol{X}_2\boldsymbol{X}_2^T\boldsymbol{V}_2}{M}$, $\widetilde{\boldsymbol{K}_A} = \frac{\boldsymbol{V}_1\boldsymbol{X}_1\boldsymbol{X}_2^T\boldsymbol{V}_2}{M}$, and $\widetilde{\boldsymbol{K}_C} = \frac{\boldsymbol{V}_1\boldsymbol{X}_1\boldsymbol{X}_2^T\boldsymbol{V}_2C^T}{M}$.

Thus the MoM estimator for genetic covariance satisfies the normal equations:

$$\begin{bmatrix} tr(\widetilde{\boldsymbol{K}_A}\widetilde{\boldsymbol{K}_A}^{\mathrm{T}}) & tr(\widetilde{\boldsymbol{K}_C}) \\ tr(\widetilde{\boldsymbol{K}_C}) & N \end{bmatrix} \begin{bmatrix} \widehat{\gamma_g} \\ \widehat{\gamma_e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T\widetilde{\boldsymbol{K}_A}\boldsymbol{y}_2 \\ \boldsymbol{y}_1^T\boldsymbol{V}_1\boldsymbol{C}\boldsymbol{V}_2\boldsymbol{y}_2 \end{bmatrix} \tag{C.3}$$

SCORE replaces $tr(\widetilde{\boldsymbol{K}_A}\widetilde{\boldsymbol{K}_A}^{\mathrm{T}})$ with an unbiased randomized estimate $\widetilde{L_B}$ using B random vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$, $z_b \in \mathbb{R}^{N_2}, b \in 1 \ldots B$ drawn independently from a distribution with zero mean and identity covariance matrix $\boldsymbol{I}_{N_2}$. The estimator of $tr(\widetilde{\boldsymbol{K}_A}\widetilde{\boldsymbol{K}_A}^{\mathrm{T}})$ is given by:

$$\widetilde{L_B} = \frac{1}{B}\frac{1}{M^2}\sum_b ||\boldsymbol{V}_1\boldsymbol{X}_1\boldsymbol{X}_2^T\boldsymbol{V}_2z_b||_2^2$$

The SCORE estimator is thus obtained by solving Equation C.3 by replacing

$tr(\widetilde{\boldsymbol{K}_A}\widetilde{\boldsymbol{K}_A}^{\mathrm{T}})$ with $\widetilde{L_B}$. $\widetilde{L_B}$ is an unbiased estimator for $tr(\widetilde{\boldsymbol{K}_A}\widetilde{\boldsymbol{K}_A}^{\mathrm{T}})$ since

$$\begin{aligned}
\mathbb{E}[\widetilde{L_B}] &= \frac{1}{B}\frac{1}{M^2}\sum_b \mathbb{E}[\boldsymbol{z}_b^T(\boldsymbol{V}_1\boldsymbol{X}_1\boldsymbol{X}_2^T\boldsymbol{V}_2)^T\boldsymbol{V}_1\boldsymbol{X}_1\boldsymbol{X}_2^T\boldsymbol{V}_2\boldsymbol{z}_b] \\
&= \frac{1}{B}\sum_b \mathbb{E}[z_b^T\widetilde{\boldsymbol{K}_A}^T\widetilde{\boldsymbol{K}_A}\boldsymbol{z}_b] \\
&= \frac{1}{B}\sum_b \mathbb{E}[tr(z_b^T\widetilde{\boldsymbol{K}_A}^T\widetilde{\boldsymbol{K}_A}\boldsymbol{z}_b)] \\
&= \frac{1}{B}\sum_b \mathbb{E}[tr(\boldsymbol{z}_b\boldsymbol{z}_b^T\widetilde{\boldsymbol{K}_A}^T\widetilde{\boldsymbol{K}_A})] \\
&= \frac{1}{B}\sum_b tr(\mathbb{E}[\boldsymbol{z}_b\boldsymbol{z}_b^T\widetilde{\boldsymbol{K}_A}^T\widetilde{\boldsymbol{K}_A}]) \\
&= \frac{1}{B}\sum_b tr(\mathbb{E}[\boldsymbol{z}_b\boldsymbol{z}_b^T]\widetilde{\boldsymbol{K}_A}^T\widetilde{\boldsymbol{K}_A}) \\
&= \frac{1}{B}\sum_b tr(\boldsymbol{I}_{N2}\widetilde{\boldsymbol{K}_A}^T\widetilde{\boldsymbol{K}_A}) \\
&= tr(\widetilde{\boldsymbol{K}_A}^T\widetilde{\boldsymbol{K}_A})
\end{aligned}$$

The projection matrix $\boldsymbol{V}_1, \boldsymbol{V}_2$ need not need be computed explicitly. While computing $\boldsymbol{X}_2^T\boldsymbol{V}_2\boldsymbol{z}b$, we only need to compute the residual of $\boldsymbol{W}_1(\boldsymbol{W}_1^T\boldsymbol{W}_1)^{-1}\boldsymbol{W}_1{}^T\boldsymbol{z}_b$, where the additional computation has the complexity of $\mathcal{O}(N_1C_1)$ where C_1 is the number of covariates, which usually is a relatively small number.

