Title
Elucidating the Genetic Architecture of Complex Traits with Variance Component Models

Author
Kim, Juhyun

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Elucidating the Genetic Architecture of Complex Traits with Variance Component Models

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Juhyun Kim

2021

ABSTRACT OF THE DISSERTATION

Elucidating the Genetic Architecture of Complex Traits with Variance Component Models

by

Juhyun Kim
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2021
Professor Hua Zhou, Chair

Variance component models are a fundamental topic in statistical genetics. These models enable us to estimate the underlying heritability of a phenotype, adjust for confounding in association testing, and assess the strength of effects of a set of genetic markers on a phenotype. Under the overarching theme of variance component models, this dissertation aims to elucidate the genetic architecture of complex diseases and traits by developing and applying variance component model-based methods to analyze high-dimensional genomic data. In the first half of the dissertation, we propose a variance component selection framework that jointly models and prioritizes a set of genetic markers that are associated with quantitative traits. The second half of the dissertation is devoted to quantifying the heritability of diabetes complications. We use various heritability estimation methods, some of which are based on variance component models.

The dissertation of Juhyun Kim is approved.

Gang Li

Janet Sinsheimer

Kenneth L. Lange

Hua Zhou, Committee Chair

University of California, Los Angeles

2021

To God and my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Looking back, the past six years I spent as a graduate student could have gone wrong as they coincided with many disasters. In 2016, a shooting on campus put the campus on lockdown and frightened everyone. And the wildfires in 2017, 2018, 2019, and 2020. I still remember the days of the dark, orange brown sky with soot and ash falling like snow. Last but not least, COVID-19 that shut down the world and made 2020 a year like no other. I survived through all of them, and I am here writing this acknowledgments because of all the people who enabled it.

First and foremost, I would like to thank my dissertation advisor, Dr. Hua Zhou, for his guidance, support, patience, and availability over the past five years. Along with Dr. Hua Zhou, Dr. Jin Zhou served as the de facto co-advisor through all chapters in the dissertation, although not officially in my committee. I have learned and grown tremendously under their mentorship. I also want to thank Drs. Judong Shen, Anran Wang, and Devan V. Mehrotra from Merck & Co., Inc. for the IMPROVE-IT data analysis in Chapter 3 and their helpful comments throughout Chapters 2 and 3. I would also like to express my gratitude to my committee members—Drs. Ken Lange, Janet Sinsheimer, and Gang Li—for their valuable suggestions and comments.

My doctoral work was financially supported by the UCLA Graduate Division through various fellowships and the NIH-funded Genomic Analysis and Interpretation Training Program. I want to thank Drs. Jeanette Papp and Eric Sobel for their support and mentorship throughout the training program.

This paragraph is dedicated to the OpenMendel group, which I was fortunate to be part of. I am grateful for the opportunities given to me to contribute to the OpenMendel packages and present at the Lange Symposium and the American Society of Human Genetics meeting, not to mention having had a community of intelligent and extremely nice people.

Thank you to my friends whose presence made this journey pleasant—Tracey Chan who

commiserated with me from year one and opened her home filled with snacks for work and exercise, Christina Lau who joined Tracey and me and had to hear about graduate student life for the past six years, Yu Shi who studied with me in the basement for the doctoral qualifying exam and served as a terrific tour guide in Japan, Claire Kim who is a wonderful friend, confidant and whose company I enjoyed the most in the COVID era, Howon Ryu who was my best friend in the department and a delightful travel mate on the Iberian Peninsula. Thanks to Kevin who inspired me and kept me going in my last year.

Finally, a big thank-you to my family for their support and prayers. My dad and mom did not exactly know what my research was about, but they knew exactly how to encourage and motivate their little girl. My sister and my brother have been there for me every step of the way from the early days of "Will I ever graduate?" to the final days.

VITA

2015	B.S. (Mathematics), University of California, Los Angeles
2018-2019	Statistical Consultant, Institute for Digital Research and Education (IDRE), University of California, Los Angeles
2019	Co-op Statistician, GlaxoSmithKline, Collegeville, PA
2015–2020	Teaching Assistant, Biostatistics Department, University of California, Los Angeles

PUBLICATIONS

A. Peer-reviewed journals

Zhou, H., Sinsheimer, J. S., Bates, D. M., Chu, B. B., German, C. A., Ji, S. S., Keys, K. L., **Kim, J.**, Ko, S., Mosher, G. D., Papp, J. C., Sobel, E. M., Zhai, J., Zhou, J. J., and Lange, K. (2020). OPENMENDEL: A cooperative programming project for statistical genetics. *Human Genetics*, 139(1):61-71.

Kim, J., Zhang, Y., Day, J., and Zhou, H. (2018). MGLM: An R package for multivariate categorical data analysis. *The R Journal*, 10(1):73-90.

Gaines, B., **Kim, J.**, and Zhou, H. (2018). Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861-871.

Zhai, J., **Kim, J.**, Knox K. S., Twigg H. L. III, Zhou, H., and Zhou, J. J. (2018). Variance component selection with applications to microbiome taxonomic data. *Frontiers in Microbiology*, 9:509.

B. Under review

Kim, J., Shen, J., Wang, A., Mehrotra, D. V., Ko, S., Zhou, J. J., and Zhou, H. (2021). VCSEL: Prioritizing SNP-set by penalized variance component selection. (under review)

C. In preparation

Kim, J., Jensen, A., Klimentidis, Y. C., Sun, Y., Zhou, H., Reaven, P., and Zhou, J. J. (2021). Systematic heritability and heritability enrichment analysis for diabetes complications in ACCORD and UK Biobank Studies. (in preparation)

CHAPTER 1

Introduction

Proposed by R. A. Fisher (1918), variance component models have been a fundamental topic in statistical genetics. These models enable us to estimate the genetic contribution to variation in a phenotype (e.g., height, weight, blood pressure, test score), adjust for confounding in association testing, assess the strength of effects of a set of genetic markers on a phenotype, and compute genetic correlations. The idea behind variance components lies in the decomposition of the overall variance of the outcome into particular sources. Consider the following standard variance component model where $\boldsymbol{y}$ is a $n \times 1$ vector of quantitative measurements taken from n individuals and $\boldsymbol{X}$ is a $n \times p$ matrix of p covariates (e.g., intercept, sex, age). Denoting $\boldsymbol{G}$ to be a genetic relationship matrix or a kernel matrix, depending on the context, and $\boldsymbol{I}_n$ to be a $n \times n$ identity matrix, we assume the following:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_g^2 \boldsymbol{G} + \sigma_e^2 \boldsymbol{I}_n), \tag{1.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects parameters and σ_g^2 and σ_e^2 are scalar parameters representing genetic and environmental variances, respectively. Because the overall variance of our quantitative trait, $\text{Var}(\boldsymbol{y})$, is divided into genetic and environmental components, σ_g^2 and σ_e^2 are called variance components. This model is general and flexible in it can easily incorporate other covariance terms such as sample relatedness or shared environment effects.

Under the overarching theme of variance component models, this dissertation aims to understand the genetic contributions to complex diseases and traits by developing and applying methods involving variance component models to analyze high-dimensional genomic

data. We address two main scenarios in which variance component models are employed: (i) testing for an association between a set of genetic markers and a phenotype and (ii) estimating the underlying heritability of a trait.

The first scenario arises from genetic association studies, which aim to find genetic variations contributing to a particular complex disease or trait. Traditional genome-wide association studies (GWAS) test one variant at a time for association with a disease. Despite the deluge of discoveries, common (minor allele frequency (MAF) > 0.05) genetic variants (usually single nucleotide polymorphisms (SNP)) identified in GWAS remain insufficient for explaining much of the genetic contribution to complex traits. One hypothesis behind this so-called "missing heritability" problem posits that low-frequency and rare variants (MAF ≤ 0.05) can explain the remaining disease risk or trait variability. When faced with low-frequency or rare variants, however, the classical single marker test is seriously underpowered. One strategy to alleviate the power issue is a gene- or region-based test, in which SNPs within a gene or region are aggregated and tested for the joint effect of variants on the risk of complex disease (for review, see Lee et al., 2014). The use of variance component models is widely accepted in the SNP-set analysis. One popular variance component model-based method for such analysis is the score-based Sequence Kernel Association Test (SKAT) method (Wu et al., 2011). Under this framework, testing for association corresponds to testing the null hypothesis $\sigma_g^2 = 0$ in the notation of (1.1).

The SKAT gave rise to a plethora of methods for rare variant association analyses in various settings. Gene Association with Multiple Traits (GAMuT; Broadaway et al., 2016), Dual Kernel-based Association Test (DKAT; Zhan et al., 2017) and Multi-SKAT (Dutta et al., 2019) allow multiple phenotype test for rare variants with kernels for genotypes and phenotypes. Methods that can incorporate gene-environment interactions are available as well (Chen et al., 2014; Lin et al., 2016; Lim et al., 2020). However, these methods are marginal testing procedures that are limited to inspecting one SNP-set at a time. Chapters 2 and 3 address the limitations of marginal testing approaches and introduce methods that

jointly model multiple SNP-sets and identify the sets that are relevant to quantitative trait(s).

The other scenario that demonstrates the utility of variance component models in quantitative genetics is heritability. The narrow-sense heritability, typically referred to simply as the "heritability," is defined as the proportion of phenotypic variance that can be attributed to additive genetic variance. In the notation of (1.1), heritability equals to

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}.$$

Estimating the heritability of a trait is often an initial step for understanding complex traits with many contributing factors. Heritability can suggest researchers how much to consider hereditary influences when they want to learn about causes for a particular trait. In the absence of genetic data, heritability has historically been estimated based on family or twin-based designs. However, the studies based on family relationships had severe flaws and were subject to confounding by shared environment effects (Keller and Coventry, 2005; Tenesa and Haley, 2013).

Advances in array-based and whole-genome sequencing techniques have led to the estimation of single-nucleotide polymorphism (SNP) heritability using large-scale genetic datasets (Yang et al., 2010). SNP heritability measures the degree to which a given set of measured SNPs explains the phenotypic variance. Yang et al. (2010) used a single variance component to capture all common array SNPs. However, common genetic variants do not capture all the variation in heritability, as mentioned earlier. In order to incorporate dense genetic markers from imputed data or whole genome sequencing data, multiple variance component approaches were adopted (Yang et al., 2011b, 2015). In the multiple-component approach, SNPs are binned into different SNP-property categories (e.g., MAF, Linkage Disequilibrium (LD)), and the heritability estimate is obtained by summing estimates across the categories. In addition to stratifying SNPs based on their properties, methods have been developed to partition the heritability based on SNP functional annotations (Finucane et al., 2015; Zhou,

2017; Hao et al., 2018). Of note, stratified LD score regression (S-LDSC) uses summary statistics from GWAS to compute heritability and heritability enrichment in functional categories. Since the heritability of complex traits is not distributed evenly across the whole genome (Maurano et al., 2012; Trynka et al., 2013), examining regions enriched for heritability can provide insights into functional categories that contribute to heritability. We partake in this endeavor to understand complex traits by quantifying the SNP heritability of diabetes complications in Chapter 4.

The contents of Chapter 2 and Chapter 3 are adapted from the following manuscript:

> Kim, J., Shen, J., Wang, A., Mehrotra, D. V., Ko, S., Zhou, J. J., and Zhou, H. (2021). VCSEL: Prioritizing SNP-set by penalized variance component selection. (under review)

Chapter 4 is adapted from:

> Kim, J., Jensen, A., Klimentidis, Y. C., Sun, Y., Zhou, H., Reaven, P., and Zhou, J. J. (2021). Systematic heritability and heritability enrichment analysis for diabetes complications in ACCORD and UK Biobank Studies. (in preparation)

The remainder of this dissertation is organized as follows: Chapter 2 presents a majorization-minimization (MM) algorithm that identifies relevant variance components given a multivariate response model with potentially many variance components. In Chapter 3 we extend the method introduced in Chapter 2 to incorporate interaction terms in a univariate response setting. Chapter 4 shifts the focus from MM algorithms and investigates the genetic components behind the development and progression of diabetes complications. This is achieved by quantifying heritability and heritability enrichment through different approaches, including the variance component model-based method. We conclude this dissertation with Chapter 5.

CHAPTER 2

Variance component selection for multivariate response model

2.1 Introduction

The limited success of genome-wide association studies (GWAS) has diverted attention away from common genetic variants, usually denoted by minor allele frequency (MAF) > 0.05. Instead, rare variants (MAF ≤ 0.05) are believed to play an important role in elucidating many common diseases and complex traits (Bodmer and Bonilla, 2008; Manolio et al., 2009; Bansal et al., 2010; Rivas et al., 2011; Gibson, 2012; Gudmundsson et al., 2012; Zuk et al., 2014; Lee et al., 2014). Although association test for common variants in a GWAS analysis is often conducted one variant at a time, this approach results in low statistical power in rare-variant association studies due to their prevalence and extremely low frequency (Li and Leal, 2008; Madsen and Browning, 2009; Zuk et al., 2014). As a remedy, many have proposed single nucleotide polymorphism (SNP) set analysis, also known as gene set, pathway, or region-based analysis (Wu et al., 2010; Dering et al., 2011). In these analyses, variants are binned into a biologically relevant unit such as a gene, pathway, or sliding window, and tested for association with complex traits. Compared to the classical single-variant-based approach, SNP-set analysis enjoys increased power as it reduces multiple comparison burden and aggregates weak signals (Rivas and Moutsianas, 2015).

In addition to the high polygenicity—influenced by a large number of genetic variants with small effects—many complex traits are inherently multi-phenotypic. For example, blood

pressure is evaluated by both systolic and diastolic pressure measurements. Obesity is determined not only by body mass index but also by waist circumference and body fat percentage. As one indicator may reveal one susceptibility gene over other indicators, it is important to jointly analyze multiple phenotype data in the analysis (Suo et al., 2013). In addition, GWAS have unveiled that many loci affect more than one trait or disease—a phenomenon known as pleiotropy (Sivakumaran et al., 2011; Solovieff et al., 2013). Testing one phenotype at a time, albeit simple and intuitive, fails to exploit the underlying shared genetic architecture of multiple phenotypes and is also subject to multiple testing penalties. On the other hand, multi-trait analyses can increase statistical power to detect association and provide important insights into pathways that certain traits or diseases share (Suo et al., 2013; Hackinger and Zeggini, 2017).

A plethora of marginal test based methods are available to detect associations of a SNP-set with multiple traits, which are termed cross-phenotype associations. For example, Maity et al. (2012); Lee et al. (2017b); Wu and Pankow (2016); Broadaway et al. (2016); Zhan et al. (2017); Dutta et al. (2019) take region-based approaches, in which variants are grouped based on pre-specified criteria and tested for cross-phenotype effects. Notably, Multi-SKAT (Dutta et al., 2019) provides a general mixed effect model-based framework for joint analysis of multiple continuous phenotypes, unlike most methods that make specific assumptions about the effects of the variants on multiple phenotypes. However, to our best knowledge, no existing methods investigate sets of genetic variants simultaneously.

Here we propose a method for jointly modeling multiple SNP-sets and selecting groups that are relevant to multiple traits while adjusting for covariates. Suppose we have observations from n individuals with d continuous phenotypes, represented by $n \times d$ matrix, and m SNP-sets. Multivariate response model with $n \times d$ response matrix $\widetilde{\boldsymbol{Y}}$ and $n \times p$ covariate matrix $\boldsymbol{X}$ assumes a multivariate normal model

$$\operatorname{vec} \widetilde{\boldsymbol{Y}} \sim N(\operatorname{vec}(\boldsymbol{X}\boldsymbol{B}), \boldsymbol{\Sigma}_1 \otimes \widetilde{\boldsymbol{V}}_1 + \cdots + \boldsymbol{\Sigma}_m \otimes \widetilde{\boldsymbol{V}}_m + \boldsymbol{\Sigma}_0 \otimes \boldsymbol{I}_n), \tag{2.1}$$

where $\boldsymbol{B}$ is the unknown $p \times d$ fixed effects parameters matrix, $\boldsymbol{\Sigma}_i$ are unknown $d \times d$ positive semidefinite variance component matrices, and $\widetilde{\boldsymbol{V}}_i$ are known $n \times n$ kernel matrices for genotypes. The vec $\widetilde{\boldsymbol{Y}}$ operator in (2.1) creates an $nd \times 1$ vector from a matrix $\widetilde{\boldsymbol{Y}}$ by stacking its column vectors, and $\otimes$ indicates Kronecker product.

As our interest lies in estimating variance components, we adopt the restricted (or residual) maximum likelihood estimation (REML) approach (Thompson et al., 1962; Patterson and Thompson, 1971; Harville, 1977; Khuri and Sahai, 1985; Robinson, 1987; Searle et al., 1992). In the notation of (2.1), REML first projects $\widetilde{\boldsymbol{Y}}$ to the null space of $\boldsymbol{X}$ and then estimates variance components based on the projected responses. If the columns of the matrix $\boldsymbol{A}$ span the null space of $\boldsymbol{X}^T$ and $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$, then REML estimates parameter $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m)$ by maximizing the log-likelihood of the redefined response matrix $\boldsymbol{Y} = \boldsymbol{A}^T\widetilde{\boldsymbol{Y}}$ whose distribution is as follows:

$$\text{vec}\, \boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_1 \otimes \boldsymbol{V}_1 + \cdots + \boldsymbol{\Sigma}_m \otimes \boldsymbol{V}_m + \boldsymbol{\Sigma}_0 \otimes \boldsymbol{I}_{n-p}). \tag{2.2}$$

where $\boldsymbol{V}_i = \boldsymbol{A}^T\widetilde{\boldsymbol{V}}_i\boldsymbol{A}, i = 1, \ldots, m$. Note that fixed effects have been eliminated.

As there are no closed-form expressions for the REML, we rely on numerical techniques. There are several iterative optimization methods for finding MLE and REML, including Newton's method (Lindstrom and Bates, 1988), Fisher's scoring algorithm, and the expectation-maximization (EM) algorithm (Dempster et al., 1977; Laird and Ware, 1982; Laird et al., 1987; Lindstrom and Bates, 1988; Bates and Pinheiro, 1998). Despite their respective advantages, they suffer from either numerical instability, high computational cost or slow convergence. Zhou et al. (2019) address this issue with a minorization-maximization (MM) algorithm that is simple to implement and numerically efficient. Zhai et al. (2018) implements an MM algorithm for penalizing variance components in microbiome data analysis, but it is limited to lasso penalty and a univariate response. The recent paper (Schaid et al., 2020) applies a similar method as Zhai et al. (2018) to the genetic association setting, but

still restricted to the univariate response setting.

Since SNPs within a gene/pathway/moving window are treated as a unit, this can be considered a group selection problem with each set being a group and SNP being a variable. Several methods have been proposed to take advantage of grouping structures in variables. Group lasso method (Bakin, 1999; Yuan and Lin, 2006) allows group selection by either including or excluding all variables in the group in the model. Bi-level selection or sparse group method (Huang et al., 2009; Breheny and Huang, 2009; Zhou et al., 2010; Simon et al., 2013) enables both group-wise and within group sparsity. However, these approaches are designed for selecting mean, or fixed effects, hence inappropriate when genetic effects are modeled as random effects.

There exists a considerable body of literature on random effect selection. Lin (1997) proposes score tests to detect the significance of individual variance components. To select important random effects, each component is tested separately, followed by some stepwise procedures. Chen and Dunson (2003), Bondell et al. (2010), Fan and Li (2012), and Peng and Lu (2012) consider random effect selection for longitudinal models where observations are divided into independent subjects with a vector of random effects corresponding to each subject. The vectors of random effect are independent and identically distributed with a covariance matrix, which could be a function of one variance component. For these methods, selecting important random effects is essentially limited to within one variance component as it removes rows or columns of covariance matrix or selects components within random effect vectors. No existing method performs a simultaneous selection of random effects at group level to our best knowledge.

In this chapter, we develop a novel penalization method for group selection where each group is treated as random effects. To that end, we devise a general MM-based optimization framework that incorporates both convex and non-convex penalties into variance component models and applies to the analysis of univariate and multivariate traits, respectively.

The remainder of this chapter is organized as follows: Section 2.2 introduces the multi-

variate response variance component model. In Section 2.3, we present the VCSEL algorithm that selects variance components in the realm of multivariate response (VCSEL-M). We illustrate the performance of our methods with simulation studies in Section 2.4 for VCSEL-M and defer the details for the univariate response VCSEL methods to Appendix A.1. In Section 2.5, the proposed methods are applied to the UK-biobank whole exome sequencing study data.

2.2 Multivariate response variance component model

Consider the model (2.2) where $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m$ are known positive semidefinite matrices. Here $\boldsymbol{V}_i$ is a genotype kernel matrix for the i-th variance component. Different choices of kernels can be readily incorporated in $\boldsymbol{V}_i$. As defined in Dutta et al. (2019), one popular choice would be $\boldsymbol{G}_i\boldsymbol{W}_i\boldsymbol{W}_i\boldsymbol{G}_i^T$ where $\boldsymbol{G}_i$ is a genotype matrix corresponding to i-th SNP group and $\boldsymbol{W}_i = \text{diag}(w_1, \ldots, w_q)$ contains the weights of q variants in $\boldsymbol{G}_i$. It corresponds to SKAT and implies that the effects of SNPs in i-th SNP-set are independent. Another choice is $\boldsymbol{G}_i\boldsymbol{W}_i\mathbf{1}\mathbf{1}^T\boldsymbol{W}_i\boldsymbol{G}_i^T$, which corresponds to the Burden test and implies that the effects of SNPs in i-th SNP set are in the same direction. Note that $\mathbf{1}$ denotes a vector of ones. In our simulation studies and real data analysis, we adopt the SKAT genotype kernel and/or the Burden test genotype kernel.

We denote the overall covariance matrix in the model by $\boldsymbol{\Omega}$, i.e.

$$\boldsymbol{\Omega}(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}_1 \otimes \boldsymbol{V}_1 + \cdots + \boldsymbol{\Sigma}_m \otimes \boldsymbol{V}_m + \boldsymbol{\Sigma}_0 \otimes \boldsymbol{I}_{n-p},$$

and assume it to be positive definite. To find estimates of $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m)$, we take a

penalization approach by minimizing the penalized negative log-likelihood function

$$\begin{aligned} &- L(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m) + \sum_{i=1}^{m} P_\lambda(\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i)}) \\ =&\frac{1}{2}\ln\det\boldsymbol{\Omega} + \frac{1}{2}(\mathrm{vec}\boldsymbol{Y})^T\boldsymbol{\Omega}^{-1}\mathrm{vec}\boldsymbol{Y} + \sum_{i=1}^{m} P_\lambda(\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i)}), \end{aligned} \tag{2.3}$$

where P_λ is a penalty term imposing sparsity on variance components for a given tuning parameter λ. Below we derive iterative procedures for lasso (Tibshirani, 1996) and minimax concave penalty (MCP) (Zhang et al., 2010); only a slight modification is needed to accommodate other penalty functions. In practice, we normalize $\boldsymbol{V}_i$ to have unit Frobenius norm to put the kernel matrices on the equal footing in penalty because the varying number of variants involved in each $\boldsymbol{V}_i$ leads to higher magnitude for sets with a large number of variants compared to those with a small number of variants.

While $\boldsymbol{V}_i$ measures genetic similarity between subjects in the i-th SNP group and is assumed fully known, it is worthwhile noting that no assumptions have been made about $\boldsymbol{\Sigma}_i$, which resides in the phenotype space and reflects how effect sizes of each variant on each phenotype are correlated. Different choices of $\boldsymbol{\Sigma}_i$ have been proposed in Dutta et al. (2019). If one does have *a priori* knowledge about phenotype structure, the algorithm simplifies to the univariate case. For example, if effect sizes of each variant in a SNP-set on different phenotypes are assumed homogeneous, we may write $\boldsymbol{\Sigma}_i = \sigma_i^2 \mathbf{1}_d\mathbf{1}_d^T$, where σ_i^2 is a scalar-valued i-th variance component and $\mathbf{1}_d$ is a $d \times 1$ vector of 1's. Then $\boldsymbol{\Omega} = \sum_{i=1}^m \sigma_i^2(\mathbf{1}_d\mathbf{1}_d^T \otimes \boldsymbol{V}_i) + \sigma_0^2(\mathbf{1}_d\mathbf{1}_d^T \otimes \boldsymbol{I}_{n-p})$, where σ_0^2 is a scalar-valued residual variance component. Since $(\mathbf{1}_d\mathbf{1}_d^T \otimes \boldsymbol{V}_i)$ is a known covariance matrix for i-th group, the problem amounts to estimating $\sigma_i^2, i = 0, 1, \ldots, m$.