C.2 Jackknife Standard Error

In order to compute the standard error using block Jackknife [49], we partition the standardized $N\times M$ genotype matrix $\boldsymbol{X}$ into J non-overlapping blocks, $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(J)}$ where $\boldsymbol{X}^{(j)} \in \{1, \ldots, J\}$ is an $N \times \frac{M}{J}$ matrix.

We define $\widehat{\rho_{g(j)}}$ to be the estimator of genetic correlation computed after excluding

genotype block $\boldsymbol{X}^{(j)}$ from $\boldsymbol{X}$. Also, we define $\overline{\widehat{\rho_{g(j)}}} \equiv \frac{1}{J}\sum_j \widehat{\rho_{g(j)}}$

Thus, the jackknife estimate of the standard error is given as

$$\widehat{SE}(\widehat{\rho_g}) = \left[\frac{J-1}{J}\sum_{j=1}^{J}(\widehat{\rho_{g(j)}} - \overline{\widehat{\rho_{g(j)}}})^2\right]^{\frac{1}{2}} \tag{C.4}$$

Bibliography

[1] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era?concepts and misconceptions. *Nature reviews genetics*, 9(4):255, 2008.

[2] Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, 2008.

[3] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203, 2006.

[4] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, sep 2011.

[5] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284, 2015.

[6] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407, 2014.

[7] Wouter Van Rheenen, Wouter J Peyrot, Andrew J Schork, S Hong Lee, and

Naomi R Wray. Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics*, 20(10):567–581, 2019.

[8] Masato Akiyama, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genome-wide association study identifies 112 new loci for body mass index in the japanese population. *Nature genetics*, 49(10):1458, 2017.

[9] S Hong Lee, Stephan Ripke, Benjamin M Neale, Stephen V Faraone, Shaun M Purcell, Roy H Perlis, Bryan J Mowry, Anita Thapar, Michael E Goddard, John S Witte, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature genetics*, 45(9):984, 2013.

[10] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236, 2015.

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

[12] Gleb Kichaev and Bogdan Pasaniuc. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *The American Journal of Human Genetics*, 97(2):260–271, 2015.

[13] Laura Almasy and John Blangero. Multipoint quantitative-trait linkage analysis

in general pedigrees. *The American Journal of Human Genetics*, 62(5):1198–1211, 1998.

[14] David Houle. Comparing evolvability and variability of quantitative traits. *Genetics*, 130(1):195–204, 1992.

[15] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[16] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

[17] Robert Makowsky, Nicholas M Pajewski, Yann C Klimentidis, Ana I Vazquez, Christine W Duarte, David B Allison, and Gustavo de Los Campos. Beyond missing heritability: prediction of complex traits. *PLoS genetics*, 7(4):e1002051, 2011.

[18] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7):507, 2013.

[19] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[20] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet*, 2010.

[21] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[22] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 47(12):1385, 2015.

[23] Matti Pirinen, Peter Donnelly, and Chris CA Spencer. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, pages 369–390, 2013.

[24] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[25] Guo-Bo Chen. Estimating heritability of complex traits from genome-wide association studies using ibs-based haseman–elston regression. *Frontiers in genetics*, 5:107, 2014.

[26] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics*, 2(1):3–19, 1972.

[27] Robert C Elston, Sarah Buxbaum, Kevin B Jacobs, and Jane M Olson. Haseman and elston revisited. *Genetic epidemiology*, 19(1):1–17, 2000.

[28] Wei-Min Chen, Karl W Broman, and Kung-Yee Liang. Quantitative trait linkage analysis by generalized estimating equations: Unification of variance components and haseman-elston regression. *Genetic epidemiology*, 26(4):265–272, 2004.

[29] Brendan Bulik-Sullivan. Relationship between ld score and haseman-elston regression. *bioRxiv*, page 018283, 2015.

[30] Edo Liberty and Steven W Zucker. The mailman algorithm: A note on matrix–vector multiplication. *Information Processing Letters*, 109(3):179–182, 2009.