2.3 Estimation algorithm

The MM principle involves majorizing the objective function $f(\boldsymbol{\theta})$ by a surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ around the current iterate $\boldsymbol{\theta}^{(t)}$ of a search (Lange et al., 2000; Hunter and Lange, 2004; Lange, 2016). The superscript t indicates the iteration number. Majorization is defined by the following two conditions

$$\begin{aligned} f(\boldsymbol{\theta}^{(t)}) &= g(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \\ f(\boldsymbol{\theta}) &\leq g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}), \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}^{(t)}. \end{aligned}$$

In other words, the surface $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ lies above the surface $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ and is tangent to it at the point $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Construction of the majorizing function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ constitutes the first M of the MM algorithm. The second M of the algorithm minimizes the surrogate $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ rather than $f(\boldsymbol{\theta})$. If $\boldsymbol{\theta}^{(t+1)}$ denotes the minimizer of $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, then this action forces the descent property $f(\boldsymbol{\theta}^{(t+1)}) \leq f(\boldsymbol{\theta}^{(t)})$. This fact follows from the inequalities

$$f(\boldsymbol{\theta}^{(t+1)}) \leq g(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \leq g(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) = f(\boldsymbol{\theta}^{(t)}),$$

reflecting the definition of $\boldsymbol{\theta}^{(t+1)}$ and the tangency condition. Monotonicity of MM iterates obliterates the need for line search and lends itself to the remarkable numerical stability of the MM algorithm.

We derive a majorizing function of the penalized loss function (2.3) by working on its three individual terms separately. For the penalty term, we first specialize to the lasso penalty then indicate the generalizations to other penalties.

1. Log-determinant term. The concavity of the map $\boldsymbol{X} \mapsto \ln\det \boldsymbol{X}$ and the supporting hyperplane inequality establish the majorization

$$\ln\det \boldsymbol{\Omega}^{(t)} + \mathrm{tr}[\boldsymbol{\Omega}^{-(t)}(\boldsymbol{\Omega} - \boldsymbol{\Omega}^{(t)})] \geq \ln\det \boldsymbol{\Omega}. \tag{2.4}$$

2. Quadratic form term. When $\boldsymbol{V}_i$ for all i are positive definite, hence invertible, convexity of the matrix function $(\boldsymbol{X}, \boldsymbol{Y}) \mapsto \boldsymbol{X}^T \boldsymbol{Y}^{-1} \boldsymbol{X}$ where $\boldsymbol{Y} \succ \boldsymbol{0}$ implies

$$\begin{aligned} \boldsymbol{\Omega}^{(t)} \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega}^{(t)} &= m \left(\frac{1}{m} \sum_{i=0}^{m} \boldsymbol{\Sigma}_i^{(t)} \otimes \boldsymbol{V}_i \right) \left(\frac{1}{m} \sum_{i=0}^{m} \boldsymbol{\Sigma}_i \otimes \boldsymbol{V}_i \right)^{-1} \left(\frac{1}{m} \sum_{i=0}^{m} \boldsymbol{\Sigma}_i^{(t)} \otimes \boldsymbol{V}_i \right) \\ &\preceq m \sum_{i=0}^{m} \frac{1}{m} (\boldsymbol{\Sigma}_i^{(t)} \otimes \boldsymbol{V}_i)(\boldsymbol{\Sigma}_i \otimes \boldsymbol{V}_i)^{-1} (\boldsymbol{\Sigma}_i^{(t)} \otimes \boldsymbol{V}_i) \\ &= \sum_{i=0}^{m} (\boldsymbol{\Sigma}_i^{(t)} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{(t)}) \otimes \boldsymbol{V}_i, \end{aligned} \tag{2.5}$$

or equivalently

$$\boldsymbol{\Omega}^{-1} \preceq \boldsymbol{\Omega}^{-(t)} \left[\sum_{i=0}^{m} (\boldsymbol{\Sigma}_i^{(t)} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{(t)}) \otimes \boldsymbol{V}_i \right] \boldsymbol{\Omega}^{-(t)}. \tag{2.6}$$

For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \preceq \boldsymbol{B}$ means $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite. The equality (2.5) follows from the identities $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$ and $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD})$. The nonsingularity assumption on $\boldsymbol{V}_i$ can be relaxed by substituting $\boldsymbol{V}_{\epsilon,i} = \boldsymbol{V}_i + \epsilon \boldsymbol{I}_n$ for $\boldsymbol{V}_i$ and sending ϵ to 0.

3. Lasso penalty term. The majorization on the lasso penalty

$$\sqrt{\mathrm{tr} \boldsymbol{\Sigma}_i^{(t)}} + \frac{1}{2\sqrt{\mathrm{tr} \boldsymbol{\Sigma}_i^{(t)}}} (\mathrm{tr} \boldsymbol{\Sigma}_i - \mathrm{tr} \boldsymbol{\Sigma}_i^{(t)}) \geq \sqrt{\mathrm{tr} \boldsymbol{\Sigma}_i} \tag{2.7}$$

follows from the concavity of the map $x \mapsto \sqrt{x}$ and the support hyperplane inequality.

Merging (2.4), (2.6) and (2.7) generates the overall majorizing function

$$\begin{aligned} g(\boldsymbol{\Sigma} \mid \boldsymbol{\Sigma}^{(t)}) =& \frac{1}{2} \sum_{i=0}^{m} \left\{ \mathrm{tr} \left[\boldsymbol{\Omega}^{-(t)} (\boldsymbol{\Sigma}_i \otimes \boldsymbol{V}_i) \right] + (\mathrm{vec} \boldsymbol{R}^{(t)})^T \left[(\boldsymbol{\Sigma}_i^{(t)} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{(t)}) \otimes \boldsymbol{V}_i \right] (\mathrm{vec} \boldsymbol{R}^{(t)}) \right\} \\ & + \frac{1}{2} \sum_{i=1}^{m} \frac{\lambda}{\sqrt{\mathrm{tr} \boldsymbol{\Sigma}_i^{(t)}}} \mathrm{tr} \boldsymbol{\Sigma}_i + c^{(t)}, \end{aligned} \tag{2.8}$$

where $\mathrm{vec}\, \boldsymbol{R}^{(t)} = \boldsymbol{\Omega}^{-(t)} \mathrm{vec}(\boldsymbol{Y})$ with $\boldsymbol{R}^{(t)}$ being a matrix of size $n \times d$ and $c^{(t)}$ is a constant impertinent to the parameters $\boldsymbol{\Sigma}_i$. Parameters $\boldsymbol{\Sigma}_i$ are nicely separated in (2.8) so we only need to minimize m individual functions

$$\begin{aligned} g_i^{(t)}(\boldsymbol{\Sigma}_i) &= \frac{1}{2} \left\{ \mathrm{tr} \left[\boldsymbol{\Omega}^{-(t)} (\boldsymbol{\Sigma}_i \otimes \boldsymbol{V}_i) \right] + \mathrm{tr}(\boldsymbol{R}^{(t)T} \boldsymbol{V}_i \boldsymbol{R}^{(t)} \boldsymbol{\Sigma}_i^{(t)} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{(t)}) + \frac{\lambda}{\sqrt{\mathrm{tr} \boldsymbol{\Sigma}_i^{(t)}}} \mathrm{tr} \boldsymbol{\Sigma}_i \right\} \\ &= \frac{1}{2} \left\{ \mathrm{tr} \left[\boldsymbol{\Omega}^{-(t)} (\boldsymbol{\Sigma}_i \otimes \boldsymbol{V}_i) \right] + \mathrm{tr}(\boldsymbol{\Sigma}_i^{(t)} \boldsymbol{R}^{(t)T} \boldsymbol{V}_i \boldsymbol{R}^{(t)} \boldsymbol{\Sigma}_i^{(t)} \boldsymbol{\Sigma}_i^{-1}) + \frac{\lambda}{\sqrt{\mathrm{tr} \boldsymbol{\Sigma}_i^{(t)}}} \mathrm{tr} \boldsymbol{\Sigma}_i \right\} \end{aligned} \tag{2.9}$$

to update $\boldsymbol{\Sigma}_i$. The first equation follows from the Kronecker identities $(\mathrm{vec}\boldsymbol{A})^T \mathrm{vec}\boldsymbol{B} = \mathrm{tr}(\boldsymbol{A}^T \boldsymbol{B})$ and $\mathrm{vec}(\boldsymbol{CDE}) = (\boldsymbol{E}^T \otimes \boldsymbol{C})\mathrm{vec}(\boldsymbol{D})$. The first trace in the second equation of (2.9) is linear in $\boldsymbol{\Sigma}_i$ with the coefficient of entry $(\boldsymbol{\Sigma}_i)_{jk}$ equal to

$$\mathrm{tr}(\boldsymbol{\Omega}_{jk}^{-(t)} \boldsymbol{V}_i) = \mathbf{1}_n^T (\boldsymbol{V}_i \odot \boldsymbol{\Omega}_{jk}^{-(t)}) \mathbf{1}_n,$$

where $\boldsymbol{\Omega}_{jk}^{-(t)}$ is the (j,k)-th $n \times n$ block of $\boldsymbol{\Omega}^{-(t)}$ and $\odot$ is the Hadamard (elementwise) product. The matrix $\boldsymbol{M}_i$ of these coefficients can be written as

$$\boldsymbol{M}_i \;=\; (\boldsymbol{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \boldsymbol{V}_i) \odot \boldsymbol{\Omega}^{-(t)}] (\boldsymbol{I}_d \otimes \mathbf{1}_n).$$

Setting the derivative of (2.9) to zeros yields the stationarity condition

$$\boldsymbol{M}_i + \frac{\lambda}{\sqrt{\text{tr}\boldsymbol{\Sigma}_i^{(t)}}}\boldsymbol{I}_d = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Sigma}_i^{(t)}\boldsymbol{\Sigma}_i^{-1}, \tag{2.10}$$

which is a Riccati equation admitting the explicit solution

$$\boldsymbol{\Sigma}_i^{(t+1)} = \boldsymbol{L}_i^{-(t)T}[\boldsymbol{L}_i^{(t)T}(\boldsymbol{\Sigma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Sigma}_i^{(t)})\boldsymbol{L}_i^{(t)}]^{1/2}\boldsymbol{L}_i^{-(t)}$$

in terms of the Cholesky factor $\boldsymbol{L}_i^{(t)}$ of the matrix on the left hand side of (2.10).

Algorithm 1 summarizes the MM algorithm for lasso penalized multivariate variance components model (VCSEL-M-lasso). Each iteration computes $m + 1$ Cholesky factorizations and symmetric square roots of $d \times d$ positive semidefinite matrices. In most applications, d is a small number. Our convergence criteria are based on the change in objective function (2.3) (the penalized negative log-likelihood function) values. The procedure is repeated until the relative change in the objective function value is less than a tolerance value ($10^{-6}\times$[|objective function value at the current iterate| + 1] by default). For tuning parameters, we first locate the tuning parameter λ value, after which all variance component estimates turn zero—which we denote maximum λ. Then we create a solution path using a set number of equidistant tuning parameter values from 0 to maximum λ.

Nonconvex penalties reduce the bias by applying less shrinkage to the large nonzero components. As an example, we illustrate with the MCP. An extra tuning parameter $\gamma > 1$ controls the concavity of the penalty. In our case where $\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)}$ is nonnegative, MCP is defined as

$$P_\gamma(\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)}; \lambda) = \begin{cases} \lambda\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} - \frac{\text{tr}(\boldsymbol{\Sigma}_i)}{2\gamma}, & \text{if } \sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if } \sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} > \gamma\lambda \end{cases}. \tag{2.11}$$

MCP converges to lasso penalty as $\gamma \to \infty$. When $\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} > \gamma\lambda$, (2.11) is a constant that

Input : $\boldsymbol{Y}$, $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m, \lambda$
Output: $\hat{\boldsymbol{\Sigma}}_0, \hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_m$
Initialize $\boldsymbol{\Sigma}_i^{(0)}$ positive definite, $i = 1, \ldots, m$
repeat
  $\boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^m \boldsymbol{\Sigma}_i^{(t)} \otimes \boldsymbol{V}_i + \boldsymbol{\Sigma}_0^{(t)} \otimes \boldsymbol{I}$
  $\boldsymbol{R}^{(t)} \leftarrow \text{reshape}(\boldsymbol{\Omega}^{-(t)}\text{vec}\boldsymbol{Y}, n, d)$
  for $i = 1, \ldots, m$ **do**
    Cholesky $\boldsymbol{L}_i^{(t)}\boldsymbol{L}_i^{(t)T} \leftarrow (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T[(\boldsymbol{1}_d\boldsymbol{1}_d^T \otimes \boldsymbol{V}_i) \odot \boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d \otimes \boldsymbol{1}_n) + \frac{\lambda}{\sqrt{\text{tr}\boldsymbol{\Sigma}_i^{(t)}}}\boldsymbol{I}_d$
    $\boldsymbol{\Sigma}_i^{(t+1)} \leftarrow \boldsymbol{L}_i^{-(t)T}[\boldsymbol{L}_i^{(t)T}(\boldsymbol{\Sigma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Sigma}_i^{(t)})\boldsymbol{L}_i^{(t)}]^{1/2}\boldsymbol{L}_i^{-(t)}$
  end
  Cholesky $\boldsymbol{L}_0^{(t)}\boldsymbol{L}_0^{(t)T} \leftarrow (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T[(\boldsymbol{1}_d\boldsymbol{1}_d^T \otimes \boldsymbol{I}_n) \odot \boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d \otimes \boldsymbol{1}_n)$
  $\boldsymbol{\Sigma}_0^{(t+1)} \leftarrow \boldsymbol{L}_0^{-(t)T}[\boldsymbol{L}_0^{(t)T}(\boldsymbol{\Sigma}_0^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{R}^{(t)}\boldsymbol{\Sigma}_0^{(t)})\boldsymbol{L}_0^{(t)}]^{1/2}\boldsymbol{L}_0^{-(t)}$
until *objective value converges*;

Algorithm 1: VCSEL algorithm for lasso penalized multivariate response variance component model (2.3) (VCSEL-M-lasso).

does not involve $\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)}$. Focusing on the region $\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} \le \gamma\lambda$ and using the concavity of the map $x \mapsto \sqrt{x}$, we obtain a majorization of the MCP penalty

$$\begin{aligned} P_\gamma(\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)}; \lambda) &= \lambda\sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} - \frac{\text{tr}(\boldsymbol{\Sigma}_i)}{2\gamma} \\ &\le \lambda \left(\sqrt{\text{tr}(\boldsymbol{\Sigma}_i^{(t)})} + \frac{1}{2\sqrt{\text{tr}(\boldsymbol{\Sigma}_i^{(t)})}} \left[\text{tr}(\boldsymbol{\Sigma}_i) - \text{tr}(\boldsymbol{\Sigma}_i^{(t)})\right] \right) - \frac{\text{tr}(\boldsymbol{\Sigma}_i)}{2\gamma} \\ &= \frac{1}{2} \left(\frac{\lambda}{\sqrt{\text{tr}(\boldsymbol{\Sigma}_i^{(t)})}} - \frac{1}{\gamma} \right) \text{tr}(\boldsymbol{\Sigma}_i) + c^{(t)}. \end{aligned} \tag{2.12}$$

Inequalities for log-determinant (2.4) and quadratic form terms (2.6) and (2.12) produce a

majorizing function for MCP penalty

$$g(\boldsymbol{\Sigma} \mid \boldsymbol{\Sigma}^{(t)}) = \frac{1}{2}\sum_{i=0}^{m}\left\{\text{tr}\left[\boldsymbol{\Omega}^{-(t)}(\boldsymbol{\Sigma}_i \otimes \boldsymbol{V}_i)\right] + \text{tr}(\boldsymbol{\Sigma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Sigma}_i^{(t)}\boldsymbol{\Sigma}_i^{-1})\right\} + \frac{1}{2}\sum_{i=1}^{m}\begin{cases}\left(\frac{\lambda}{\sqrt{\text{tr}\boldsymbol{\Sigma}_i^{(t)}}} - \frac{1}{\gamma}\right)\text{tr}\boldsymbol{\Sigma}_i & \text{if } \sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} \le \gamma\lambda \\ \gamma\lambda^2 & \text{if } \sqrt{\text{tr}(\boldsymbol{\Sigma}_i)} > \gamma\lambda\end{cases}, \tag{2.13}$$

which admits an analytical update similar to that for lasso. Algorithm 2 summarizes the MM algorithm for MCP penalized multivariate response variance component model (VCSEL-M-MCP).

Input : $\boldsymbol{Y}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_m, \lambda, \gamma$
Output: $\widehat{\boldsymbol{\Sigma}}_0, \widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_m$
Initialize $\boldsymbol{\Sigma}_i^{(0)}$ positive definite, $i = 1, \ldots, m$
repeat
  $\boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^{m}\boldsymbol{\Sigma}_i^{(t)} \otimes \boldsymbol{V}_i + \boldsymbol{\Sigma}_0^{(t)} \otimes \boldsymbol{I}$
  $\boldsymbol{R}^{(t)} \leftarrow \text{reshape}(\boldsymbol{\Omega}^{-(t)}\text{vec}\boldsymbol{Y}, n, d)$
  for $i = 1, \ldots, m$ **do**
    if $\sqrt{\text{tr}(\boldsymbol{\Sigma}_i^{(t)})} \le \gamma\lambda$ **then**
      Cholesky $\boldsymbol{L}_i^{(t)}\boldsymbol{L}_i^{(t)T} \leftarrow (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T[(\boldsymbol{1}_d\boldsymbol{1}_d^T \otimes \boldsymbol{V}_i) \odot \boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d \otimes \boldsymbol{1}_n) + \left(\frac{\lambda}{\sqrt{\text{tr}\boldsymbol{\Sigma}_i^{(t)}}} - \frac{1}{\gamma}\right)\boldsymbol{I}_d$
    else
      Cholesky $\boldsymbol{L}_i^{(t)}\boldsymbol{L}_i^{(t)T} \leftarrow (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T[(\boldsymbol{1}_d\boldsymbol{1}_d^T \otimes \boldsymbol{V}_i) \odot \boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d \otimes \boldsymbol{1}_n)$
    end
    $\boldsymbol{\Sigma}_i^{(t+1)} \leftarrow \boldsymbol{L}_i^{-(t)T}[\boldsymbol{L}_i^{(t)T}(\boldsymbol{\Sigma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Sigma}_i^{(t)})\boldsymbol{L}_i^{(t)}]^{1/2}\boldsymbol{L}_i^{-(t)}$
  end
  Cholesky $\boldsymbol{L}_0^{(t)}\boldsymbol{L}_0^{(t)T} \leftarrow (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T[(\boldsymbol{1}_d\boldsymbol{1}_d^T \otimes \boldsymbol{I}_n) \odot \boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d \otimes \boldsymbol{1}_n)$
  $\boldsymbol{\Sigma}_0^{(t+1)} \leftarrow \boldsymbol{L}_0^{-(t)T}[\boldsymbol{L}_0^{(t)T}(\boldsymbol{\Sigma}_0^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{R}^{(t)}\boldsymbol{\Sigma}_0^{(t)})\boldsymbol{L}_0^{(t)}]^{1/2}\boldsymbol{L}_0^{-(t)}$
until *objective value converges*;

Algorithm 2: VCSEL algorithm for MCP penalized multivariate response variance component model (2.3) (VCSEL-M-MCP).

2.4 Simulation studies

We conduct simulation studies to examine the selection performance of the proposed methods. We compare with R package Multi-SKAT (Dutta et al., 2019) for multivariate response model. Multi-SKAT is a marginal approach that tests one SNP-set at a time and makes a formal inference. This contrasts with our method that encompasses multiple SNP-sets in a joint model and provides rankings. For readers interested in the results on a univariate response, we summarize the results in Appendix A.2, in which we compare the selection performance of VCSEL to the group lasso. The group lasso is a group selection method designed for selecting fixed effects. Interestingly, the proposed penalized variance component model outperforms group lasso even when the data is generated from a fixed effects model, not to mention under a variance component model.

Both the lasso and MCP penalties are demonstrated for multivariate trait and interaction models. Unless otherwise specified, $\gamma = 2.0$ is used for the MCP penalty. We use the area under the Precision-Recall curve (auPRC) to evaluate performance. Similar to Receiver Operator Characteristic (ROC) curves, Precision-Recall (PR) curves (recall on the x-axis and precision on the y-axis) illustrate the tradeoff between precision and recall for varying cutoff values (Manning and Schütze, 1999; Raghavan et al., 1989). *Precision* is defined as the number of true positives over the total number of declared positives, while *recall* is defined as the number of true positives over the number of true positives plus the number of false negatives. A PR curve closer to the upper right corner, which corresponds to 100% precision and 100% recall, generally represents a better classifier. Since we want to take the influence of all cutoff values into account, we report auPRC, which is an aggregate measure of performance across all tuning parameter values and has a range of [0, 1]. An auPRC close to 1 indicates that the classifier returns accurate results (high precision) and most of all positive results (high recall).

Although ROC curves are the most popular metric for binary classifiers, PR curves

are more suitable when the class distribution is highly skewed, usually negative instances outnumbering positive instances (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). In fact, PR curves have been cited as an alternative in unbalanced datasets (Craven and Bockhorst, 2005; Bunescu et al., 2005; Davis et al., 2005; Goadrich et al., 2004; Kok and Domingos, 2005; Singla and Domingos, 2005). As we expect the number of positive variance components to be greatly exceeded by that of zero variance components, we deem auPRC to be an appropriate metric.

For the marginal testing method, we calculate the auPRC by ranking all genes by their p-values and assuming that each gene enters the solution path from smallest to largest. For example, the gene with the smallest p-value enters the solution path first, and the gene with the largest value would be the last one to enter the solution path.

For a sample of size n, we form genotype matrix $\boldsymbol{G}$ by randomly pairing $2n$ haplotypes drawn from a haplotype pool (SKAT.haplotypes in the SKAT R-package). The genotype values in matrix G are coded as 0, 1 and 2, representing the number of minor alleles while an additive genetic model is assumed. Assuming that there are m SNP-sets, we partition $\boldsymbol{G}$ into m submatrices of pre-specified window length:

$$\boldsymbol{G} = \left[\boldsymbol{G}_1 \middle| \boldsymbol{G}_2 \middle| \cdots \middle| \boldsymbol{G}_m\right],$$

where $\boldsymbol{G}_i \in \mathbb{R}^{n \times q_i}, i = 1, \ldots, m$, represents the i-th SNP-set.

We fix the number of positive variance components, excluding the residual variance component, to be 5. We calculate each auPRC over 100 tuning parameter values and report the average auPRCs along with their standard errors across 20 replicates.

2.4.1 Simulation studies for multiple traits

Here we compare selection performance of Algorithm 1 and MultiSKAT (Dutta et al., 2019) package in R. We generate three phenotypes ($n = 2000$, $d = 3$) from the following:

$$\text{vec}(\boldsymbol{Y}) = \text{vec}(\boldsymbol{XB}) + \boldsymbol{L}_{\boldsymbol{\Omega}}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}_{nd}), \tag{2.14}$$

where $\boldsymbol{L}_{\boldsymbol{\Omega}}$ is the lower triangular Cholesky factor of $\boldsymbol{\Omega} = \sum_{i=1}^{m} \boldsymbol{\Sigma}_i \otimes \boldsymbol{V}_i + \boldsymbol{\Sigma}_0 \otimes \frac{1}{\sqrt{n}}\boldsymbol{I}_n$. Depending on the genotype kernel, $\boldsymbol{V}_i$ equals to $\frac{1}{||\boldsymbol{G}_i\boldsymbol{W}_i\boldsymbol{W}_i\boldsymbol{G}_i^T||_F}\boldsymbol{G}_i\boldsymbol{W}_i\boldsymbol{W}_i\boldsymbol{G}_i^T$ (SKAT genotype kernel) or $\frac{1}{||\boldsymbol{G}_i\boldsymbol{W}_i\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}_i\boldsymbol{G}_i^T||_F}\boldsymbol{G}_i\boldsymbol{W}_i\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}_i\boldsymbol{G}_i^T$ (Burden test genotype kernel) where $\boldsymbol{W}_i$ is diagonal matrix whose entry equals to the weights $w_k = Beta(\text{MAF}_k; 1, 25)$ with MAF_k being the minor allele frequency of the k-th genetic variant (Wu et al., 2011). We use this weight since it is the default version in MultiSKAT package. We set $\boldsymbol{X}$ to be a $n \times 1$ matrix of 1s and $\boldsymbol{B}$ to be a $1 \times d$ matrix of 0.5s. For non-zero variance components $\boldsymbol{\Sigma}_i$, we incorporate two structures proposed in Dutta et al. (2019). The first choice is $\boldsymbol{\Sigma}_i = \boldsymbol{1}_d\boldsymbol{1}_d^T$, which implies that effect sizes of a variant on d different phenotypes are homogeneous. Hence it is called homogeneous kernel. The second structure is $\boldsymbol{\Sigma}_i = \boldsymbol{I}_d$, also known as heterogeneous kernel, which assumes that effect sizes of a variant on different phenotypes are heterogeneous or independent. Non-zero variance component matrices are spread across all m groups to create a scenario of low linkage disequilibrium (LD) between causal SNP-sets or variance components:

$$\boldsymbol{\Sigma}_i \quad = \quad \begin{cases} \boldsymbol{1}_d\boldsymbol{1}_d^T \text{ or } \boldsymbol{I}_d & \text{if } i = 1, 10, 20, 30, 40 \ (m = 40) \\ & \text{if } i = 1, 25, 50, 75, 100 \ (m = 100) \\ \boldsymbol{I}_d & \text{if } i = 0 \\ \boldsymbol{0} & \text{else.} \end{cases}$$

In this case, causal genes, or signal variance components are dispersed, hence there is little correlation among causal genes. One notable difference between VCSEL-M and Multi-SKAT

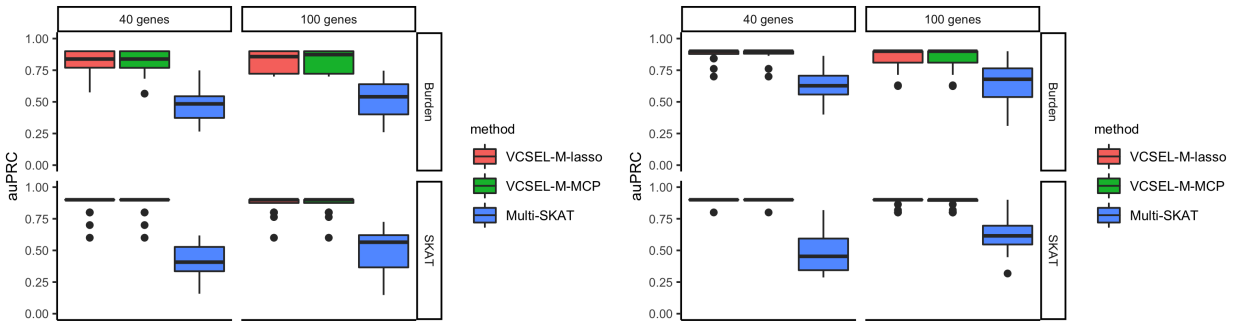

Figure 2.1: The auPRCs of VCSEL-M-lasso, VCSEL-M-MCP and Multi-SKAT under 40 and 100 genes and different genotype kernels for models with 6 non-zero variance components and 3 simulated traits ($d = 3$), using haplotype data from the SKAT R-package. The left and right panels assume $\boldsymbol{\Sigma}_i = \mathbf{1}_d\mathbf{1}_d^T$ and $\boldsymbol{\Sigma}_i = \boldsymbol{I}_d$, respectively, for non-zero variance components.

is that Multi-SKAT does not estimate $\boldsymbol{\Sigma}_i$ while VCSEL-M estimates $\boldsymbol{\Sigma}_i$. In fact, Multi-SKAT requires one to provide phenotype kernel structure, which is $\boldsymbol{\Sigma}_i$ in our notation, for testing association between a SNP-set and multiple phenotypes. In our simulations, we supply the ground truth $\boldsymbol{\Sigma}_i$, whether it be $\mathbf{1}_d\mathbf{1}_d^T$ or $\boldsymbol{I}_d$, when calling Multi-SKAT, hence giving an advantage to the Multi-SKAT method.

Figure 2.1 and Table 2.1 describe simulation results. Overall, our methods perform as well as Multi-SKAT, if not better. Despite having the ground truth $\boldsymbol{\Sigma}_i$ as an input argument, Multi-SKAT does not perform well when phenotype kernel has a homogeneous structure, as seen in the left panel of Figure 2.1.

Genotype kernel	Phenotype kernel	No. genes	VCSEL-M-lasso	VCSEL-M-MCP	MultiSKAT
$\boldsymbol{G}_i\boldsymbol{W}_i\mathbf{1}_{q_i}\mathbf{1}_{q_i}^T\boldsymbol{W}_i\boldsymbol{G}_i^T$ (Burden)	$\boldsymbol{\Sigma}_i = \mathbf{1}_d\mathbf{1}_d^T$	100 (2kb/gene)	0.82 (0.019)	0.82 (0.020)	0.52 (0.032)
		40 (5kb/gene)	0.82 (0.020)	0.82 (0.021)	0.48 (0.034)
	$\boldsymbol{\Sigma}_i = \boldsymbol{I}_d$	100 (2kb/gene)	0.84 (0.021)	0.84 (0.021)	0.65 (0.035)
		40 (5kb/gene)	0.87 (0.012)	0.88 (0.012)	0.63 (0.029)
$\boldsymbol{G}_i\boldsymbol{W}_i\boldsymbol{I}_{q_i}\boldsymbol{W}_i\boldsymbol{G}_i^T$ (SKAT)	$\boldsymbol{\Sigma}_i = \mathbf{1}_d\mathbf{1}_d^T$	100 (2kb/gene)	0.86 (0.017)	0.86 (0.017)	0.48 (0.041)
		40 (5kb/gene)	0.87 (0.018)	0.87 (0.018)	0.42 (0.030)
	$\boldsymbol{\Sigma}_i = \boldsymbol{I}_d$	100 (2kb/gene)	0.88 (0.008)	0.88 (0.009)	0.62 (0.031)
		40 (5kb/gene)	0.90 (0.005)	0.90 (0.005)	0.48 (0.038)

Table 2.1: The auPRCs of VCSEL-M-lasso, VCSEL-M-MCP, and Multi-SKAT across varying size and number of genes, using SKAT.haplotypes data from the SKAT R-package. In parentheses are standard deviation/$\sqrt{\text{no. replicates}}$.

2.5 Real data analysis

To test the multivariate response model, we apply our methods to the genetic data from the UK Biobank exome sequencing study (Sudlow et al., 2015). By doing so, we aim to identify genes associated with two quantitative lipid traits: high-density lipoprotein cholesterol (HDL-C) and low-density lipoprotein cholesterol (LDL-C). For the analysis, we only use measurements from the initial assessment visit. We regress each phenotype separately on age, age^2, sex, and the top five principal components and inverse normal transform respective residuals. The transformed residuals are used as our response variables. For our samples, we extract self-reported white British individuals (data field 21000: Ethnic background) with no genetic kinship to other participants (data field 22021: Genetic kinship to other participants) and without any medication for cholesterol, blood pressure, diabetes or exogenous hormones at baseline (data field 6153 and 6177: Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones). After removing individuals with missing values, we have 18,020 samples and genotype information of 8,959,608 variants, which are grouped into 26,395 genes based on the annotation information from SnpEff software (Cingolani et al., 2012) with GRCh38 human reference genome. We then remove monoallelic variants, common variants with MAF > 0.05, variants in sex chromosome from the analysis. Finally, we have the data of 18,020 individuals and genotype information of 4,312,036 low-frequency/rare (MAF ≤ 0.05) variants in 25,460 genes with at least three of those variants in each gene. Because the number of genes is too large, we first screen 25,460 genes down to 200 genes according to their p-values from Multi-SKAT omnibus approach that combines results across three pre-specified phenotype kernels (homogeneous, heterogeneous, and phenotype covariance kernels). Then we carry out a penalized estimation of the 200 variance components in the joint model (2.1) using the Burden test genotype kernel. This is akin to the sure independence screening strategy by Fan and Lv (2008), which entails large-scale screening accompanied by moderate-scale variable selection. Genes are ranked according to the order they appear in the solution path. Figure 2.2 illustrates the solution paths obtained

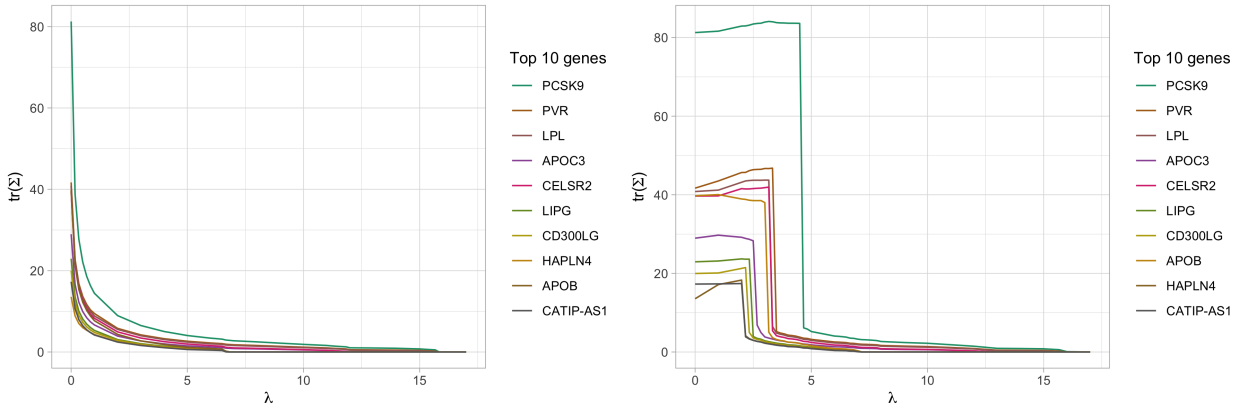

Figure 2.2: Solution paths of VCSEL-M-lasso (left) and VCSEL-M-MCP (right) methods in the analysis of 200 genes and two lipid measurements (HDL-C, LDL-C).

from VCSEL-M-lasso and VCSEL-M-MCP methods, along with their corresponding lists of the top ten genes in the order they appear in the solution path. Table 2.2 lists the top 10 genes together with their marginal p-values from Multi-SKAT. Most genes that are highly ranked by VCSEL methods—*PCSK9*, *PVR*, *LPL*, *APOC3*, *CELSR2*, *LIPG*, *CD300LG*, and *APOB* in the top 10 list—have their marginal test p-values under the false discovery rate (FDR) $< 5\%$ threshold and/or are known to play a role in modulating lipid levels (Benn et al., 2005; Cohen et al., 2005; Heid et al., 2008; Abifadel et al., 2009; Wallace et al., 2008; Tachmazidou et al., 2013; Lange et al., 2014; Holmen et al., 2014; Surakka et al., 2015). VCSEL methods identify genes that are not deemed significant by marginal testing but have association evidence in the literature. *HAPLN4* has been shown significant association with LDL-C and total cholesterol levels (Southam et al., 2017) and *APOC4* with HDL-C, LDL-C (Hoffmann et al., 2018; Wojcik et al., 2019).

2.6 Discussion

We defer the discussion until the end of Chapter 3.

Lasso Rank	MCP Rank	Gene	Marginal p-value	# Variants
1	1	*PCSK9*	3.37×10^{-20}	353
2	2	*PVR*	3.56×10^{-20}	111
3	4	*LPL*	5.73×10^{-18}	198
4	3	*APOC3*	2.04×10^{-7}	61
5	5	*CELSR2*	4.05×10^{-13}	986
6	6	*LIPG*	2.36×10^{-13}	225
7	7	*CD300LG*	6.56×10^{-10}	189
8	9	*HAPLN4*	2.86×10^{-3}	141
9	8	*APOB*	5.33×10^{-11}	947
10	10	*CATIP-AS1*	2.81×10^{-3}	16
11	11	*APOC4*	1.34×10^{-4}	74

Table 2.2: Top genes selected by the lasso and MCP penalized variance component model are tallied with their marginal p-values from the Multi-SKAT omnibus test in an association study of 200 genes and bivariate trait: HDL-C and LDL-C.

CHAPTER 3

Variance component selection for models with interaction terms

3.1 Introduction

Adverse drug reactions (ADR) pose a serious threat to public health, responsible for as many as 100,000 deaths and an estimated $136 billion annually in the United States according to U.S. Food & Drug Administration (2018). ADRs may occur from misuse–for example, inappropriate dosage, prolonged administration of a drug, or polypharmacy. Another important factor contributing to ADRs is genetics (He and Allen, 2010, and references therein).

Genomic differences among people place some individuals at grave risk of harm from certain medication while others may benefit from the same drug. For that reason, detecting those genetic variants that contribute to variability in treatment responses is the main objective in pharmacogenetic (PGx) studies. A number of methods have been proposed to test interaction effect or jointly test the genetic main effect and the interaction effect (Broadaway et al., 2015; Chen et al., 2014; Zhao et al., 2019; Yang et al., 2019; Zhang et al., 2020). However, they are limited to testing a single SNP-set. Hence, in this chapter, we outline an algorithm to incorporate SNP-set-treatment or -environment interaction terms in a univariate trait variance component model, motivated by pharmacogenomic studies.

If there are m genes under consideration, we have $2m + 1$ variance components in total, including the residual variance component, because each gene is associated with two variance components, one for the gene itself and the other for the interaction between gene and

treatment. For the i-th SNP-set, σ_{i1} and σ_{i2} denote the genetic effect and interaction effect variance components, respectively. Let $\boldsymbol{G}_i$ be the corresponding genotype matrix and $\boldsymbol{T} = \text{diag}(t_1, \ldots, t_n)$ be a diagonal matrix where $t_i \in \{0, 1\}$ indicates treatment status. Then linear weighted kernels associated with σ_{i1} and σ_{i2} are $\boldsymbol{V}_{i1} = \boldsymbol{G}_i \boldsymbol{W}_i \boldsymbol{G}_i^T$ and $\boldsymbol{V}_{i2} = \boldsymbol{T}\boldsymbol{G}_i \boldsymbol{W}_i \boldsymbol{G}_i^T \boldsymbol{T}^T$ respectively. The matrix $\boldsymbol{W}_i = \text{diag}(w_1, \ldots, w_q)$ contains the weights of the q variants in the i-th SNP-set. We remind readers that linear weighted kernels can be readily replaced by other choices of kernels. Note that $\boldsymbol{T}$ matrices are not limited to binary values. For example, one can swap diagonal entries in $\boldsymbol{T}$ matrix with environmental variable values, which are often continuous. Simulation studies 3.3 demonstrate this option of continuous values.

For a given response vector $\boldsymbol{y}$, the penalized loglikelihood augmented by group penalty on two variance components of each gene can be written as

$$f(\boldsymbol{\sigma}) \;=\; \frac{1}{2}\log\det \boldsymbol{\Omega}(\boldsymbol{\sigma}) + \frac{1}{2}\boldsymbol{y}^T[\boldsymbol{\Omega}(\boldsymbol{\sigma})]^{-1}\boldsymbol{y} + \sum_{i=1}^{m} P_\lambda(\sigma_{i1}, \sigma_{i2}), \tag{3.1}$$

where $\boldsymbol{\Omega}(\boldsymbol{\sigma}) = \sum_{i=1}^{m}(\sigma_{i1}^2 \boldsymbol{V}_{i1} + \sigma_{i2}^2 \boldsymbol{V}_{i2}) + \sigma_0^2 \boldsymbol{I}_n$ and $\boldsymbol{\sigma} = (\sigma_0, \sigma_{i1}, \sigma_{i2}, i = 1, \ldots, m)$ collects all $2m + 1$ variance components. We introduce two routes to constructing interaction models: 1) include/exclude main effects and interaction term together as a pair (VCSEL-I) and 2) enforce hierarchy restriction that only allows interaction term into the model when the corresponding main effect is included (VCSEL-Ih).

3.2 Estimation algorithm

3.2.1 All-in/all-out (VCSEL-I)

Often in the discovery phase, genetic main effect and gene-treatment interaction effect are jointly tested. This approach examines association between the trait of interest and genetic marker while accounting for gene-treatment interaction. To majorize the group lasso penalty on a pair of variance components, we apply the support hyperplane inequality to the concave

map $x \mapsto \sqrt{x}$

$$P_\lambda(\sigma_{i1}, \sigma_{i2}) = \lambda\sqrt{\sigma_{i1}^2 + \sigma_{i2}^2} \leq \frac{\lambda}{2}\frac{\sigma_{i1}^2 + \sigma_{i2}^2}{\sqrt{\sigma_{i1}^{(t)2} + \sigma_{i2}^{(t)2}}} + c^{(t)},$$

where $c^{(t)}$ is an irrelevant constant. Combining with the univariate case of inequalities (2.4) and (2.6), the surrogate function given t-th iterate $\boldsymbol{\sigma}^{(t)}$ is

$$\begin{aligned} g(\boldsymbol{\sigma}|\boldsymbol{\sigma}^{(t)}) &= \sum_{i=1}^{m}\sum_{j=1}^{2}\left[\frac{\sigma_{ij}^2}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_{ij}) + \frac{1}{2}\frac{\sigma_{ij}^{4(t)}}{\sigma_{ij}^2}\boldsymbol{y}^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_{ij}\boldsymbol{\Omega}^{-(t)}\boldsymbol{y} + \lambda\frac{\sigma_{ij}^2}{2\sqrt{\sigma_{i1}^{(t)2} + \sigma_{i2}^{(t)2}}}\right] \\ &\quad + \frac{\sigma_0^2}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}) + \frac{1}{2}\frac{\sigma_0^{4(t)}}{\sigma_0^2}\boldsymbol{y}^T\boldsymbol{\Omega}^{-2(t)}\boldsymbol{y}. \end{aligned}$$

Then the update $\sigma_{ij}^{(t+1)}$ for $i = 1, \ldots, m$ and $j = 1, 2$ is

$$\sigma_{ij}^{(t+1)} = \sigma_{ij}^{(t)}\left[\frac{\boldsymbol{y}^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_{ij}\boldsymbol{\Omega}^{-(t)}\boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_{ij}) + \lambda/\sqrt{\sigma_{i1}^{(t)2} + \sigma_{i2}^{(t)2}}}\right]^{1/4}.$$

Algorithm 3 summarizes the VCSEL algorithm for the all-in/all-out interaction with lasso penalty (VCSEL-I-lasso). A similar algorithm for MCP penalty (VCSEL-I-MCP) is summarised in Algorithm 4.

Input : $\boldsymbol{y}$, $\boldsymbol{V}_{11}, \boldsymbol{V}_{12}, \ldots, \boldsymbol{V}_{m1}, \boldsymbol{V}_{m2}, \lambda$
Output: $\hat{\sigma}_0^2, \hat{\sigma}_{11}^2, \hat{\sigma}_{12}^2, \ldots, \hat{\sigma}_{m1}^2, \hat{\sigma}_{m2}^2$
Initialize $\sigma_0^{(0)}, \sigma_{ij}^{(0)} > 0$, $i = 1, \ldots, m$, $j = 1, 2$
repeat
  $\boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^{m}(\sigma_{i1}^{2(t)}\boldsymbol{V}_{i1} + \sigma_{i2}^{2(t)}\boldsymbol{V}_{i2}) + \sigma_0^{2(t)}\boldsymbol{I}$
  $\sigma_{ij}^{(t+1)} \leftarrow \sigma_{ij}^{(t)}\left(\frac{\boldsymbol{y}^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_{ij}\boldsymbol{\Omega}^{-(t)}\boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_{ij}) + \lambda/\sqrt{\sigma_{i1}^{(t)2} + \sigma_{i2}^{(t)2}}}\right)^{1/4}, i = 1, \ldots, m,\ j = 1, 2$
  $\sigma_0^{(t+1)} \leftarrow \sigma_0^{(t)}\left(\frac{\boldsymbol{y}^T\boldsymbol{\Omega}^{-2(t)}\boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)})}\right)^{1/4}$
until *objective value converges*;

Algorithm 3: VCSEL algorithm with lasso penalty for selecting main effect and interaction effect variance components as a pair (VCSEL-I-lasso).

Input : $\boldsymbol{y}, \boldsymbol{V}_{11}, \boldsymbol{V}_{12}, \ldots, \boldsymbol{V}_{m1}, \boldsymbol{V}_{m2}, \lambda, \gamma$
Output: $\hat{\sigma}_0^2, \hat{\sigma}_{11}^2, \hat{\sigma}_{12}^2, \ldots, \hat{\sigma}_{m1}^2, \hat{\sigma}_{m2}^2$
Initialize $\sigma_0^{(0)}, \sigma_{ij}^{(0)} > 0$, $i = 1, \ldots, m,\ j = 1, 2$
repeat
  $\boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^m (\sigma_{i1}^{2(t)} \boldsymbol{V}_{i1} + \sigma_{i2}^{2(t)} \boldsymbol{V}_{i2}) + \sigma_0^{2(t)} \boldsymbol{I}$
  for $i = 1, \ldots, m,\ j = 1, 2$ **do**
    $$\sigma_{ij}^{(t+1)} \leftarrow \begin{cases} \sigma_{ij}^{(t)} \left(\dfrac{\boldsymbol{y}' \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{ij} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{ij}) + \frac{\lambda}{\sqrt{\sigma_{i1}^{(t)2} + \sigma_{i2}^{(t)2}}} - \frac{1}{\gamma}} \right)^{1/4}, & \text{if } \sqrt{\sigma_{i1}^{(t)} + \sigma_{i2}^{(t)}} \le \gamma\lambda \\ \sigma_{ij}^{(t)} \left(\dfrac{\boldsymbol{y}' \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{ij} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{ij})} \right)^{1/4}, & \text{otherwise} \end{cases}$$
  end
  $\sigma_0^{(t+1)} \leftarrow \sigma_0^{(t)} \left(\dfrac{\boldsymbol{y}^T \boldsymbol{\Omega}^{-2(t)} \boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)})} \right)^{1/4}$
until *objective value converges*;

Algorithm 4: VCSEL algorithm with MCP penalty for selecting main effect and interaction effect variance components as a pair (VCSEL-I-MCP).

3.2.2 Hierarchical interactions (VCSEL-Ih)

In the confirmation phase of gene-drug testing, interest lies in detecting gene-treatment interaction. Choi et al. (2010) argue that for easier interpretability, interaction terms should be included only if all corresponding main effects are in the model. We integrate this idea by assuming interaction effect variance component to be a constant multiple of genetic effect counterpart, i.e. $\sigma_{i2}^2 = \gamma_i \sigma_{i1}^2$. Whenever the variance component for i-th gene σ_{i1} is equal to 0, the interaction variance component σ_{i2} is automatically set to 0. Following Choi et al. (2010), we penalize both variance component σ_{i1} and interaction parameter γ_i. Then our objective function with lasso penalty becomes

$$f(\boldsymbol{\sigma}) \;=\; \frac{1}{2} \log\det \boldsymbol{\Omega} + \frac{1}{2} \boldsymbol{y}^T \boldsymbol{\Omega}^{-1} \boldsymbol{y} + \lambda_1 \sum_{i=1}^m \sigma_{i1} + \lambda_2 \sum_{i=1}^m \gamma_i,$$

where $\boldsymbol{\Omega} = \sum_{i=1}^m (\sigma_{i1}^2 \boldsymbol{V}_{i1} + \sigma_{i2}^2 \boldsymbol{V}_{i2}) + \sigma_0^2 \boldsymbol{I} = \sum_{i=1}^m (\sigma_{i1}^2 \boldsymbol{V}_{i1} + \gamma_i \sigma_{i1}^2 \boldsymbol{V}_{i2}) + \sigma_0^2 \boldsymbol{I}$. Both λ_1 and λ_2 are tuning parameters controlling the strength of the penalty terms.