[31] Ali Pazokitoroudi, Yue Wu, Kathryn S Burch, Kangcheng Hou, Aaron Zhou, Bogdan Pasaniuc, and Sriram Sankararaman. Efficient variance components analysis across millions of genomes. *Nature communications*, 11(1):4020, August 2020.

[32] PC Sham, SS Cherny, S Purcell, and JK Hewitt. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *The American Journal of Human Genetics*, 66(5):1616–1630, 2000.

[33] PC Sham and S Purcell. Equivalence between haseman-elston and variance-components linkage analyses for sib pairs. *The American Journal of Human Genetics*, 68(6):1527–1532, 2001.

[34] Tian Ge, Chia-Yen Chen, Benjamin M Neale, Mert R Sabuncu, and Jordan W Smoller. Phenome-wide heritability analysis of the uk biobank. *PLoS genetics*, 13(4):e1006711, 2017.

[35] Chiara Sabatti, Susan K. Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G. Jones, Noah A. Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, Aimo Ruokonen, Jaana Laitinen, Eveliina Jakkula, Lachlan Coin, Clive Hoggart, Andrew Collins, Hannu Turunen, Stacey Gabriel, Paul Elliot, Mark I. McCarthy, Mark J. Daly, Marjo-Riitta Jrvelin, Nelson B. Freimer, and Leena Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, 41(1):35–46, Jan 2009.

[36] S.H. Lee, J. Yang, M.E. Goddard, P.M. Visscher, and N.R. Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012.

[37] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 47(12):1385, 2015.

[38] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

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

[40] Qiongshi Lu, Boyang Li, Derek Ou, Margret Erlendsdottir, Ryan L Powles, Tony Jiang, Yiming Hu, David Chang, Chentian Jin, Wei Dai, et al. A powerful approach to estimating annotation-stratified genetic covariance via gwas summary statistics. *The American Journal of Human Genetics*, 101(6):939–964, 2017.

[41] Doug Speed and David J Balding. Sumher better estimates the snp heritability of complex traits from summary statistics. *Nature genetics*, 51(2):277–284, 2019.

[42] Omer Weissbrod, Jonathan Flint, and Saharon Rosset. Estimating snp-based heritability and genetic correlation in case-control studies directly and with summary statistics. *The American Journal of Human Genetics*, 103(1):89–99, 2018.

[43] Zheng Ning, Yudi Pawitan, and Xia Shen. High-definition likelihood inference of genetic correlations across human complex traits. Technical report, Nature Publishing Group, 2020.

[44] Guiyan Ni, Gerhard Moser, Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, et al. Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *The American Journal of Human Genetics*, 2018.

[45] Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature genetics*, page 1, 2019.

[46] Yiliang Zhang, Youshu Cheng, Yixuan Ye, Wei Jiang, Qiongshi Lu, and Hongyu

Zhao. Comparison of methods for estimating genetic correlation between complex traits using gwas summary statistics. *bioRxiv*, 2020.

[47] MF Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[48] Yue Wu and Sriram Sankararaman. A scalable estimator of snp heritability for biobank-scale data. *Bioinformatics*, 34(13):i187–i194, 2018.

[49] Hans R Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

[50] Nazim Ghouri, David Preiss, and Naveed Sattar. Liver enzymes, nonalcoholic fatty liver disease, and incident cardiovascular disease: a narrative review and clinical perspective of prospective data. *Hepatology*, 52(3):1156–1161, 2010.

[51] Ming Gao, Yi Cheng, Yang Zheng, Weihua Zhang, Lin Wang, and Ling Qin. Association of serum transaminases with short-and long-term outcomes in patients with st-elevation myocardial infarction undergoing primary percutaneous coronary intervention. *BMC cardiovascular disorders*, 17(1):43, 2017.

[52] Kyung Mook Choi, Kyungdo Han, Sanghyun Park, Hye Soo Chung, Nam Hoon Kim, Hye Jin Yoo, Ji-A Seo, Sin Gon Kim, Nan Hee Kim, Sei Hyun Baik, et al. Implication of liver enzymes on incident cardiovascular diseases and mortality: A nationwide population-based cohort study. *Scientific reports*, 8(1):1–9, 2018.

[53] Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam

Shoresh, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, 50(4):621–629, 2018.

[54] Yiliang Zhang, Qiongshi Lu, Yixuan Ye, Kunling Huang, Wei Liu, Yuchang Wu, Xiaoyuan Zhong, Boyang Li, Zhaolong Yu, Brittany G Travers, et al. Supergnova: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome biology*, 22(1):1–30, 2021.