The already familiar majorizations (2.4), (2.6) and (2.7) yields the surrogate function

$$\begin{aligned} g(\boldsymbol{\sigma} \mid \boldsymbol{\sigma}^{(t)}) &= \sum_{i=1}^{m} \left[\frac{\sigma_{i1}^2}{2} \mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i1}) + \frac{\gamma_i \sigma_{i1}^2}{2} \mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i2}) + \frac{1}{2} \frac{\sigma_{i1}^{4(t)}}{\sigma_{i1}^2} \boldsymbol{y}^T \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i1} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y} \right. \\ &\quad \left. + \frac{1}{2} \frac{\gamma_i^{2(t)} \sigma_{i1}^{4(t)}}{\gamma_i \sigma_{i1}^2} \boldsymbol{y}^T \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i2} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y} + \frac{\lambda_1}{2\sigma_{i1}^{(t)}} \sigma_{i1}^2 + \lambda_2 \gamma_i \right] \\ &\quad + \frac{\sigma_0^2}{2} \mathrm{tr}(\boldsymbol{\Omega}^{-(t)}) + \frac{1}{2} \frac{\sigma_0^{4(t)}}{\sigma_0^2} \boldsymbol{y}^T \boldsymbol{\Omega}^{-2(t)} \boldsymbol{y}. \end{aligned}$$

We adopt the block update strategy to decrease the objective value of $g(\boldsymbol{\sigma} \mid \boldsymbol{\sigma}^{(t)})$. Given $\gamma_i = \gamma_i^{(t)}$, we update σ_{i1} by

$$\sigma_{i1}^{2(t+1)} = \sigma_{i1}^{2(t)} \sqrt{\frac{\boldsymbol{y}^T \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i1} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y} + \gamma_i^{(t)} \boldsymbol{y}^T \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i2} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i1}) + \gamma_i^{(t)} \mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i2}) + \lambda_1 / \sigma_{i1}^{(t)}}}, \quad i = 1, \ldots, m.$$

Given $\sigma_{i1} = \sigma_{i1}^{(t+1)}$, we first update the covariance matrix

$$\widetilde{\boldsymbol{\Omega}}^{(t)} = \sum_{i=1}^{m} \left(\sigma_{i1}^{2(t+1)} \boldsymbol{V}_{i1} + \gamma_i^{(t)} \sigma_{i1}^{2(t+1)} \boldsymbol{V}_{i2} \right) + \sigma_0^{2(t+1)} \boldsymbol{I},$$

then update the i-th interaction parameter by

$$\gamma_i^{(t+1)} = \gamma_i^{(t)} \sqrt{\frac{\boldsymbol{y}^T \widetilde{\boldsymbol{\Omega}}^{-(t)} \boldsymbol{V}_{i2} \widetilde{\boldsymbol{\Omega}}^{-(t)} \boldsymbol{y}}{\mathrm{tr}(\widetilde{\boldsymbol{\Omega}}^{-(t)} \boldsymbol{V}_{i2}) + 2\lambda_2 / \sigma_{i1}^{2(t+1)}}}.$$

Summary of the algorithm for this hierarchical interaction selection method with lasso penalty (VCSEL-Ih-lasso) is provided in Algorithm 5.

3.3 Simulation studies

Here we compare selection performance of Algorithm 3, 4 and rareGE (Chen et al., 2014) package in R. Unlike the proposed method that models multiple SNP-sets, rareGE is a

Input : $\boldsymbol{y}, \boldsymbol{V}_{11}, \boldsymbol{V}_{12}, \ldots, \boldsymbol{V}_{m1}, \boldsymbol{V}_{m2}, \lambda_1, \lambda_2$
Output: $\hat{\sigma}_0^2, \hat{\sigma}_{11}^2, \hat{\sigma}_{21}^2, \ldots, \hat{\sigma}_{m1}^2, \hat{\gamma}_1, \ldots, \hat{\gamma}_m$

Initialize $\sigma_0^{(0)}, \sigma_{i1}^{(0)} > 0$, $i = 1, \ldots, m$

repeat

$\quad \boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^m (\sigma_{i1}^{2(t)} \boldsymbol{V}_{i1} + \gamma_i^{(t)} \sigma_{i1}^{2(t)} \boldsymbol{V}_{i2}) + \sigma_0^{2(t)} \boldsymbol{I}$

$\quad \sigma_{i1}^{2(t+1)} \leftarrow \sigma_{i1}^{2(t)} \sqrt{\frac{\boldsymbol{y}' \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i1} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y} + \gamma_i^{(t)} \boldsymbol{y}' \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i2} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_{i2}) + \lambda_1 / \sigma_{i1}^{(t)}}}, \ i = 1, \ldots, m$

$\quad \sigma_0^{2(t+1)} \leftarrow \sigma_0^{2(t)} \sqrt{\frac{\boldsymbol{y}' \boldsymbol{\Omega}^{-2(t)} \boldsymbol{y}}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)})}}$

$\quad \widetilde{\boldsymbol{\Omega}}^{(t)} \leftarrow \sum_{i=1}^m \left[\sigma_{i1}^{2(t+1)} \boldsymbol{V}_{i1} + \gamma_i^{(t)} \sigma_{i1}^{2(t+1)} \boldsymbol{V}_{i2} \right] + \sigma_0^{2(t+1)} \boldsymbol{I}$

$\quad \gamma_i^{(t+1)} \leftarrow \gamma_i^{(t)} \sqrt{\frac{\boldsymbol{y}' \widetilde{\boldsymbol{\Omega}}^{-(t)} \boldsymbol{V}_{i2} \widetilde{\boldsymbol{\Omega}}^{-(t)} \boldsymbol{y}}{\mathrm{tr}(\widetilde{\boldsymbol{\Omega}}^{-(t)} \boldsymbol{V}_{i2}) + 2\lambda_2 / \sigma_{i1}^{2(t+1)}}}, i = 1, \ldots, m$

until *objective value converges*;

Algorithm 5: VCSEL algorithm for selecting variance components, incorporating hierarchy of interaction terms (VCSEL-Ih-lasso).

marginal approach that tests one SNP-set at a time and makes a formal inference. As explained in Section 2.4, we use the auPRC as a metric to compare the proposed method and the rareGE. We generate a phenotype from

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{L}_{\boldsymbol{\Omega}} \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$$

where $n = 500$. Here covariate matrix $\boldsymbol{X}$ is a 500×3 matrix whose first column is a vector of 1's, second column is generated from $N(50, 5^2)$, and third column from $N(25, 4^2)$, which mimic covariate matrix in simulation studies of Chen et al. (2014). $\boldsymbol{L}_{\boldsymbol{\Omega}}$ is the lower triangular Cholesky factor of $\boldsymbol{\Omega} = \sum_{j=1}^2 \sum_{i=1}^m \sigma_{ij}^2 \boldsymbol{V}_{ij} + \frac{\sigma_0^2}{\sqrt{n}} \boldsymbol{I}_n$. Following the default option of rareGE package, we set

$$\begin{aligned} \boldsymbol{V}_{i1} &= \frac{1}{||\boldsymbol{G}_i \boldsymbol{W}_i \boldsymbol{G}_i^T||_F} \boldsymbol{G}_i \boldsymbol{W}_i \boldsymbol{G}_i^T \\ \boldsymbol{V}_{i2} &= \frac{1}{||\boldsymbol{E}\boldsymbol{G}_i \boldsymbol{W}_i \boldsymbol{G}_i^T \boldsymbol{E}_i||_F} \boldsymbol{E}\boldsymbol{G}_i \boldsymbol{W}_i \boldsymbol{G}_i^T \boldsymbol{E}, \end{aligned}$$

where $\boldsymbol{W}_i$ diagonal matrix whose entry equals to to the commonly used weights $\sqrt{w_k} = Beta(\text{MAF}_k; 1, 25)$ with MAF_k being the MAF of the k-th genetic variant (Wu et al., 2011). $\boldsymbol{E}$ is a diagonal matrix whose entries coincide with that of the second column in $\boldsymbol{X}$. $\boldsymbol{G}_i$ is a submatrix of genotype matrix we form from haplotypes data in R SKAT package, as explained in Section 2.4. We restrict $\boldsymbol{G}_i$ to only include SNPs with MAF less than 0.05 for fair comparison with rareGE method. This constraint leads to the number of SNPs ranging from 18 to 51 with a median of 33 for groups with window length of 5kb and that ranging from 3 to 29 with a median of 13 for groups with window length of 2kb. We set the effect strength of non-zero variance components to be 2.236. Two scenarios are simulated. The first is low LD setting:

$$\sigma_{i1} = \sigma_{i2} = \begin{cases} 2.236 & i = 1, 11, 20, 30, 40 \ (m = 40) \\ & i = 1, 26, 50, 75, 100 \ (m = 100) \\ 1.0 & i = 0 \\ 0.0 & \text{else.} \end{cases} . \tag{3.2}$$

The second is high LD setting, where the first 5 variance components are set to be non-zero:

$$\sigma_{i1} = \sigma_{i2} = \begin{cases} 2.236 & i = 1, 2, 3, 4, 5 \\ 1.0 & i = 0 \\ 0.0 & \text{else.} \end{cases} \tag{3.3}$$

In Supplementary Materials A.3, we quantify the correlations between SNP-sets in these high/low LD settings via the canonical correlation analysis. The true fixed effects parameter values are set to be $\boldsymbol{\beta} = (0.5, 0.1, 0.05)^T$.

As seen in Figure 3.1, VCSEL-I method is competitive against rareGE. The outperformance of VCSEL-I method is more dramatic under the low LD scenario, probably because

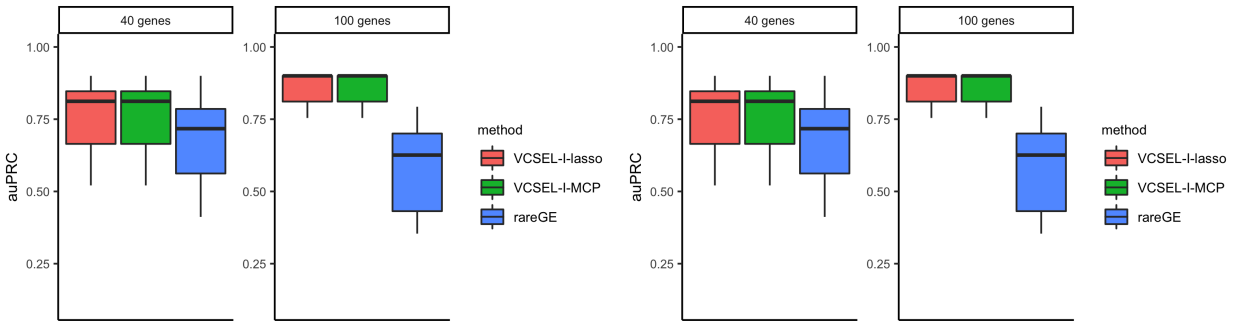

Figure 3.1: The auPRCs of VCSEL-I-lasso, VCSEL-I-MCP and rareGE under 40 and 100 genes for models with 6 non-zero variance components, using haplotype data from R SKAT package. True variance component values in the left panel mimic low LD scenario (3.2) while those in the right panel mimic high LD scenario (3.3).

the marginal test rareGE is not able to jointly model the multiple SNP-sets.

3.4 Real data analysis

Next, we apply our methods to the GWAS of Ezetimibe response in IMPROVE-IT (IMProved Reduction of Outcomes: Vytroin Efficacy International Trial), which is a phase 3b, multicenter, double-blind, randomized study to establish the clinical benefit and safety of Vytorin (Ezetimibe/Simvastatin tablet) versus Simvastatin mono-therapy in high-risk subjects (Cannon et al., 2015). In this PGx study using IMPROVE-IT clinical data, we are interested in discovering genes associated with 1) the efficacy of Vytorin treatment for 2,808 European patients who receive a greater benefit compared with the Simvastatin mono-therapy and 2) the joint efficacy of Ezetimibe/Simvastatin treatment and the Simvastatin mono-therapy treatment for 5,661 European patients. The endpoint for this gene-based variance component selection analysis is LDL-C fold-change at 1-month. The standard GWAS quality control and SNP imputation are conducted. We focus on the low frequency variants ($0.01 \leq \text{MAF} \leq 0.05$) after imputation (with imputation quality scores $r^2 > 0.5$) and putatively functional variants with consequences as non-synonymous, splice-site, non-sense, and frameshift variants annotated from the GEMINI software (Paila et al., 2013). Missing

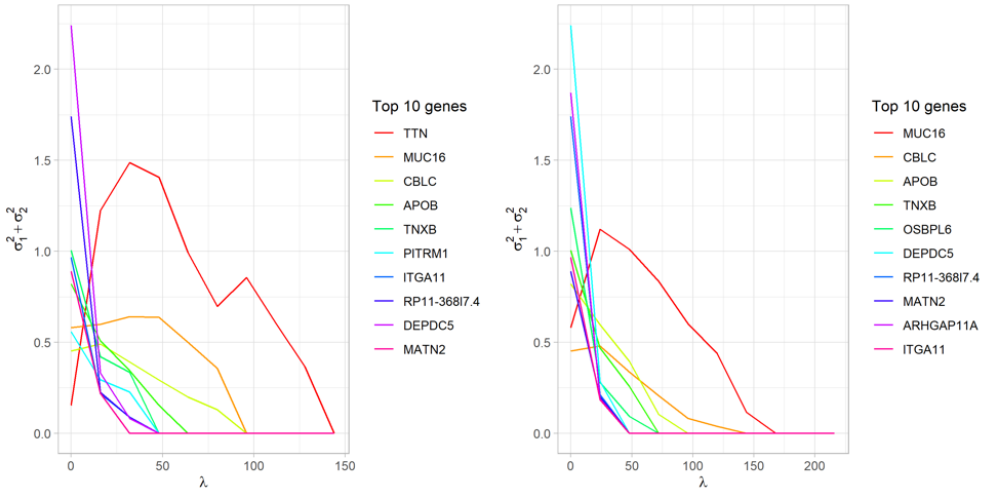

Figure 3.2: Solution paths of VCSEL-I-lasso (left) and VCSEL-I-MCP (right) methods in the analysis of 200 genes and the LDL-C response of all the patients receiving the Vytorin (Ezetimibe/Simvastatin tablet) treatment and Simvastatin mono-therapy in the IMPROVE-IT PGx study.

genotypes are imputed by their column mean. In total, there are 208,123 low frequency variants in 2,572 genes with at least two low frequency variants in each gene. The covariate matrix includes age, gender, prior lipid lowering therapy, early Acute Coronary Syndrome (ACS) trial, high risk ACS diagnosis, and the top five principal components calculated from the GWAS data to adjust for population structure. Because the number of genes is too large, we first screen the 2,572 genes down to 200 genes according to their marginal p-values from SKAT-O (Lee et al., 2012) for the analysis of Vytorin treatment effect and the other 200 genes according to their marginal p-values from the Composite Kernel Association Test (CKAT) (Zhang et al., 2020) for the analysis of Ezetimibe/Simvastatin treatment and the Simvastatin mono-therapy treatment joint effects. Then we analyze the two sets of the 200 genes by penalized estimation of the 200 variance components respectively.

Figure 3.2 illustrates the solution paths from VCSEL-I-lasso and VCSEL-I-MCP methods, along with their corresponding lists of the top ten genes in the order they appear in the solution path for the analysis of Ezetimibe/Simvastatin treatment and the Simvastatin mono-therapy treatment joint effects. The top five genes selected by the VCSEL-I-lasso

method are *TTN*, *MUC16*, *CBLC*, *APOB* and *TNXB*, and those selected by the VCSEL-I-MCP method are *MUC16*, *CBLC*, *APOB* and *TNXB* and *OSBPL6*. *CBLC* and *TNXB* are selected by both methods and have been shown to associate with statins response in literature. More specifically, similar as the *BCAM* gene, *CBLC* gene, close to *BCAM* gene, has been shown to associate with the response to statins (LDL-C change) and multiple non-drug-response LDL-C related traits as well (Postmus et al., 2014; Deshmukh et al., 2012). In addition, *TNXB* gene also shows significant association with the non-drug-response LDL-C trait in the literature.

3.5 Discussion

Chapter 2 and 3 provide a variance component selection framework for identifying SNP-sets associated with quantitative traits, particularly for multivariate traits and SNP-set-treatment interactions. Simulation studies and real data analyses have testified to the competitiveness of the proposed methods, compared to the traditional marginal tests.

Additionally, our VCSEL methods can adjust for sample relatedness by augmenting the model with a kinship matrix. More precisely, borrowing the notation of (2.2), the model becomes

$$\operatorname{vec} \boldsymbol{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_1 \otimes \boldsymbol{V}_1 + \cdots + \boldsymbol{\Sigma}_m \otimes \boldsymbol{V}_m + \boldsymbol{\Sigma}_g \otimes \boldsymbol{\Phi} + \boldsymbol{\Sigma}_0 \otimes \boldsymbol{I}_{n-p}),$$

where $\boldsymbol{\Phi}$ is the kinship matrix, and $\boldsymbol{\Sigma}_g$ is a matrix describing the shared heritability between the phenotypes. Along with the residual variance component $\boldsymbol{\Sigma}_0$, coheritability variance component $\boldsymbol{\Sigma}_g$ would remain in the model without any regularization.

While chiefly motivated by association testing in genetics, we envision the VCSEL methods to be applicable beyond genetics. For instance, in random effects ANOVA with many factors, each represented by a variance component, one may wish to select factors that are

relevant to the response. This ANOVA scenario has been alluded in Appendix A.2.

There are some limitations to the proposed methods, however. First, it is difficult to conduct formal inference on the selected SNP-sets. Second, it does not apply to biobank-scale data. We recommend this method for datasets of size up to $n \times d = 50,000$ where n is the number of samples and d is the number of traits. This is because VCSEL methods involve inverting the covariance matrix $\mathbf{\Omega}$ in each iteration, which is computationally expensive. Additionally, we do not suggest jointly fitting all 20,000-25,000 genes in the human genome using our method. We recommend that the number of genes is reduced before fitting the model by the sure independence screening strategy, which has been extensively studied and investigated (Fan and Lv, 2008).

For the VCSEL methods, we focus on the ranking of genes and report the overall selection performance by auPRC. In practice, the tuning parameters can be chosen according to the extended Bayesian information criteria (Chen and Chen, 2008). Future research for VCSEL methods should entail post-selection inference and investigation of the algorithms' theoretical properties and address the limitations mentioned in.

CHAPTER 4

Systematic heritability and heritability enrichment analysis for diabetes complications in ACCORD and UK Biobank studies

4.1 Introduction

Diabetes-related complications, which involve longstanding damage to microvascular and macrovascular systems, are a significant cause of morbidity and mortality among individuals with diabetes and impose a substantial burden on both the individual and society. Macrovascular complications, particularly atherosclerotic cardiovascular disease (CVD), are the leading cause of mortality among people with diabetes (American Diabetes Association, 2018). Microvascular complications affect kidney, eyes, or nerves and can lead to blindness, renal failure, or amputation, in addition to the devastating impact on quality of life. Long duration of diabetes and poor glycemic control are two primary risk factors for vascular complications (American Diabetes Association, 1998; Diabetes Control and Complications Trial Research Group, 1993). However, the development and progression of complications are heterogeneous even in individuals with comparable glucose control and diabetes duration (Bowden, 2002).

Early heritability studies have implicated genetic factors to explain the remaining heterogeneity in the development of diabetes-related complications. Among other diabetes-associated diseases, diabetic kidney disease (DKD) has been extensively studied in family

clustering studies. Diabetic siblings of probands with DKD had approximately 2-4 times the risk of developing DKD than diabetic siblings of probands free of DKD (Borch-Johnsen et al., 1992; Quinn et al., 1996; Harjutsalo et al., 2004). Heritability analysis of renal complication in type 1 diabetes estimates that 34–59% (adjusted for sex, diabetes duration, and age at diabetes onset; 24-42% unadjusted) of the variance is explained by common genetic variants, depending on the stages or phenotype definitions of DKD (Sandholm et al., 2017). A similar unadjusted analysis of DKD in individuals with type 2 diabetes estimates SNP heritability to be 8-12%, probably because of the phenotypic heterogeneity of kidney disease in type 2 diabetes (Van Zuydam et al., 2018).

Diabetic retinopathy is the leading cause of blindness in American adults aged 18-64 years (Centers for Disease Control and Prevention, 2020). Early family and twin studies suggested high concordance of diabetic retinopathy between family members (Diabetes Control and Complications Trial Research Group, 1997; Leslie and Pyke, 1982). Of note, genetic components for the risk of diabetic retinopathy appear related more to its severity of retinopathy, rather than to the simple presence or absence of retinopathy (Diabetes Control and Complications Trial Research Group, 1997; Hallman et al., 2005). Heritability estimates from family studies range from 18% to 52% (Looker et al., 2007; Hietala et al., 2008; Arar et al., 2008), while SNP heritability of severe diabetic retinopathy due to common genetic variants is estimated at 7% (Meng et al., 2018).

Little is known about the genetic contributions to CVD heritability among individuals with diabetes. The heritability of coronary artery disease in the general population is estimated to be between 40% and 60% in family and twin studies (Zdravkovic et al., 2002, 2007; McPherson and Tybjaerg-Hansen, 2016) and around 30% in studies of unrelated individuals using common genetic variants (Simonson et al., 2011). However, the only heritability-based studies for CVD in the diabetes population come from small family studies of quantitative traits, including coronary artery calcification (Wagenknecht et al., 2001), C-reactive protein levels (Lange et al., 2006), and carotid intima-media thickness (Lange et al., 2002).

Although previous studies have demonstrated genetic components to diabetes complications, they were limited to small sample sizes and lacked dense genetic markers, which neglected the potential contribution of low-frequency or rare variants (minor allele frequency (MAF) < 0.05). Genetic components reported using only family relationships had severe flaws and were subject to confounding by shared environmental effects (Tenesa and Haley, 2013). Early genetic studies were also characterized by differential definitions of complications due to phenotypic complexity, leading to a patchwork of findings.

In the present study, we conduct a systematic heritability analysis using two well-characterized cohorts—The Action to Control Cardiovascular Risk in Diabetes trial (ACCORD) and the UK Biobank (UKB) study—with high-quality imputed genetic markers to investigate genetic components involved in the development and progression of diabetes complications. The ACCORD study is a double-blind randomized clinical trial with clinically adjudicated complication outcomes, while the UKB provides an opportunity to adopt a prospective cohort study with larger sample size. In addition to estimating heritability using genotype and imputed data, we partition heritability by functional annotations using Stratified Linkage Disequilibrium Score regression (S-LDSC; Finucane et al., 2015). Since the heritability of complex traits is not distributed evenly across the whole genome (Maurano et al., 2012; Trynka et al., 2013), examining regions enriched for heritability can provide insights into functional categories that contribute to heritability. As the rising prevalence of diabetes has led to more people at an elevated risk of serious complications, elucidating the genetic component (i.e., heritability) to the development and progression of complications in a systematic manner can provide the rationale for genetic studies, which will ultimately enhance our ability to use precision medicine to tailor disease prevention/treatment.

4.2 Research design

4.2.1 Study design and participants

The ACCORD study is a double-blind, two-by-two factorial, randomized controlled, parallel treatment trial. In the trial, 10,251 participants were assigned intensive treatment targeting an HbA1c concentration of less than 6.0% (42.1 mmol/mol) or standard treatment targeting HbA1c of 7.0-7.9% (53-62.8 mmol/mol), in addition to assignments to distinct blood pressure and lipid interventions arms (Ismail-Beigi et al., 2010; Action to Control Cardiovascular Risk in Diabetes Study Group, 2008). It was designed to evaluate the effectiveness of a more aggressive treatment target to reduce the rate of macro and microvascular complications (Zoungas et al., 2017). The ACCORD study included type 2 diabetes participants with HbA1c concentrations of 7.5% (58.5 mmol/mol) or more, and who were aged 40-79 years with a history of cardiovascular disease or 55-79 years with evidence of significant atherosclerosis, albuminuria, left ventricular hypertrophy, or at least two risk factors for cardiovascular disease (dyslipidemia, hypertension, smoking, or obesity). Details of the design and principal results of the ACCORD trial were reported previously (Ismail-Beigi et al., 2010; Action to Control Cardiovascular Risk in Diabetes Study Group, 2008).

The UKB study recruited approximately 500,000 individuals aged between 40 and 69 in 2006-2010 from the general population across the United Kingdom. Participants answered detailed demographic, socioeconomic, and health-related questions using a touch screen questionnaire. Blood, urine, and saliva samples were collected, and physical measurements were taken from participants. Historical and follow-up information is provided by linking health and medical records. Genome-wide genotype data have been collected on all participants, creating many opportunities to discover new genetic associations and the genetic bases of complex traits (Fry et al., 2017; Bycroft et al., 2018). This large-scale cohort study with linked health and medical records enables studying of diabetes complication incidence under a prospective study design.

Within UKB, we curate a diabetes cohort based on the following categories: (A) baseline (2006-2010) and repeated assessments (2012-2013) at UKB assessment centers using questionnaires, physical measurements, and biological samples; (B) health-related records: hospital inpatient, death register, algorithmically-defined outcomes, first occurrences, and primary care data. Diabetes mellitus cases were ascertained according to the criteria: (A) first occurrence provided with fields 130706, 130708, and 130714. These fields take the first date of any of the following: International Classification of Disease, Ninth and Tenth Revision (ICD-9 and ICD-10) codes for type 1, type 2, and unspecified diabetes mellitus; self-report of diabetes mellitus at a UKB Assessment center along with the interpolated date from the age of diagnosis; and a limited number of primary care codes mapped to the three-digit ICD10 code: E10, E11, and E14; (B) the first occurrence of a more extensive list of diabetes-related primary care codes. Pregnancy or malnutrition-related diabetes was excluded. After excluding individuals with non-European ancestry, a total of 26,387 non-Hispanic white (NHW) diabetes patients were identified between the ages of 49 and 82 years as of 2020. We refer to it as the UKB-Diabetes cohort (Figure 4.2). Among the UKB-Diabetes cohort, we defined the first diabetes diagnosis as the index date and the incident case to be the first occurrence of the event after the index date.

4.2.2 Outcome definitions

In the ACCORD trial, all outcomes were pre-specified and adjudicated by the outcome committee. The pre-specified ACCORD primary cardiovascular (CVD) outcome (i.e., 3-point Adverse Cardiovascular Events (MACE) including CVD mortality, nonfatal MI, and nonfatal stroke) was the first occurrence of nonfatal myocardial infarction (MI) or nonfatal stroke or death from cardiovascular causes. We expanded this primary CVD outcome by including individual outcomes of new or worsening congestive heart failure (CHF), total stroke, and major coronary heart disease (CHD). For microvascular complications, we included a broader combination of microvascular outcomes from more severe (e.g., Neph3 and Retin1)

to less advanced conditions. The following outcomes record incident cases.

- Neph1: Doubling of baseline serum creatinine or 20 mL/min per 1.73 m^2 decrease in estimated GFR.
- Neph2: Development of macroalbuminuria. UACR 33.9 mg/mmol.
- Neph3: End-stage renal disease (ESRD, i.e., initiation of dialysis or a rise of serum creatinine to 3.3 mg per deciliter (292 mol/L)).
- Neph4: Development of Neph1, Neph2, or Neph3.
- Neph5: Development of microalbuminuria. UACR 3.4 mg/mmol.
- Retin1: Retinal photocoagulation or vitrectomy to treat retinopathy.
- Retin2: Eye surgery for cataract extraction.
- Retin3: Three-line change in visual acuity.
- Retin4: Severe vision loss (Snellen fraction 20/200).

A detailed description of the pre-specification of the ACCORD outcomes was documented previously (Ismail-Beigi et al., 2010; Action to Control Cardiovascular Risk in Diabetes Study Group, 2008).

Among the UKB-Diabetes cohort, we defined incident cases as the first occurrence of the event after the first diabetes diagnosis. For each type of incident cases, we excluded (A) individuals with documented occurrences of the event before the diagnosis of diabetes (B) individuals with no HbA1c measures after a diabetes diagnosis. We defined control cases as those with no event of interest occurred during the entire observation period with at least five years of follow-up. Key phenotypes are detailed below:

- CVD: Composite for CVD. Either MI, Ischemic stroke, unstable angina, or PCI.

- MI: Myocardial infarction from self-reported, primary care, hospital admissions, or death records. Controls were required to have no evidence of certain cardiovascular diseases.
- PCI: Percutaneous coronary intervention.
- Stroke any: Either ischemic, hemorrhagic, or unspecified stroke.
- Stroke infarct: Ischemic stroke.
- DKD: Chronic/diabetic kidney disease in self-reported, primary care, hospital or death records.
- DR: Composite for diabetic eye disease in self-reported, primary care, or hospital admission records.
- Macroalbuminuria: Urine ACR 33.9 mg/mmol at either UKB visit.
- Microalbuminuria: Urine ACR 3.4 mg/mmol at either UKB visit.

A complete list of codes and data fields used in the definition of diabetes mellitus, diabetes complications, and their date of the first occurrence are found in Appendix B.1.

4.2.3 Genotyping and imputation in ACCORD and UKB

Detailed accounts on DNA extraction, genotyping, and quality control (QC) procedures in ACCORD were reported previously (Shah et al., 2016). After retrieving the ACCORD genetic study data from dbGap (Study Accession: phs001411.v1.p1), we used genetic variants genotyped on Affymetrix Axiom Biobank chips from the University of North Carolina (UNC) and merged data under two different institutional review board (IRB) protocols—HMB-IRB (73941) and DS-CDKD-IRB (73944). There were 6,291 (2,335 females and 3,956 males) with 546,800 SNPs in the merged dataset. Based on self-reported ethnicity, there were 4,369

NHW, 935 African-Americans (AA), 381 Hispanics, and 606 others. After pre-imputation QC steps, imputation was performed on the genotype data using a two-step approach: pre-phasing genotype calls and imputation. After discarding imputed SNPs with $R^2 < 0.3$ and MAF < 0.0003, the total number of SNPs was 25,667,109. Additional details on imputation procedures are provided in Appendix B.2.1 and B.2.3.

We analyzed the genotyping and imputation (version 3) data released by the UKB in 2017. Details on genotyping and imputation have been extensively described elsewhere (Bycroft et al., 2018). In summary, genome-wide genotyping was performed on all UKB participants using the UK Biobank Axiom array. Around 850,000 variants were directly genotyped, and more than 90 million variants were imputed using the merged UK10K and 1000 Genomes Phase 3 reference panels (1000 Genomes Project Consortium, 2015). Only autosomal SNPs were included for both genotype and imputed data analyses. In the analyses involving imputation data, we discarded SNPs with imputation info score > 0.3, missing genotype rate > 0.05, Hardy-Weinberg equilibrium test $p < 1 \times 10^{-6}$ and MAF < 0.0001, yielding a total of 33,932,888 autosomal SNPs.

4.3 Statistical analysis

4.3.1 Overview of methods

First, we computed a Genetic Relationship Matrix (GRM) from all autosomal SNPs in genotype data using the Relatedness Estimation in Admixed Populations (REAP) approach (Thornton et al., 2012). Then we selectively excluded one of any pair of individuals with an estimated kinship greater than the separation between full and half-siblings (estimated kinship $(1/2)^{5/2} = 0.1768$) in a way to maximize the remaining sample size (Manichaikul et al., 2010). This step was done to avoid inflation caused by cryptic relatedness. After the pruning step, we estimated heritability on the NHW samples. Based on the GRM constructed from the REAP, heritability was computed using the GREML-SC (single component GREML)

approach (Yang et al., 2010) via the software package Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011a). We adjusted for relevant covariates such as the top genetic principal components, age, or sex. Additionally, we measured genetic correlations among binary traits using the GCTA software (Yang et al., 2011a).

Next, we used several approaches to calculate the narrow sense heritability of diabetes complications from imputed datasets. First, we applied GREML-LDMS-I (Yang et al., 2015), the multiple variance components approach that bins SNPs by MAF crossed by their individual levels of linkage disequilibrium (LD). We selected this approach over other multicomponent approaches such as GREML-LDMS-R, which allocates SNPs by the MAF and regional LD, since GREML-LDMS-I was shown to be the least biased method (Evans et al., 2018). For the GREML-LDMS-I approach, we followed the design laid out in Evans et al. (2018). First, we calculated segment-based LD scores using the default settings in the GCTA software and stratified SNPs into high LD and low LD score groups using the median as a threshold. In each LD group, SNPs were further partitioned into four MAF bins. Then GRMs were computed for each of the eight groups of SNPs. Finally, we estimated the heritability of each binary phenotype with fixed covariates.

We also applied S-LDSC (Finucane et al., 2015, 2018), a method for partitioning heritability from genome-wide association study (GWAS) summary statistics. After acquiring statistics from logistic regression, we performed none cell type-specific and tissue type-specific heritability enrichment analyses. In none cell type-specific analyses, we used 53 overlapping functional categories used in Finucane et al. (2015). In tissue-type specific analyses, we used the specifically expressed gene annotations generated by Finucane et al. (2018) with the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2015). For all S-LDSC analyses, we used 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium, 2015) European population SNPs as an LD reference panel. All annotations and reference panel data were obtained from Alkes Price's group data repository (see URLs in Appendix B.4). For more details on methods, see Appendix B.2.

Characteristic	N=4318
Age at baseline, years*	63.2 ± 6.4
Years since diabetes diagnosis*	10.7 ± 7.4
HbA1c at baseline, %*	8.2 ± 1.0
Sex, %	
Female	34.4
Male	65.6
Smoked cigarettes in last 30 days, %	
Yes	12.4
No	87.6
Smoked >100 cigarettes during lifetime, %	
Yes	50.4
No	37.9
NA	11.7
CVD history at baseline, %	
Yes	36.1
No	63.9
Glycemic treatment arm, %	
Intensive	49.8
Standard	50.2

Table 4.1: Sample characteristics of the non-Hispanic white participants used in the ACCORD analyses. * Denotes mean ± standard deviation.

Characteristic	N=26387
Age in 2010, years*	60.9 ± 7.0
Age of first DM, years*	56.4 ± 12.4
BMI*	31.5 ± 5.7
HbA1c at initial visit, mmol/mol*	48.8 ± 13.3
HbA1c at repeat visit, mmol/mol*	48.6 ± 11.1
Sex, %	
Male	61.4
Female	38.6
Current/former smoking, %	
Yes	56.1
No	43.4
Missing	0.5
DM Type	
Type 1	2.9
Type 2	69.0
Unspecified	28.1

Table 4.2: Sample characteristics of the non-Hispanic white participants used in the UKB analyses. The initial visit indicates anytime between 2006 to 2010, depending on the individual. * Denotes mean ± standard deviation. DM, Diabetes mellitus.

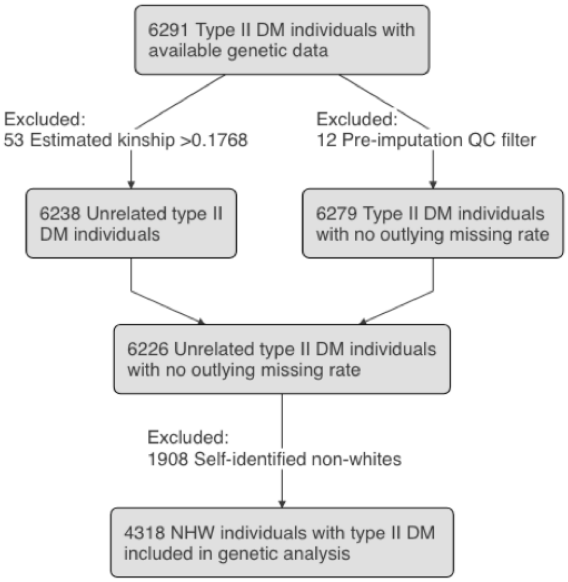

Figure 4.1: Diagram depicting a flow of participants used in the ACCORD analyses. DM, diabetes mellitus; NHW, none-Hispanic white.

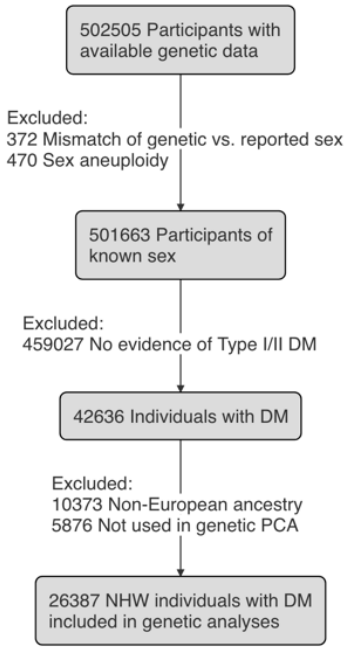

Figure 4.2: Diagram depicting a flow of participants used in the UK Biobank analyses. DM, diabetes mellitus; NHW, none-Hispanic white.

4.4 Results

Table 4.1 and Table 4.2 describe characteristics of the NHW samples used in the ACCORD and the UK Biobank analyses, respectively. Figure 4.1 and 4.2 show a breakdown of participant flow in the ACCORD and UKB analyses, respectively.

4.4.1 Heritability

First, we computed the heritability of phenotypes from the SNPs on the genotyping array using the GREML-SC approach (Yang et al., 2010). After pruning related individuals and extracting NHW samples, there remained 4,318 samples for the ACCORD and 26,387 samples for the UKB.

For the ACCORD data, we adjusted for sex, age at baseline, history of CVD at baseline, and the top five principal components. Heritability estimates for the ACCORD data are displayed as purple bars in Figure 4.3 (also see Table 4.3). Except for the phenotype primary, heritability estimates of the phenotypes are below 0.2. The estimate for the composite nephropathy outcome among type 2 diabetes (Neph4), calculated from the SNPs on the genotyping array, is 0.129 (SE 0.091), which is comparable with estimates from a similar analysis (0.12 for chronic kidney disease and 0.08 for diabetic kidney disease among type 2 diabetes subjects) (Van Zuydam et al., 2018). We also ran an additional GREML-SC analysis that includes interaction with intensive glycemic treatment arm (see Figure B.6 in Appendix B.3.

For the UKB data, the following covariates were accounted for: sex, age in 2010, and the top ten principal components. Heritability estimates from the UKB genotype data are illustrated as purple bars in Figure 4.4 (also see Table 4.4). Heritability estimates from the UKB genotype data tend to have smaller error bars than those from the ACCORD genotype data due to the larger sample size in the UKB dataset. We also observe that some corresponding phenotypes have quite disparate estimates. While the composite CVD

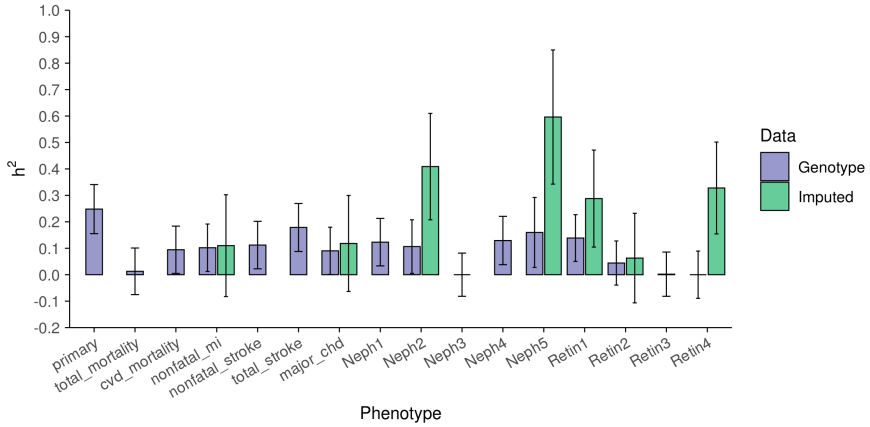

Figure 4.3: Heritability estimates and standard errors of diabetes complications using the ACCORD data. Estimates from genotype data are obtained using the GREML-SC approach. Estimates from the imputed data are using the GREML-LDMS-I method. Following covariates are adjusted for: sex, age at baseline, CVD history at baseline, and the top five genetic principal components.

phenotype from the ACCORD (primary) is 0.248 (SE 0.093), the composite CVD outcome from UKB is 0.081 (SE 0.028). This may stem from the discrepancy in sample characteristics between the two cohorts. While the UKB-Diabetes cohort is younger and relatively healthy, the ACCORD group is at high risk of CVD with a longer duration of diabetes.

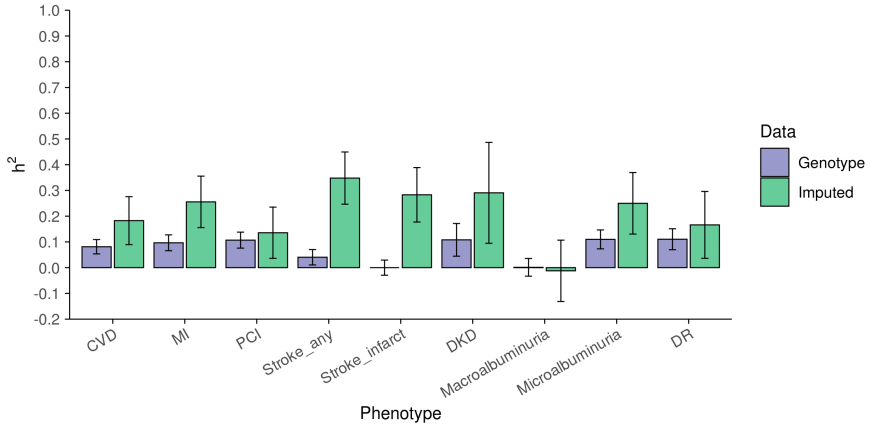

Figure 4.4: Heritability estimates and standard errors of diabetes complication outcomes using the UKB data. Estimates from genotype data are obtained using the GREML-SC approach. Estimates from the imputed data are using the GREML-LDMS-I method. Following covariates are adjusted for: sex, age in 2010, and the top ten genetic principal components. Note that the estimates are on the observed scale.

Phenotype	Proportion of cases in the sample	N	V(G)/V(p) (SE)	
			GREML-SC	GREML-LDMS
Primary	0.106	4318	0.248 (0.093)	NA
Total mortality	0.066	4318	0.013 (0.088)	NA
CVD mortality	0.028	4318	0.094 (0.089)	NA
Nonfatal MI	0.071	4318	0.102 (0.090)	0.110 (0.192)
Nonfatal stroke	0.015	4318	0.112 (0.090)	NA
Total stroke	0.018	4318	0.179 (0.091)	NA
Major CHD	0.129	4318	0.090 (0.089)	0.118 (0.181)
Neph1	0.591	4318	0.123 (0.090)	NA
Neph2	0.070	3866	0.106 (0.101)	0.409 (0.201)
Neph3	0.028	4318	0.000 (0.082)	NA
Neph4	0.616	4318	0.129 (0.091)	NA
Neph5	0.241	2912	0.160 (0.132)	0.596 (0.254)
Retin1	0.084	4318	0.139 (0.088)	0.288 (0.183)
Retin2	0.158	4318	0.044 (0.083)	0.063 (0.169)
Retin3	0.360	4318	0.002 (0.084)	NA
Retin4	0.068	4318	0.000 (0.089)	0.328 (0.174)

Table 4.3: GREML-SC and GREML-LDMS estimates using the ACCORD genotype and imputed data, respectively. NA under GREML-LDMS, the GREML analysis failed to run due to the small sample size. V(G)/V(p), proportion of phenotypic variance explained by genotypes, i.e., heritability, as observed in the study population. SE, standard error.

Phenotype	Proportion of cases in the sample	N	V(G)/V(p) (SE)	
			GREML-SC	GREML-LDMS
CVD	0.159	17540	0.081 (0.028)	0.183 (0.093)
MI	0.094	16310	0.097 (0.031)	0.256 (0.100)
PCI	0.090	16252	0.107 (0.031)	0.136 (0.100)
Stroke any	0.087	16002	0.041 (0.030)	0.348 (0.101)
Stroke infarct	0.042	15429	0.000 (0.029)	0.283 (0.106)
DKD	0.256	7707	0.108 (0.064)	0.291 (0.196)
Macroalbuminuria	0.029	13246	0.001 (0.034)	0.000 (0.119)
Microalbuminuria	0.238	13246	0.110 (0.037)	0.250 (0.119)
DR	0.541	11739	0.110 (0.041)	0.166 (0.130)

Table 4.4: GREML-SC and GREML-LDMS estimates using the UKB genotype and imputed data, respectively. V(G)/V(p), proportion of phenotypic variance explained by genotypes, i.e., heritability, as observed in the study population. SE, standard error.

Note that the h^2 estimates are on the observed scale, which does not take population prevalence into account. We display observed scale estimates because ACCORD and UKB-Diabetes studies are intervention or study with a prospective design, not ascertained case-control studies, in which the proportion of cases are often overrepresented. In fact, the sample proportion of cases and prevalence do not deviate much from each other for phenotypes with population prevalence available in the literature. For example, the proportion of DKD cases in the UKB-Diabetes cohort is 0.256, which is similar to the prevalence of any diabetic kidney disease among US adults with diabetes (0.262; 95% CI, 22.6-29.9) reported in Afkarian et al. (2016). The proportions of incident cases for primary CVD outcome and total stroke in the ACCORD group are 0.106 and 0.018, respectively, while hospital discharges record in 2016 reported that 75.3 and 13.6 per 1,000 adults with diabetes had major CVD and stroke, respectively (Centers for Disease Control and Prevention, 2020).

Estimates of the genetic correlation between selected traits are presented in Table 4.5 for the ACCORD. The standard errors are large for most pairs of traits, leading to confidence intervals including 0 and suggesting a lack of power. Despite the large standard errors, we observe high correlation estimates between Retin1 and Neph2/Neph5, which agree with the finding that links diabetic retinopathy with renal function (Xu et al., 2012). For the UKB estimates, see Table B.1 in Appendix B.3. We observe some inconsistencies between the

	Neph1	Neph2	Neph4	Neph5	Retin1
primary	-0.54 (0.41)	0.12 (0.43)	-0.53 (0.40)	0.14 (0.42)	-0.52 (0.39)
Neph1		0.47 (0.62)	0.96 (0.05)	0.25 (0.57)	0.19 (0.50)
Neph2			0.49 (0.57)	0.11 (0.62)	0.70 (0.67)
Neph4				0.33 (0.56)	0.32 (0.51)
Neph5					0.52 (0.59)

Table 4.5: Genetic correlation estimates and the standard errors between selected phenotypes using the ACCORD genotype data. Adjusted for sex, CVD history at baseline, age at baseline, and the top five genetic principal components.

results of the two datasets. While the genetic correlation between the CVD composite outcome and retinopathy outcome from the ACCORD data (primary-Retin1) shows a negative correlation, that from the UKB data (CVD-DR) shows a positive correlation. This is most likely due to the difference in phenotype definition.

On the imputed datasets, we employ the GREML-LDMS-I method. Heritability estimates of the phenotypes are provided as green bars in Figure 4.3 and 4.4 for the ACCORD and the UKB, respectively (also see Table 4.3 and Table 4.4). The heritability of diabetic kidney disease is estimated to be 0.29. Microalbuminuria estimates range from 0.24 to 0.60, while macroalbuminuria estimates are up to 0.41. Heritability estimates of diabetic retinopathy range from 0.06 to 0.33, depending on the definition of phenotype. Although less than family study estimates for broad-sense heritability—0.27 for diabetic retinopathy (Arar et al., 2008) and as high as 0.52 (SE 0.31) for proliferative retinopathy among adults with type 1 diabetes (Hietala et al., 2008), our estimates are still comparable to pedigree heritability estimates.

Of note, we observe higher estimates with more advanced retinopathy: 0.29 and 0.33 for Retin1 (retinal photocoagulation or vitrectomy) and Retin4 (severe vision loss), respectively, as opposed to 0.06 and around 0 for Retin2 (cataract extraction) and Retin3 (three-line change in visual acuity), respectively. On the other hand, diabetic nephropathy phenotypes do not exhibit such a pattern. While the heritability of macroalbuminuria phenotype from ACCORD is estimated at 0.41, that of microalbuminuria from ACCORD is at 0.60, re-

spectively. Estimates for either Neph1 or Neph3 are unavailable despite larger sample sizes (4,318 for both Neph1 and Neph3; 3,866 and 2,912 for Neph2 and Neph5, respectively). This pattern or lack thereof is consistent with earlier heritability studies that insinuated genetic components to the severity of diabetic retinopathy and presence/absence of diabetic nephropathy (Diabetes Control and Complications Trial Research Group, 1997; Hallman et al., 2005). Although we cannot confirm the trend of diabetic retinopathy in the UKB data (due to the absence of granular outcome definition for diabetic retinopathy), estimates for diabetic nephropathy from the UKB study (0.25 for microalbuminuria and close to 0 for macroalbuminuria) are in accordance with the pattern seen in the ACCORD study.

Heritability analyses using imputed data reveal a substantial contribution of low frequency/rare variants to the predisposition for complications. While the heritability of severe diabetic retinopathy from common genetic variants among individuals with type 2 diabetes was estimated to be 0.07 in a previous study (Meng et al., 2018) and up to 0.14 in our analysis (see GREML-SC results of Retin1 and Retin4 in Table 4.3), we observe higher heritability estimates of advanced diabetic retinopathy among type 2 diabetes individuals (0.29 and 0.33 for Retin1 and Retin4 in ACCORD) calculated from directly typed and imputed genetic markers. The distribution of heritability across the MAF spectrum for other complication phenotypes, including retinopathy, is found in Figure 4.5. Notably, the UKB results (Figure 4.6) show a more pronounced contribution pattern with small error bars and heritability heavily concentrated in very rare variants ($0.0003 \leq \text{MAF} < 0.0025$).

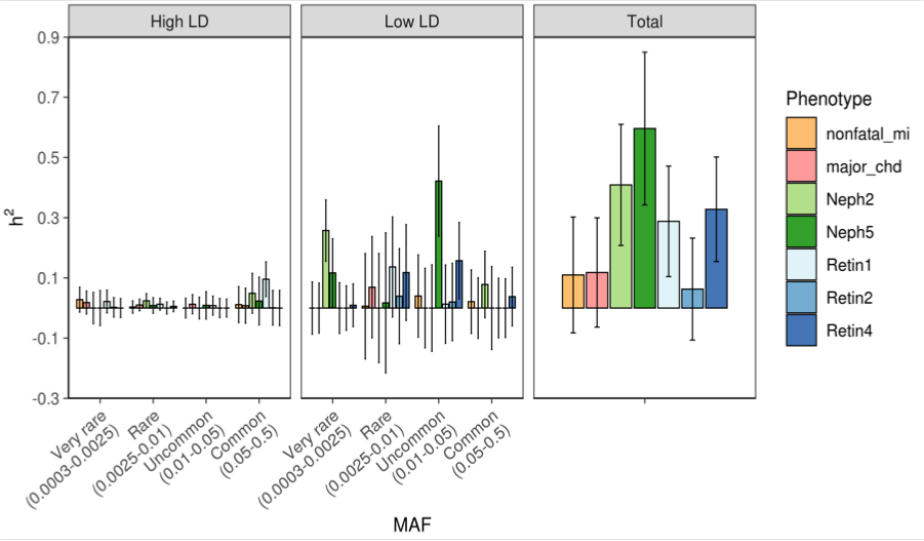

Figure 4.5: GREML-LDMS estimates using the imputed ACCORD data. GRM with eight bins (2 LD bins for each of the 4 MAF bins).

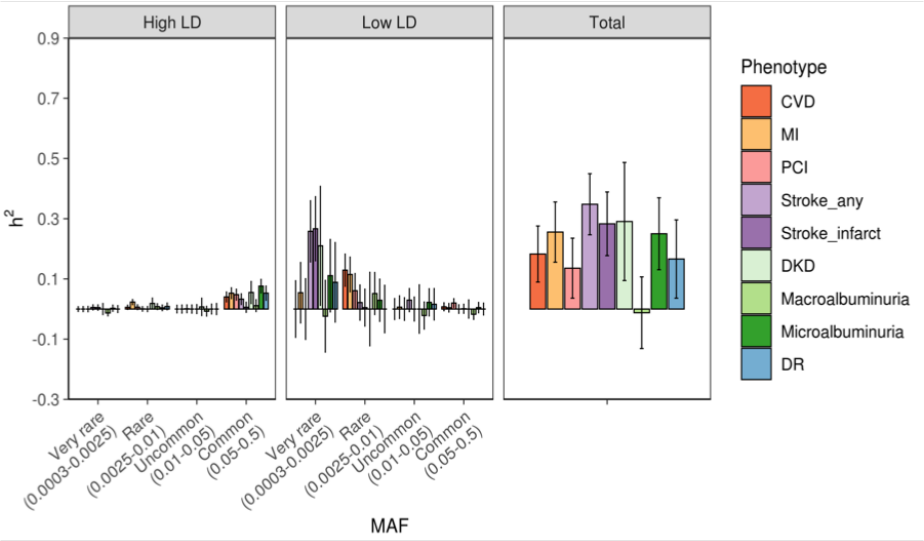

Figure 4.6: GREML-LDMS estimates using the imputed UKB data. GRM with eight bins (2 LD bins for each of the 4 MAF bins).

4.4.2 GWAS

Association results identified multiple significant peaks ($p < 5\times10^{-8}$) in the UKB-diabetes cohort. For macrovascular complications in the UKB-diabetes cohort (CVD, MI, and PCI), variants on chromosome 9p21 reached genome-wide significance. Association of the regions on chromosome 9p21 with type 2 diabetes and progression of CVD was seen previously (Helgeland et al., 2015). For DR in UKB-diabetes cohort, 22 variants on 6p21 reached genome-wide significance ($p < 5\times10^{-8}$) with rs9273367 ($p = 1.23\times10^{-9}$, $OR = 1.18$). These variants were in or near HLA regions, whose previous associations with type 1 diabetes have been well documented (Noble and Erlich, 2012). For DKD, 17 variants had $p < 5\times10^{-8}$. Eleven of these SNPs were on chromosome 3q26.31, and six were in UMOD and PDILT (lead SNP rs77924615 with $p = 7.82\times10^{-9}$, $OR = 0.75$) on chromosome 16p12.3. UMOD was previously associated with eGFR in the meta-analysis combining type 1 and type 2 diabetes patients of European and Asian ancestry (Van Zuydam et al., 2018). Although some variants were below the genome-wide significance threshold in the ACCORD cohort, they were not as prominent as in the UKB-diabetes cohort. Figures B.4 and B.5 in Appendix B.3 show Manhattan and quantile-quantile (QQ) plots for GWAS.

4.4.3 Heritability enrichment by functional annotations

We applied S-LDSC to identify disease-relevant tissues and cell types. Results for the selected ACCORD phenotypes are illustrated in Figure 4.7 (for the results on more phenotypes, see Figures B.7 and B.8 in Appendix B.3). Renal failure or ESRD phenotype (Neph3) exhibit skin-specific (sun-exposed skin $p = 4.82 \times 10^{-4}$; non-sun-exposed skin $p = 4.29 \times 10^{-3}$) and brain-specific enrichments (brain cerebellar hemisphere $p = 1.99 \times 10^{-3}$). The enrichment mentioned earlier captures dermatologic manifestations of ESRD (47). Macrovascular complications (primary and major CHD) show enrichments in EBV transformed lymphocytes ($p = 1.38 \times 10^{-3}$ and $p = 2.25 \times 10^{-3}$, respectively). This finding of macrovascular

complications reflects their mechanism involving inflammatory cells (e.g., monotypes and T lymphocytes) (48). Despite the larger sample size, no tissues are enriched for heritability of diabetic complications from the UKB (see Figures B.9 and B.10 in Appendix B.3). We note that the lack of significant enrichment findings from S-LDSC methods may stem from the difference in cohort characteristics (healthier participants from UKB and non-adjudicated outcomes from biobank studies).

We also conducted the S-LDSC analysis partitioning heritability into 53 (overlapping) categories that are not specific to any cell type. The annotations were derived from the 'full baseline model' in Finucane et al. (2015). Figure 4.8 illustrates enrichment estimates for selected annotations and traits from the ACCORD data. None of the categories for any phenotype passed the Bonferroni significance threshold (threshold $p = 0.05/53 = 9.43 \times 10^{-4}$). Given the small sample size we had for phenotypes (e.g., ranging from 2,912 samples for Neph5 to 4,318 samples for primary), we were aware of the power limitations. Some categories are still noteworthy, however. Promoter region showed enrichment in the retinopathy phenotype (Retin1;$p = 2.82 \times 10^{-2}$) and H3K27ac showed enrichment in the composite nephropathy phenotype (Neph4; $p = 4.64 \times 10^{-2}$).

Figure 4.9 reports enrichment estimates from the UKB data. Only the coding region passed Bonferroni-corrected significant enrichment (threshold $p = 0.05/53 = 9.43 \times 10^{-4}$) in the diabetic kidney disease phenotype (DKD; $p = 6.55 \times 10^{-4}$). Though only nominally significant, H3K9ac is enriched in microalbuminuria phenotype ($p = 0.04$). H3K9ac enrichment agrees with the findings from Salem et al. (2019) that the top signal (TAMM41) for microalbuminuria is close to the histone marks—H3K27ac, H3K9ac, H3k4me1.

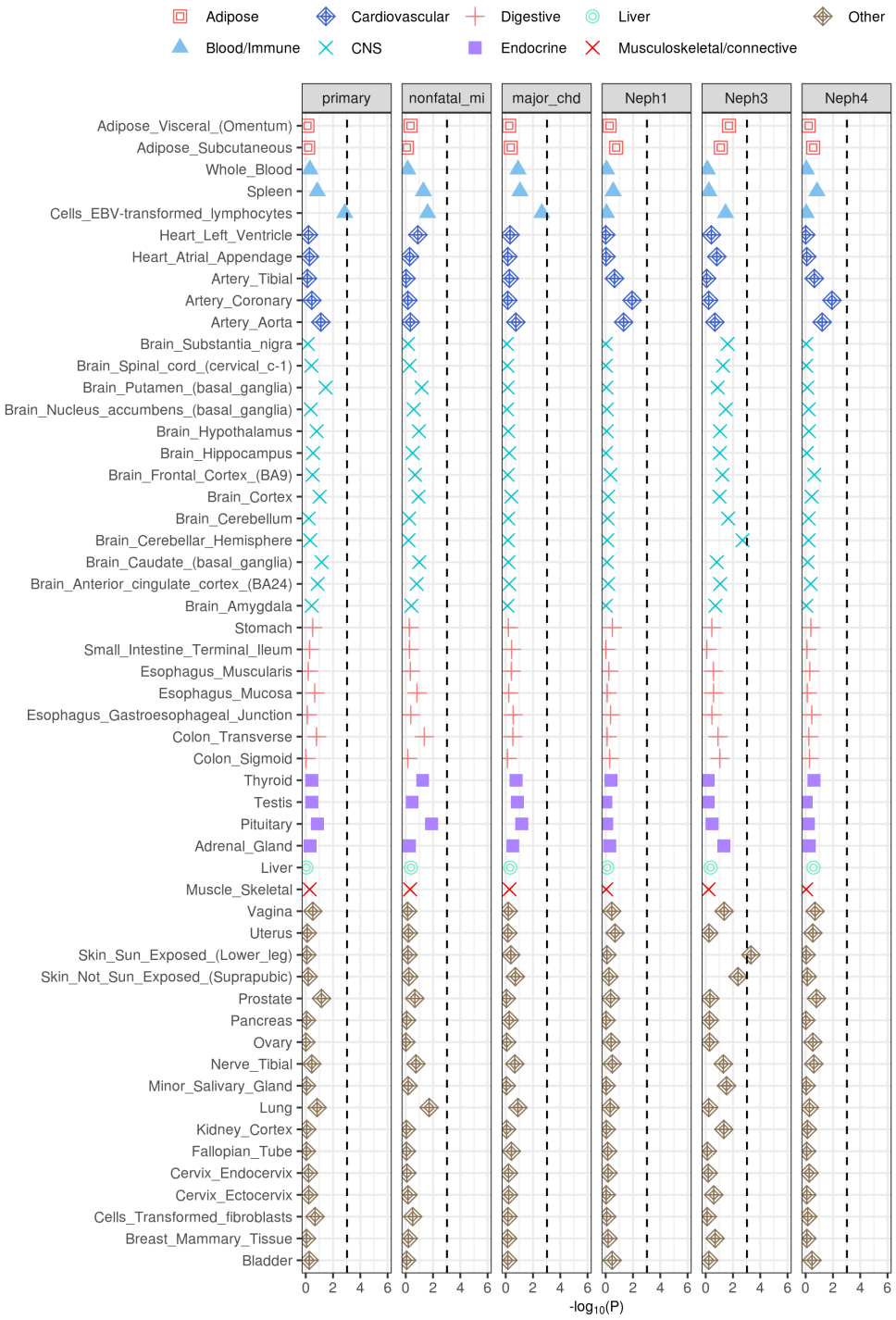

Figure 4.7: Enrichment of the selected ACCORD phenotypes in tissue-specific gene expression annotations used in Finucane et al. (2018). The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).

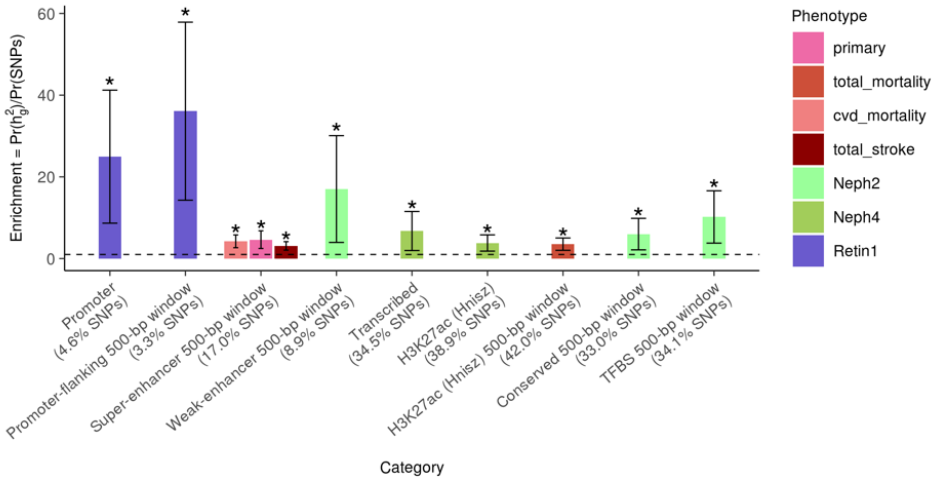

Figure 4.8: Enrichment estimates for selected annotations and traits using the ACCORD imputed data. The dashed line represents no enrichment (enrichment=1). One asterisk indicates nominal significance at $p < 0.05$. TFBS, Transcription factor binding site.

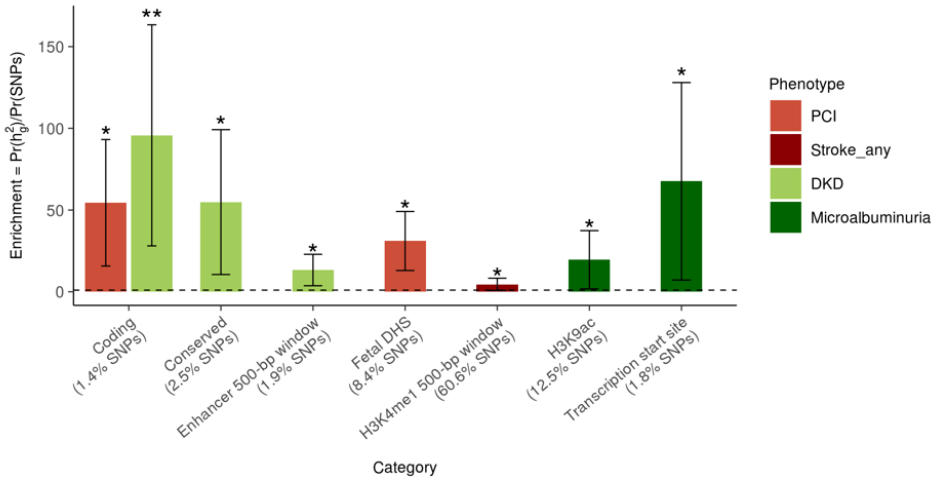

Figure 4.9: Enrichment estimates for selected annotations and traits using the UKB imputed data. The dashed line represents no enrichment (enrichment=1). One asterisk indicates nominal significance at $p < 0.05$, while two asterisks denote significance at $p < 0.001$. DHS, DNase I hypersensitivity sites.

4.5 Discussion

We have used two independent studies with imputed data to estimate the heritability for diabetes complications. Although a meta-analysis from the two studies would have increased the sample size, we conducted two separate analyses to reduce the risk of phenotypic heterogeneity. Our analyses show some discordance in findings between the two data sets. Heritability estimates obtained from the GREML-LDMS-I using imputed datasets tend to be larger in the ACCORD study than the UKB study despite a larger sample size in the UKB-Diabetes cohort. Additionally, no tissue enrichment is observed from the UKB-Diabetes cohort. Inherent differences in the two studies may provide a basis for the discordant findings. First, the ACCORD is a clinical trial that offers adjudicated outcomes in a well-controlled clinical trial setting. In contrast, UKB studies are conducted within a real-life cohort and based on electronic medical records, which are known for high noise and potential for bias. Second, the ACCORD cohort consisted of adults at increased risk for CVD with a longer duration of diabetes and higher glycated hemoglobin level (Table 4.1). On the other hand, the UKB participants were younger and relatively healthy (Table 4.2).

Notwithstanding these limitations, this heritability analysis still represents the first systematic investigation of SNP heritability for diabetes complications. It adds to the existing heritability analyses primarily based on family or small cohort studies.

CHAPTER 5

Conclusion

In this dissertation, we have developed and applied variance component model-based methodology for understanding the genetic architecture of complex traits and diseases. Chapter 2 introduced a method that prioritizes SNP-sets in a joint multivariate variance component model. Each SNP-set corresponded to a variance component, and model selection was achieved by incorporating either convex or non-convex penalties. We extended this approach to incorporate SNP-set-treatment or -environment interactions in Chapter 3. The algorithms we devised were based on the majorization-minimization (MM) principle. Through simulation studies, we demonstrated the competitiveness of our methods in model selection performance, compared to the commonly used marginal testing and group penalization methods. We also applied our methods to a real whole exome sequencing study and a real pharmacogenomics study. As we saw that some top ranked genes by our methods were detected as insignificant by the marginal testing, our methods can provide alternative insights for biologists to prioritize follow-up studies and develop polygenic risk score models.

For the proposed VCSEL methods, two future directions come to mind. One direction is to develop post-selection inference procedures. In Chapters 2 and 3, we only focused on the ranking of genes and reported the overall selection performance by the area under the Precision-Recall curve. Inference tools would allow us to assess the strength of the model after penalization. Another direction is to scale the methods to accommodate biobank-scale data. The computational bottleneck of the current VCSEL methods is arises from an inversion of the covariance matrix at each iteration. Computationally efficient approaches

that can handle large-scale data would widen the applicability.

Chapter 4 surveyed the SNP heritability of diabetes macrovascular complications and microvascular complications. Heritability estimates of diabetic retinopathy ranged from 0.06 to 0.33, depending on the definition of phenotype. The heritability of diabetic kidney disease was estimated to be 0.29. Microalbuminuria estimates ranged from 0.24 to 0.60 while macroalbuminuria estimates were up to 0.41. Heritability analyses using imputed data revealed a substantial contribution of low-frequency/rare variants to the predisposition for complications. This analysis represents the first systematic investigation of SNP heritability for diabetes complications. It adds to the existing heritability analyses primarily based on family or small cohort studies.

Appendix A

A.1 Simulation studies for univariate trait

In Section 2.4, we presented simulation results for the multivariate trait model. Here we present more details on the simulation results for univariate trait model. For univariate response, we compare with the group lasso and SKAT, using R packages `gglasso` (Yang and Zou, 2014) and `SKAT` (Lee et al., 2017a), respectively. In implementing VCSEL algorithm, we employ the lasso (Chen et al., 2001; Tibshirani, 1996), adaptive lasso (Zou, 2006) and MCP (Zhang et al., 2010) penalties, which are denoted as VCSEL-lasso, VCSEL-adlasso, and VCSEL-MCP, respectively. For the adaptive lasso penalty, we take the two-stage approach: the initial estimation of $\widetilde{\boldsymbol{\sigma}}$ is made using no penalty, which in turn generates weights $w_i = |\widetilde{\sigma}_i|^{-1}, i = 1, \ldots, m$ that are held constant across all values of tuning parameter λ.

We simulate univariate phenotype vector $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ from

$$\boldsymbol{y} = \boldsymbol{L}_{\boldsymbol{\Omega}}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n),$$

where $\boldsymbol{L}_{\boldsymbol{\Omega}}$ is the lower triangular Cholesky factor of $\boldsymbol{\Omega} = \sigma_1^2\boldsymbol{V}_1 + \cdots + \sigma_m^2\boldsymbol{V}_m + \frac{\sigma_0^2}{\sqrt{n}}\boldsymbol{I}_n$. For simplicity, we adopt the linear kernel function for $\boldsymbol{V}_i$ and divide the kernel by its Frobenius norm. In other words, $\boldsymbol{V}_i = \frac{1}{||\boldsymbol{G}_i\boldsymbol{G}_i^T||_F}\boldsymbol{G}_i\boldsymbol{G}_i^T, i = 1, \cdots, m$, where $||\cdot||_F$ is the Frobenius norm.

Throughout simulations, we set the residual variance $\sigma_0 = 1.0$. Furthermore, we fix sample size n to be 500 and the number of positive variance components, excluding σ_0, to be 5. Those positive variance components are spread evenly across all m sets to create a scenario of low linkage disequilibrium between causal variants.

We exploit two different haplotype pools: one from SKAT.haplotypes in the SKAT R package and the other simulated from the cosi2 simulator (Shlyakhter et al., 2014). The former contains predominantly rare variants while the latter consists of common and rare variants in similar proportions.

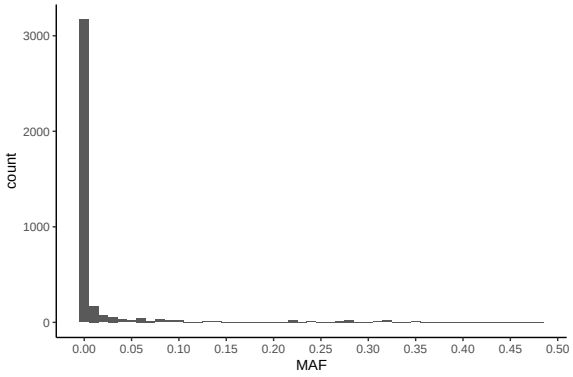

Figure A.1: Histogram of MAFs for the 3,845 SNPs in haplotype matrix from the SKAT R package.

Haplotype dataset from the SKAT R package

Generated by the calibration coalescent model (COSI) mimicking the linkage disequilibrium structure of European ancestry, the haplotype matrix in the SKAT R package contains 10,000 haplotypes over 200kb region. Most SNPs in the matrix are rare, with 82% of variants having MAF ≤ 0.005 and 8.7% having MAF > 0.05. Figure A.1 illustrates the distribution of MAFs for all SNPs in the pool.

We define one set to be 5kb, 2kb, or 1kb long, which translates to 40, 100, or 200 groups, respectively. Suppose this set represents a gene, then each gene of length 5kb, 2kb, and 1kb contains approximately 42, 17, and 8 non-monomorphic SNPs on average, respectively (Table A.1).

	Min.	Mean	Max.
1kb/gene	1	8	24
2kb/gene	6	17	31
5kb/gene	24	42	69

Table A.1: The minimum, mean, and maximum numbers of SNPs that are not monoallelically expressed within a gene for SKAT haplotype data when a gene is defined to be 1kb, 2kb, or 5kb long.

Table A.2: The auPRCs of VCSEL-lasso, VCSEL-adlasso, VCSEL-MCP, group-lasso and SKAT. We set the number of replicates as 20 and number of tuning parameters as 100. In parentheses are standard deviation $/\sqrt{\text{no. replicates}}$. For VCSEL-MCP, $\gamma = 2.69$ is used.

no. genes	VCSEL-lasso	VCSEL-adlasso	VCSEL-MCP	group-lasso	SKAT
40 (5kb/gene)	0.76 (0.02)	0.82 (0.02)	0.76 (0.02)	0.27 (0.02)	0.40 (0.03)
100 (2kb/gene)	0.69 (0.04)	0.72 (0.04)	0.68 (0.04)	0.21 (0.02)	0.24 (0.02)
200 (1kb/gene)	0.56 (0.04)	0.65 (0.03)	0.56 (0.04)	0.12 (0.01)	0.17 (0.02)

We set the effect strength of non-zero variance component to be 2.236. In other words,

$$\sigma_i = \begin{cases} 2.236 & i = 1, 11, 20, 30, 40 \ (m = 40) \\ & i = 1, 26, 50, 75, 100 \ (m = 100) \\ & i = 1, 51, 100, 150, 200 \ (m = 200) \\ 1.0 & i = 0 \\ 0.0 & \text{else.} \end{cases}$$

The simulation results are summarized in Figure A.2 with boxplots and Table A.2 with the average auPRCs and standard errors. Overall, we observe VCSEL-adlasso > VCSEL-MCP > VCSEL-lasso $\gg$ SKAT > group-lasso in terms of selection performance.

Haplotype pool generated from the cosi2 simulator

The haplotype matrix in this section is generated by cosi2 simulator (Shlyakhter et al., 2014). It contains 1,200 haplotypes over 10Mb region. Proportions of common and rare variants are more or less equal: 38.5% of those have MAF > 0.05, 31.9% have MAF ≤ 0.005, and

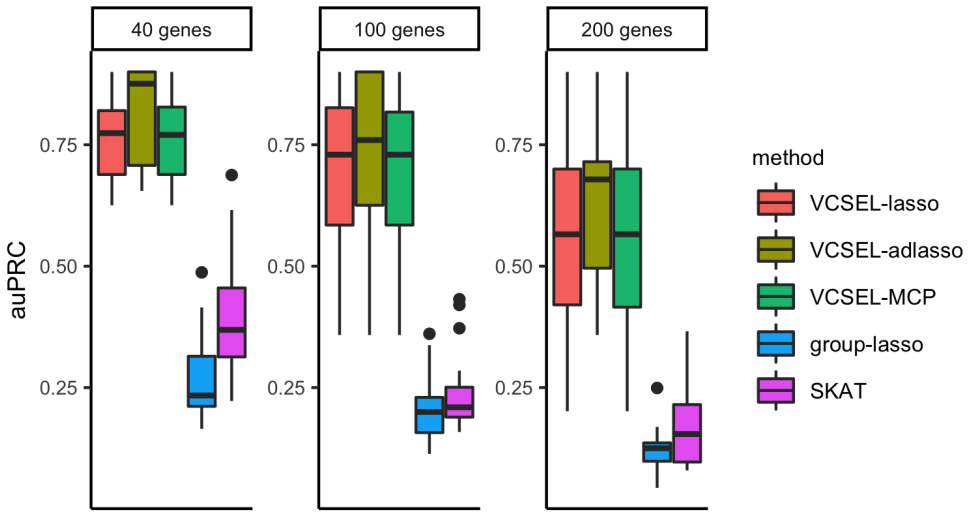

Figure A.2: The auPRCs of VCSEL-lasso, VCSEL-adlasso, VCSEL-MCP($\gamma = 2.69$), group-lasso, and SKAT under different number of genes for models with 5 non-zero variance components, using haplotype data from the SKAT R package. Three different numbers of groups are compared: $m = 40, 100, 200$.

	Min.	Mean	Max.
10kb/gene	435	540	673
20kb/gene	932	1080	1230
50kb/gene	2414	2704	3041

Table A.3: The minimum, mean, and maximum numbers of SNPs that are not monoallelically expressed within a gene for cosi2 haplotype data when a gene is defined to be 10kb, 20kb or 50kb long.

the rest (29.6%) have MAF between 0.005 and 0.05. Figure A.3 illustrates the distribution of MAFs for all SNPs.

As before, we vary the number of genes and gene sizes. We consider three cases where there are 50, 100, and 200 genes, in which each gene is 20kb, 10kb, or 50kb long, respectively. On average, genes of size 20kb, 10kb, and 50kb long contain 1080, 540, and 2704 non-monomorphic SNPs, respectively (Table A.3). We set the effect strength of non-zero variance

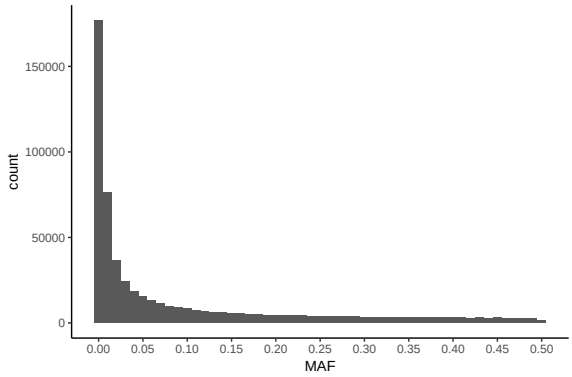

Figure A.3: Histogram of MAFs for 553,862 SNPs in haplotype matrix generated from the cosi2 simulator.

component to be 2.236. In other words,

$$\sigma_i = \begin{cases} 2.236 & \text{i=1, 13, 26, 38, 50 (m=50)} \\ & \text{i=1, 26, 50, 75, 100 (m=100)} \\ & \text{i=1, 51, 100, 150, 200 (m=200)} \\ 1.0 & \text{i=0} \\ 0.0 & \text{else.} \end{cases}$$

The simulation results are summarized in Figure A.4 with boxplots and Table A.4 with the average auPRCs and standard errors. We again observe a similar pattern as in Section A.1 in terms of selection performance, but the difference between our method and competing methods is not as drastic. We attribute this discrepancy to different proportions of common/rare variants and different gene sizes. Advantages of our method appear to be amplified when there are more rare variants present and gene sizes are smaller.

no. genes	VCSEL-lasso	VCSEL-adlasso	VCSEL-MCP	group-lasso	SKAT
50 (20kb/gene)	0.83 (0.02)	0.82 (0.02)	0.83 (0.02)	0.81 (0.02)	0.70 (0.02)
100 (10kb/gene)	0.72 (0.03)	0.79 (0.02)	0.72 (0.03)	0.62 (0.03)	0.57 (0.04)
200 (50kb/gene)	0.67 (0.05)	0.67 (0.04)	0.67 (0.05)	0.59 (0.05)	0.60 (0.05)

Table A.4: The auPRCs of VCSEL-lasso, VCSEL-MCP, and Multi-SKAT across varying size and number of genes, using haplotype data from the cosi2 simulator. We set n = 500, number of replicates = 20, number of tuning parameters = 100. In parentheses are standard deviation $/\sqrt{\text{no. replicates}}$.

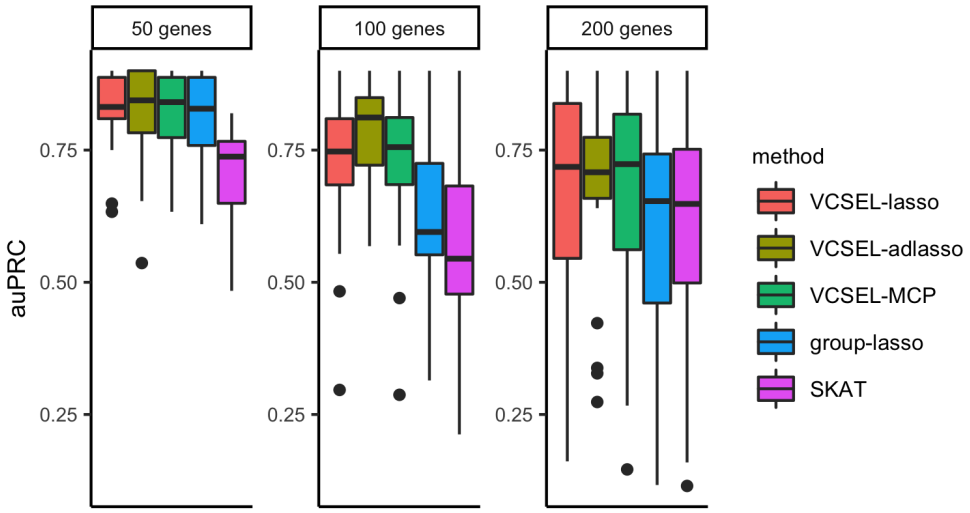

Figure A.4: The auPRCs of VCSEL-lasso, VCSEL-adlasso, VCSEL-MCP, group-lasso, and SKAT under different number of genes for models with 5 non-zero variance components, using haplotype data from the cosi2 simulator. Three different numbers of groups are compared: $m = 50, 100, 200$.

A.2 Simulation studies for univariate response model - extra results

In this section, we further expand the simulations for the univariate response model to address two questions. First, besides the genetic association study setting, how does VCSEL perform in random effects ANOVA (factorial analysis of variance) model with many factors? In the ANOVA model, the effects of each factor is modeled as random effects and correspond to one variance component. Second, how does VCSEL perform in the group selection in the fixed effects models, for both genetic association studies and factorial ANOVA.

In this section we only compare selection performance of the VCSEL method to group lasso (Yuan and Lin, 2006), a group selection method, given data generated from different univariate response models. We simulate $n \times 1$ response vector $\boldsymbol{y}$ both from a random effects model,

$$\boldsymbol{y} \sim N(\boldsymbol{0}_n, \boldsymbol{\Omega}), \tag{A.1}$$

where $\boldsymbol{\Omega} = \sum_{i=1}^{m} \sigma_i^2 \boldsymbol{V}_i + \sigma_0^2 \boldsymbol{I}_n$ and $\boldsymbol{V}_i = \frac{1}{||\boldsymbol{X}_i \boldsymbol{X}_i^T||_F} \boldsymbol{X}_i \boldsymbol{X}_i^T$, and a fixed effects model,

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n). \tag{A.2}$$

For random effects model (A.1), we set true variance components to be $\sigma_0^2 = 1.0$, $\sigma_i^2 = 9.0$ if $i = 1, 2, 3$, and 0.0 if else. For fixed effects model (A.2), true parameter $\boldsymbol{\beta}$ are generated from normal distribution, $\boldsymbol{\beta}_{[(qi+1):(qi+q)]} \sim N(0, 9\boldsymbol{I}), i = 0, 1, 2$, where q denotes group size. The rest of values in $\boldsymbol{\beta}$ are set to be 0. Two simulation designs are motivated by genetics and analysis of variance (ANOVA) scenario and are distinguished by a $n \times mq$ matrix

$$\boldsymbol{X} = \left[\boldsymbol{X}_1 \middle| \boldsymbol{X}_2 \middle| \cdots \middle| \boldsymbol{X}_m \right],$$

where $\boldsymbol{X}_i \in \mathbb{R}^{n \times q}, i = 1, \ldots, m$. In both settings, we fix $n = 100$ and experiment with 6

different groups ($m = 10, 20, 30, 50, 70, 100$) and 4 different group sizes ($q = 40, 50, 70, 100$). We generate 50 replications and obtain 50 area under Precision-Recall curve (auPRC) values at each combination of group size, number of groups, and method. We remind readers that auPRC value ranges from 0 to 1, and higher auPRC value signals both high precision (low false positive rate) and high recall (low false negative rate).

In genetics scenario, $\boldsymbol{X}$ is a matrix of genotype data with allele counts, which we obtain from `SNP_data29a.bin` in option 24 example of Mendel software (Lange et al., 2013). Here each $\boldsymbol{X}_i$ represents i-th SNP-set which includes q variants. The results from the data generated by a random effects model are summarized in Table A.5 and Figure A.5. Similarly, the results from the data generated by a fixed effects are model are displayed in Table A.6 and Figure A.6.

On the other hand, in ANOVA scenario, $\boldsymbol{X}_i$ represents i-th factor with q levels, and its k-th row indicates which level k-th subject belongs to in factor i. Interactions are not considered for simplicity. The results from the data generated by a random effects model are summarized in Table A.7 and Figure A.7. Similarly, the results from the data generated by a fixed effects are model are displayed in Table A.8 and Figure A.8.

In both scenarios, we observe that auPRC values decrease with increasing number of groups and group size, which translates to increasing number of parameters. The overall performance trend in all scenarios is VCSEL-adlasso > VCSEL-lasso > group-lasso. This trend is reasonable for data generated from random effects model since group-lasso assumes fixed effects model. On that note, it is interesting to point out that VCSEL methods remains competitive even under the scenario where data are generated from fixed effects model.

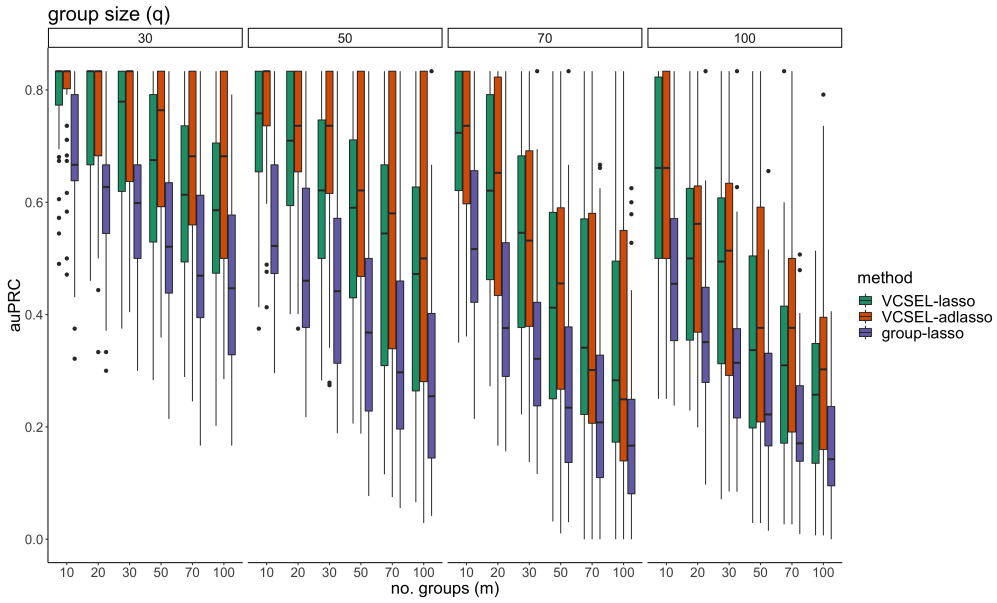

Figure A.5: The auPRCs of VCSEL-lasso, VCSEL-adlasso, and group-lasso under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given the data generated from the random effects model (A.1) in genetics setting.

q	Method	m = number of groups					
		10	20	30	50	70	100
30	VCSEL-lasso	0.78 (0.01)	0.74 (0.02)	0.71 (0.02)	0.65 (0.02)	0.61 (0.02)	0.57 (0.02)
	VCSEL-adlasso	0.79 (0.01)	0.75 (0.02)	0.74 (0.02)	0.71 (0.02)	0.66 (0.02)	0.65 (0.02)
	group-lasso	0.67 (0.02)	0.61 (0.02)	0.58 (0.02)	0.52 (0.02)	0.49 (0.02)	0.45 (0.02)
50	VCSEL-lasso	0.72 (0.02)	0.69 (0.02)	0.62 (0.02)	0.57 (0.03)	0.51 (0.03)	0.45 (0.03)
	VCSEL-adlasso	0.77 (0.01)	0.72 (0.02)	0.70 (0.02)	0.61 (0.03)	0.55 (0.04)	0.52 (0.04)
	group-lasso	0.57 (0.02)	0.49 (0.02)	0.46 (0.02)	0.38 (0.03)	0.34 (0.03)	0.29 (0.03)
70	VCSEL-lasso	0.70 (0.02)	0.61 (0.03)	0.54 (0.03)	0.43 (0.03)	0.39 (0.03)	0.35 (0.03)
	VCSEL-adlasso	0.70 (0.02)	0.61 (0.03)	0.54 (0.03)	0.45 (0.03)	0.38 (0.03)	0.34 (0.03)
	group-lasso	0.53 (0.02)	0.42 (0.02)	0.36 (0.03)	0.29 (0.03)	0.24 (0.02)	0.19 (0.02)
100	VCSEL-lasso	0.63 (0.02)	0.51 (0.03)	0.47 (0.03)	0.38 (0.03)	0.31 (0.02)	0.26 (0.02)
	VCSEL-adlasso	0.64 (0.02)	0.54 (0.03)	0.50 (0.03)	0.41 (0.03)	0.36 (0.03)	0.30 (0.03)
	group-lasso	0.48 (0.02)	0.37 (0.02)	0.32 (0.02)	0.25 (0.02)	0.19 (0.02)	0.16 (0.01)

Table A.5: The mean auPRCs across 50 replications under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given data generated from random effects model (A.1) in genetics setting. In parentheses are standard deviation $/\sqrt{\text{no. replicates}}$.

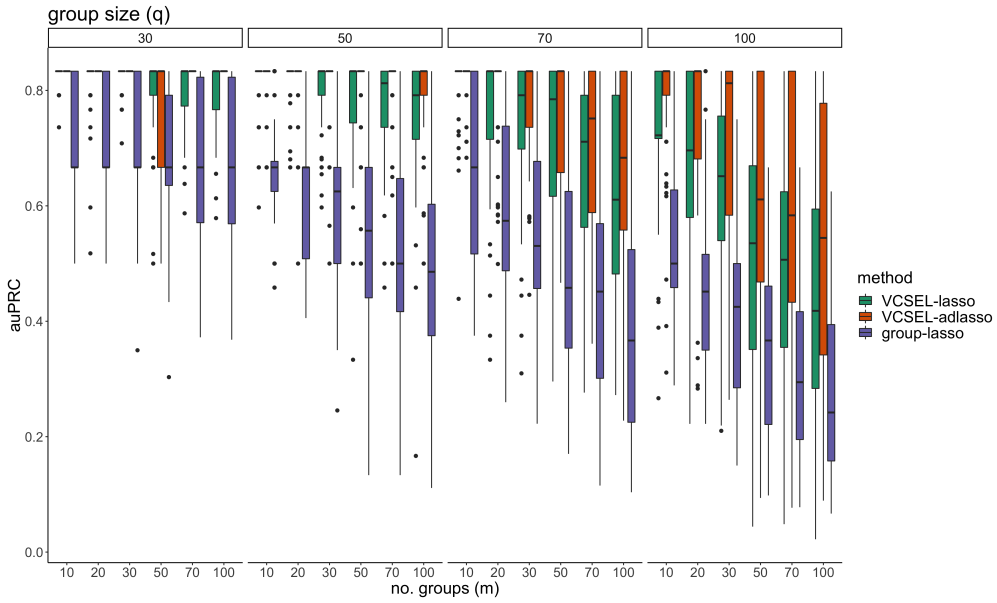

Figure A.6: The auPRCs of VCSEL-lasso, VCSEL-adlasso, and group-lasso under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given data generated from fixed effects model (A.2) in genetics setting.

q	Method	m = no. of groups					
		10	20	30	50	70	100
30	VCSEL-lasso	0.83 (0.00)	0.81 (0.01)	0.82 (0.00)	0.79 (0.01)	0.80 (0.01)	0.80 (0.01)
	VCSEL-adlasso	0.83 (0.00)	0.83 (0.00)	0.83 (0.00)	0.77 (0.02)	0.83 (0.00)	0.83 (0.00)
	group-lasso	0.72 (0.01)	0.72 (0.01)	0.70 (0.02)	0.68 (0.02)	0.67 (0.02)	0.66 (0.02)
50	VCSEL-lasso	0.81 (0.01)	0.81 (0.01)	0.80 (0.01)	0.78 (0.01)	0.76 (0.01)	0.75 (0.02)
	VCSEL-adlasso	0.82 (0.01)	0.82 (0.01)	0.81 (0.01)	0.80 (0.01)	0.80 (0.01)	0.79 (0.01)
	group-lasso	0.66 (0.02)	0.62 (0.02)	0.60 (0.02)	0.56 (0.02)	0.52 (0.02)	0.49 (0.02)
70	VCSEL-lasso	0.81 (0.01)	0.76 (0.02)	0.74 (0.02)	0.70 (0.02)	0.66 (0.02)	0.62 (0.03)
	VCSEL-adlasso	0.82 (0.00)	0.79 (0.01)	0.78 (0.01)	0.75 (0.02)	0.71 (0.02)	0.66 (0.02)
	group-lasso	0.66 (0.02)	0.59 (0.02)	0.57 (0.02)	0.49 (0.03)	0.46 (0.03)	0.40 (0.03)
100	VCSEL-lasso	0.72 (0.02)	0.67 (0.03)	0.62 (0.03)	0.52 (0.03)	0.50 (0.03)	0.45 (0.03)
	VCSEL-adlasso	0.78 (0.02)	0.74 (0.02)	0.70 (0.02)	0.62 (0.03)	0.60 (0.03)	0.53 (0.03)
	group-lasso	0.53 (0.02)	0.47 (0.02)	0.42 (0.02)	0.36 (0.02)	0.32 (0.02)	0.27 (0.02)

Table A.6: The mean auPRCs across 50 replications under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given data generated from fixed effects model (A.2) in genetics setting. In parentheses are standard deviation $/\sqrt{\text{no. replicates}}$.

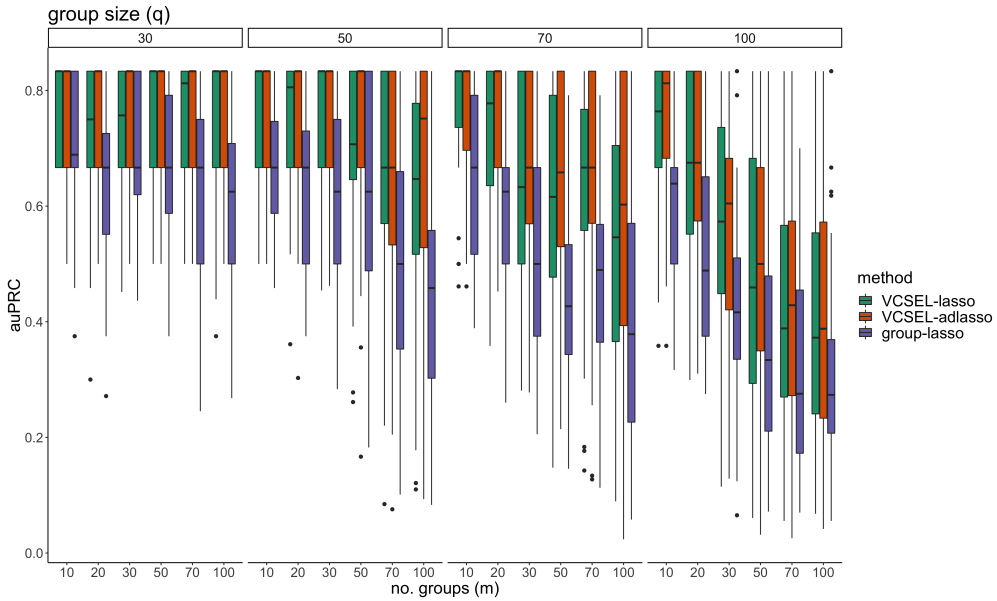

Figure A.7: The auPRCs of VCSEL-lasso, VCSEL-adlasso, and group-lasso under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given data generated from random effects model (A.1) in ANOVA setting.

q	Method	m = no. of groups					
		10	20	30	50	70	100
30	VCSEL-lasso	0.71 (0.01)	0.73 (0.02)	0.70 (0.02)	0.71 (0.01)	0.75 (0.02)	0.74 (0.01)
	VCSEL-adlasso	0.71 (0.01)	0.72 (0.01)	0.64 (0.02)	0.69 (0.02)	0.75 (0.01)	0.72 (0.02)
	group-lasso	0.69 (0.02)	0.71 (0.02)	0.65 (0.02)	0.62 (0.02)	0.70 (0.02)	0.68 (0.01)
50	VCSEL-lasso	0.76 (0.01)	0.78 (0.01)	0.73 (0.02)	0.74 (0.02)	0.70 (0.02)	0.67 (0.02)
	VCSEL-adlasso	0.75 (0.01)	0.77 (0.01)	0.75 (0.02)	0.73 (0.02)	0.71 (0.02)	0.68 (0.02)
	group-lasso	0.67 (0.02)	0.67 (0.02)	0.61 (0.02)	0.63 (0.02)	0.54 (0.02)	0.50 (0.03)
70	VCSEL-lasso	0.77 (0.01)	0.67 (0.02)	0.71 (0.02)	0.66 (0.02)	0.64 (0.02)	0.59 (0.03)
	VCSEL-adlasso	0.77 (0.01)	0.69 (0.02)	0.73 (0.02)	0.65 (0.03)	0.67 (0.02)	0.57 (0.03)
	group-lasso	0.67 (0.02)	0.51 (0.02)	0.51 (0.02)	0.50 (0.03)	0.47 (0.02)	0.44 (0.03)
100	VCSEL-lasso	0.72 (0.02)	0.70 (0.02)	0.63 (0.02)	0.55 (0.03)	0.44 (0.03)	0.39 (0.03)
	VCSEL-adlasso	0.71 (0.02)	0.71 (0.02)	0.64 (0.02)	0.56 (0.03)	0.45 (0.03)	0.42 (0.04)
	group-lasso	0.60 (0.02)	0.58 (0.03)	0.50 (0.02)	0.41 (0.03)	0.32 (0.03)	0.28 (0.02)

Table A.8: The mean auPRCs across 50 replications under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given data generated from fixed effects model (A.2) in ANOVA setting. In parentheses are standard deviation $/\sqrt{\text{no. replicates}}$.

q	Method	m = no. of groups					
		10	20	30	50	70	100
30	VCSEL-lasso	0.77 (0.01)	0.72 (0.02)	0.74 (0.01)	0.76 (0.02)	0.75 (0.01)	0.73 (0.02)
	VCSEL-adlasso	0.77 (0.01)	0.74 (0.02)	0.75 (0.02)	0.75 (0.02)	0.74 (0.02)	0.77 (0.01)
	group-lasso	0.72 (0.02)	0.65 (0.02)	0.68 (0.02)	0.67 (0.02)	0.63 (0.02)	0.61 (0.02)
50	VCSEL-lasso	0.76 (0.01)	0.74 (0.02)	0.75 (0.02)	0.69 (0.02)	0.65 (0.03)	0.61 (0.03)
	VCSEL-adlasso	0.77 (0.01)	0.73 (0.02)	0.76 (0.02)	0.74 (0.02)	0.66 (0.03)	0.67 (0.03)
	group-lasso	0.67 (0.02)	0.63 (0.02)	0.62 (0.02)	0.59 (0.03)	0.49 (0.03)	0.44 (0.03)
70	VCSEL-lasso	0.78 (0.01)	0.72 (0.02)	0.62 (0.02)	0.6 (0.03)	0.62 (0.03)	0.52 (0.03)
	VCSEL-adlasso	0.77 (0.01)	0.75 (0.02)	0.66 (0.03)	0.64 (0.03)	0.65 (0.03)	0.58 (0.04)
	group-lasso	0.66 (0.02)	0.59 (0.02)	0.53 (0.03)	0.44 (0.02)	0.46 (0.02)	0.41 (0.03)
100	VCSEL-lasso	0.74 (0.02)	0.65 (0.02)	0.57 (0.03)	0.49 (0.03)	0.41 (0.03)	0.41 (0.03)
	VCSEL-adlasso	0.75 (0.02)	0.66 (0.02)	0.57 (0.03)	0.50 (0.03)	0.43 (0.03)	0.44 (0.03)
	group-lasso	0.61 (0.02)	0.51 (0.02)	0.42 (0.02)	0.36 (0.03)	0.32 (0.02)	0.31 (0.02)

Table A.7: The mean auPRCs across 50 replications under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given data generated from random effects model (A.1) in ANOVA setting. In parentheses are standard deviation $/\sqrt{\text{no. replicates}}$.

A.3 Canonical correlations of SNP-sets in simulation studies

In this section, we conduct a canonical correlation analysis to quantify the linkage equilibrium (LD) in the high/low LD settings in Simulation Studies in Section 3.3 and A.1. Recall that we construct a genotype matrix $\boldsymbol{G}$ for 500 observations from the haplotype matrix in the SKAT R package and then partition $\boldsymbol{G}$ into submatrices or SNP-sets. For simplicity, we look at the case where the constructed genotype matrix $\boldsymbol{G}$ is broken into 40 submatrices $\boldsymbol{G}_i$, $i = 1, \ldots, 40$ by 5 kb region. In the low LD setting, we set causal SNP-sets to be distanced ($i = 1, 11, 20, 30, 40$) while causal SNP-sets are neighboring in the high LD setting ($i = 1, 2, 3, 4, 5$). To justify the claim that causal SNPs are in a high/low LD, we calculate canonical correlations of causal SNP-sets $\boldsymbol{G}_j$ and $\boldsymbol{G}_k$ where $j, k \in \{1, 2, 3, 4, 5\}$ in high LD and $j, k \in \{1, 11, 20, 30, 40\}$ in low LD, respectively. We choose canonical correlation since it measures the associations among two sets of multidimensional variables. For both high LD and low LD cases, we repeat the calculations over 10 replicates of varying $\boldsymbol{G}$ matrix and obtain the mean canonical correlation values, which are illustrated through heat maps in Figure A.9 and A.10. By visual comparison, we observe that heat maps of causal SNP-sets in low LD setting are in general lighter than those in high LD setting, which suggests higher

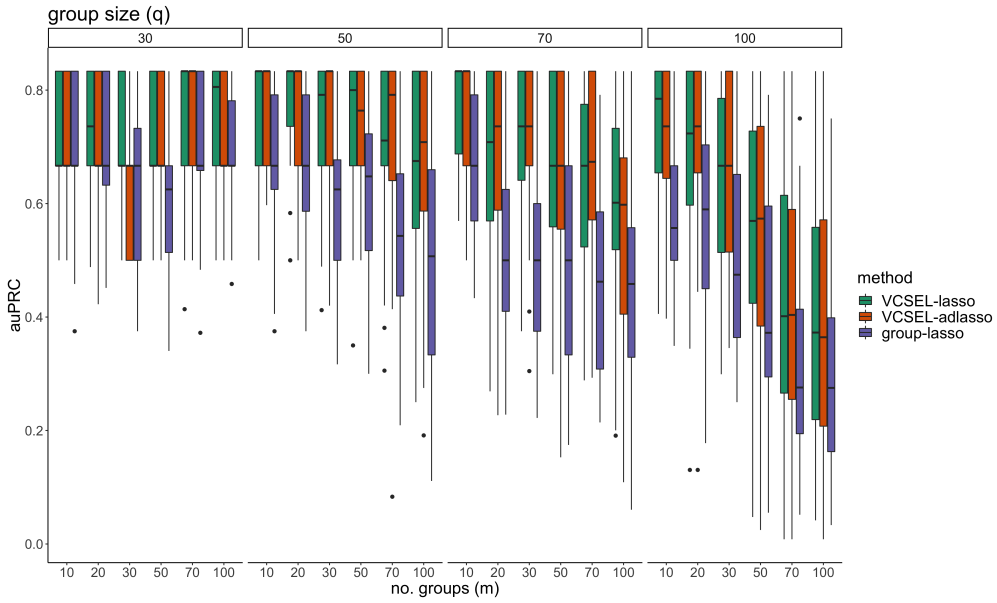

Figure A.8: The auPRCs of VCSEL-lasso, VCSEL-adlasso, and group-lasso under different number of groups ($m = 10, 20, 30, 50, 70, 100$) and group size ($q = 40, 50, 70, 100$), given data generated from fixed effects model (A.2) in ANOVA setting.

correlation in causal SNP-sets in high LD setting.

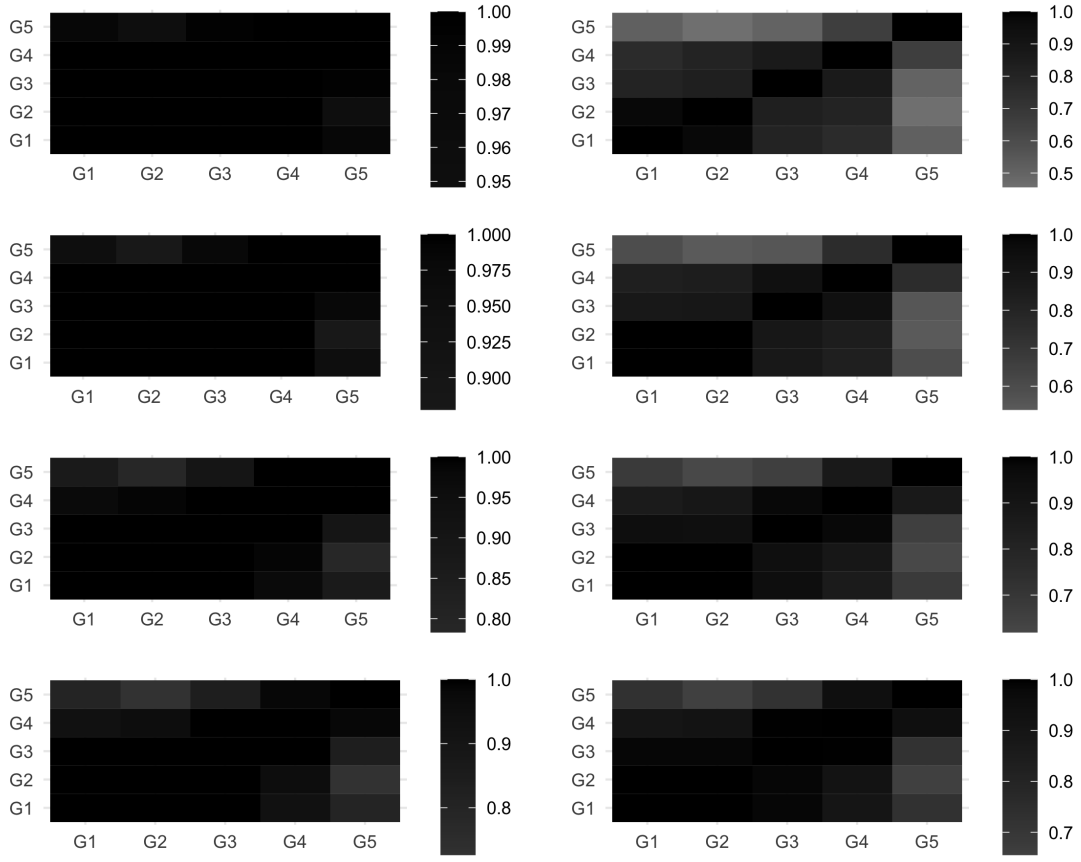

Figure A.9: Heat maps of the mean canonical correlations between SNP-sets $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$ where $i, j \in \{1, 2, 3, 4, 5\}$ across 10 replicates. From top left in counterclockwise order, each heat map corresponds to canonical correlation for the first to the eighth canonical variate pair. Darker color represents higher canonical correlation.

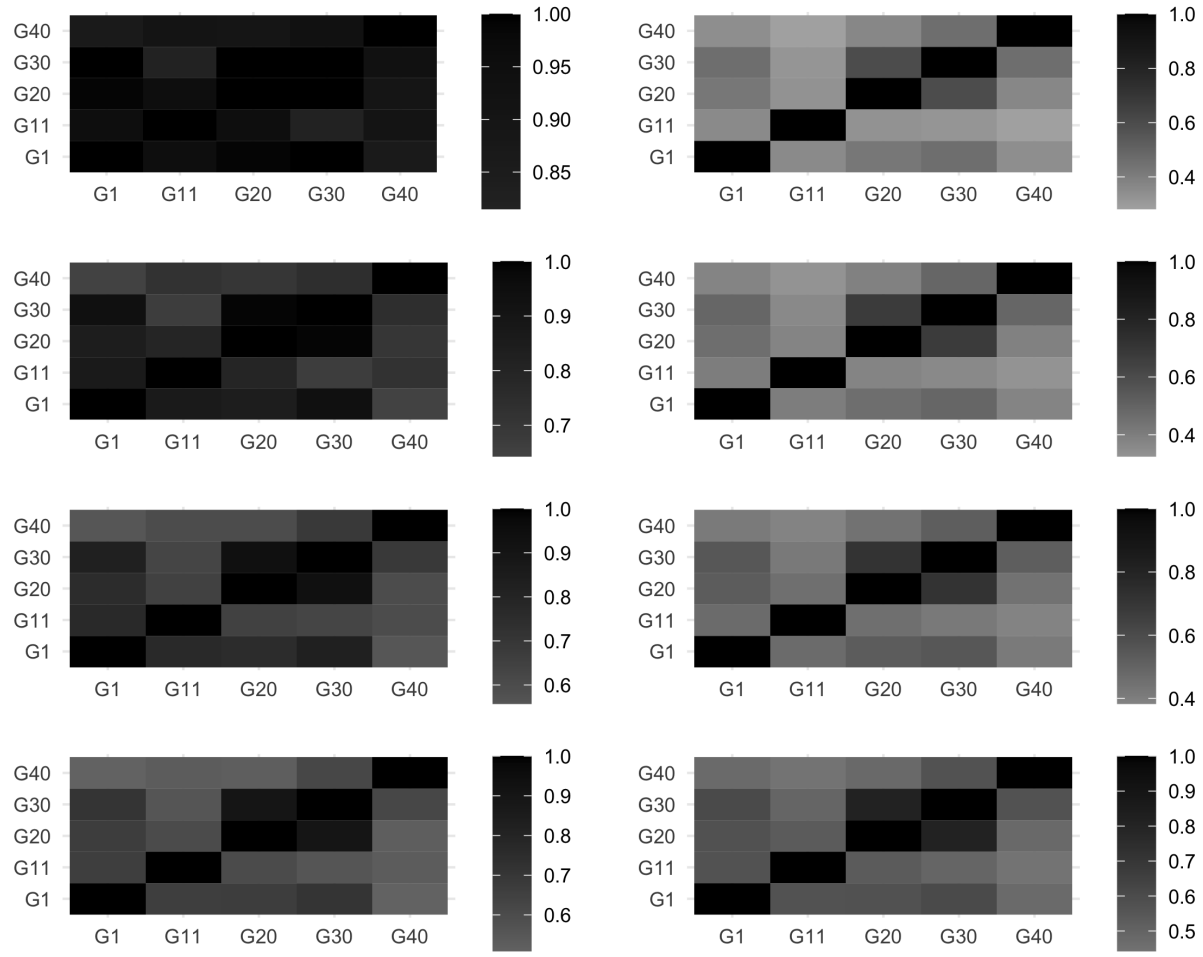

Figure A.10: Heat maps of the mean canonical correlations between SNP-sets $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$ where $i, j \in \{1, 11, 20, 30, 40\}$ across 10 replicates. From top left in counterclockwise order, each heat map corresponds to canonical correlation for the first to the eighth canonical variate pair. Darker color represents higher canonical correlation.

Appendix B

B.1 UKB phenotype definition

Myocardial infarction (MI). Cases of MI were identified by the International Classification of Disease, Ninth and Tenth Revision (ICD-10) code family I21 (Acute MI). The primary source for this code was UKB's field 131298 (Date I21 first reported). This field gathers information from hospital admissions, death records, primary care, and self-reported outcomes from surveys taken at UK Biobank assessment centers at initiation into the study and maps them to three-digit ICD-10 categories. To obtain the most up-to-date information, we also gathered this ICD-10 code family directly from hospital admission and death records. We also included cases of MI identified through UKB's algorithmically defined outcome (field 42000). Controls were required to have no evidence of certain cardiovascular diseases. Unstable angina. Cases of unstable angina were identified by the ICD-10 code I20.0, extracted from hospital admissions and death records.

Ischemic stroke (Stroke infarct). Cases of ischemic stroke were identified in a manner similar to MI, using a combination of UKB's first occurrence field 131366 (Date I63 first reported (cerebral infarction), the algorithmically defined outcome for ischemic stroke (field 42008), and the ICD-10 code I63 in hospital admission or death records. Controls were required to have no evidence of cerebrovascular disease (ICD10 codes I6*, G45*, G46*).

Stroke (Stroke any). Stroke was taken to be the first occurrence of either ischemic or hemorrhagic stroke, or of unspecified stroke via UKB fields 42006 (algorithmically defined stroke), 131368 (unspecified stroke), or ICD10 code I64. Controls were required to have no evidence of cerebrovascular disease (ICD10 codes I6*, G45*, G46*).

Percutaneous coronary intervention (PCI). Cases of PCI were identified through OPCS4 codes K40, K41, K42, K43, K44, K45, K46, K483, K49, K501, K75, K76, and UKB self-report codes 1070 (coronary angioplasty) and 1095 (coronary bypass grafts). Controls were not to have self-reported any non-coronary revascularization procedures.

Composite CVD (CVD). A composite CVD event consisted of either MI, ischemic stroke, unstable angina, or PCI. The first date of CVD was taken as the first date of any of these events. Controls were required to satisfy all of the conditions for each component outcome.

Macroalbuminuria/Microalbuminuria. Urine Albumin:Creatinine ratio (UACR) was calculated using UKB fields 30700 (urine creatinine), 30500 (urine microalbumin), and 30505 (reason for missing urine microalbumin). UACR above 33.9 was considered macroalbuminuria, while above 3.4 was considered microalbuminuria. In cases where urine microalbumin was below detectable levels, albuminuria status was inferred from urine creatinine where possible.

Chronic/Diabetic kidney disease (DKD). DKD was identified through UKB's algorithmically defined end-stage renal disease (field 42026, previously described), ICD10 codes E1*.2 (diabetes mellitus with renal complications), E18[0345] (chronic kidney disease stage 3-5, end-stage), N08.3 (glomerular disorders in diabetes mellitus) in hospital or death records, self-reported diabetic kidney disease, two or more consecutive eGFR (EPI creatinine) < 60 mL/min/1.73m2 measured 90+ days apart from either UK Biobank Assessment Center or primary care data. The date of the first DKD was taken as the first occurrence of any of the previous codes/events. Controls were required not to have micro/macroalbuminuria or a list of exclusion codes. Controls were required to have at least five years of follow-up since their diabetes diagnosis, and cases were required to have more than five years between their date of diabetes diagnosis and first DKD.

Diabetic eye disease (DR). DR was determined using the ICD10 codes E1*.3 (diabetes mellitus with ophthalmic complications), H36.0 (diabetic retinopathy), and H28.0 (Diabetic Cataract), as well as a set of primary care codes. Since most cases were identified through

primary care data, controls were required to have this data available in order to reduce misclassification. Controls were also required not to have any glaucoma, cataract, or non-diabetic/unspecified retinopathy.

B.2 Methods

B.2.1 Genotyping in ACCORD and UKB

After downloading data from dbGap, we used genetic variants genotyped on Affymetrix Axiom Biobank 1 chips from the University of North Carolina (UNC) and merged data under two different institutional review board (IRB) protocols—HMB-IRB (73941) and DS-CDKD-IRB (73944). There were 6,291 (2,335 females and 3,956 males) with 546,800 SNPs in the merged dataset. Based on self-reported ethnicity, there were 4,369 non-Hispanic whites (NHW), 935 African-Americans (AA), 381 Hispanics, and 606 others. We checked the validity of self-reported ethnicity by running the ADMIXTURE software (Alexander and Lange, 2011) with K=4, categorizing each individual into a group with the highest probability, and comparing the categories against self-reported ethnicity (see Figure B.1). We can infer that the ADMIXTURE ancestry groups 1, 2, 3, and 4 represent NHW, AA, Other, and Hispanic, respectively. Considering that Hispanics are a highly genetically heterogeneous admixed group, the distribution in ADMIXTURE ancestry group 4 (Figure B.1) appears reasonable.

Genome-wide genotyping was performed on all UK Biobank participants using the UK Biobank Axiom Array.

B.2.2 Heritability estimation using genotype data

ACCORD. We calculated a Genetic Relationship Matrix (GRM) using SNPs from all autosomes. The GRM uses SNP data to measure the relatedness between each pair of individuals

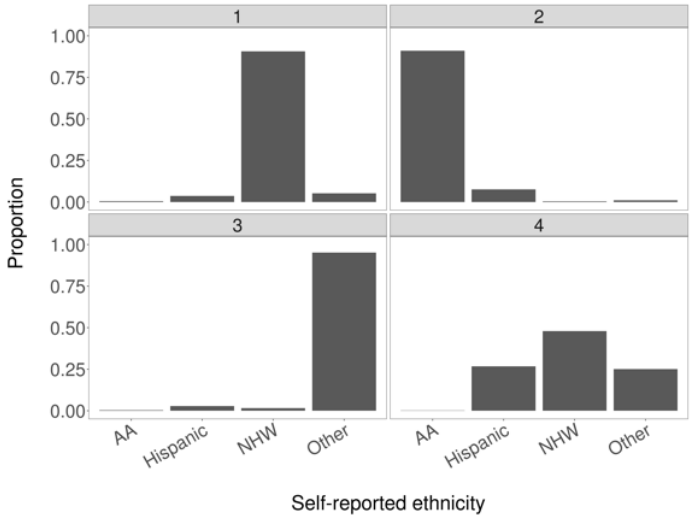

Figure B.1: Bar graph indicating the percentage of self-reported ethnicity groups categorized into each ADMIXTURE bin. Each individual is binned based on the largest proportion from ADMIXTURE.

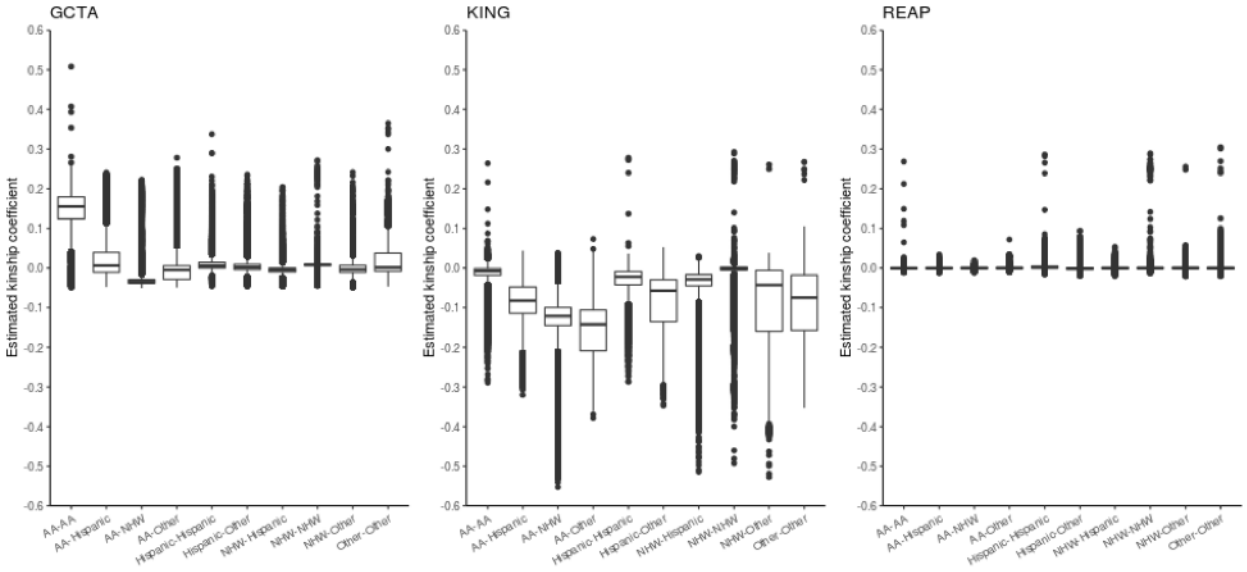

Figure B.2: Estimated kinship coefficients from software packages GCTA, KING, and REAP. Estimates from GCTA have been divided by 2 for comparability with those from other packages.

in our sample. This GRM replaces the known information about relatedness found in pedigrees. While the ACCORD trial did not deliberately recruit related individuals, we took a step to avoid inflation caused by cryptic (i.e., unknown) relatedness. We selectively exclude one of any pair of individuals with an estimated kinship greater than the separation between full and half-siblings (estimated kinship $(1/2)^{5/2}= 0.1768$) in a way to maximize the remaining sample size (Manichaikul et al., 2010; Marvel et al., 2017). Initially, we used the software package Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011a) to construct the GRM. However, the degree of relatedness calculated by GCTA appears inflated (see Figure B.2). The inflation may be mainly due to population heterogeneity in the data. Next, we try Kinship-based INference for Genome-wide association studies (KING; Manichaikul et al., 2010). As seen in Figure B.2, estimated kinship-coefficient values from KING are systematically negative, which ultimately leads the GRM to be not positive semi-definite. Finally, we use Relatedness Estimation in Admixed Populations (REAP; Thornton et al., 2012), which produces more robust results. The REAP approach requires individual ancestry proportions and allele frequencies for each ancestral population. Both proportions were obtained using the ADMIXTURE software (Alexander et al., 2009), with the number of ancestral populations specified as four (K=4). The number four was chosen because there were four different self-reported ethnic groups (NHW, AA, Hispanic and other).

We only extract NHW samples after pruning related individuals, which leaves us with 4,329 samples. With the GRM constructed from REAP, heritability is estimated via GCTA (Yang et al., 2011a) that uses a liability model (Falconer, 1965; Lee et al., 2011). We adjust for sex, CVD history at baseline, age at baseline, and the top five genetic principal components. An additional analysis that incorporates interaction with glycemic intensive treatment arm (intensive=1, standard=0) is shown in Figure B.6. We also estimate the genetic correlation between binary traits via the GCTA software (Yang et al., 2011a; Lee et al., 2011). The following covariates are adjusted for: sex, CVD history at baseline, age at baseline, and the top five genetic principal components.

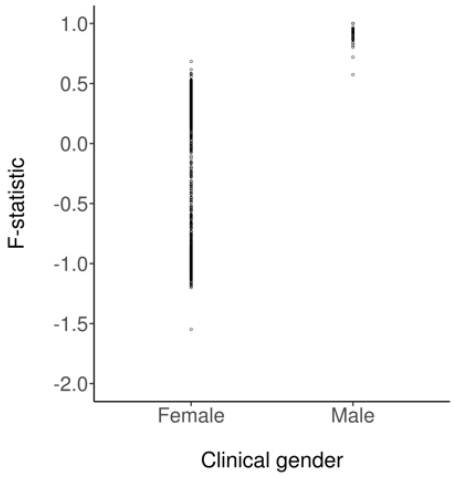

Figure B.3: Distribution of F (inbreeding) coefficients against clinical gender.

UKB. We extract the NHW diabetes cohort (n=26,387) and compute the GRM via the REAP approach, for which necessary proportions are obtained from the ADMIXTURE software with K=3. No individuals are pruned out under the relatedness threshold (0.1768). We estimate heritability using the GREML-SC approach while adjusting for sex, age in 2010, and the top ten genetic principal components. Also calculated using the UKB genotype data are genetic correlations between phenotypes (see Table B.1).

B.2.3 Imputation

ACCORD. Prior to imputation, we perform quality control steps on the data. First, we check if there are mismatches between genetic gender and clinical gender. We run plinkv1.9 `--check-sex` option along with `--split-x` and make an F-statistic against sex-label plot (see Figure B.3). As expected, we see a big tight clump near 1 for males while a more widely dispersed set of values centered near 0 (Chang, 2020). Even though some individuals do not pass the default threshold set in plink, we decide not to remove any individuals since the data exhibit an expected pattern.

Next, following the procedure in Marvel et al. (2017), we compute HWE values for each of the self-reported ethnicity groups: NHW, AA, Hispanic, and other. Any SNPs deviating

from p value 1×10^{-5} in at least two of the four groups are excluded. This step reduces the number of variants to 542,847. Additionally, we check alleles to allow only A, C, G, T and exclude SNPs with a missing rate $> 3\%$ and monomorphic sites (MAF < 0.0000001). We also exclude individuals with a genotype missing rate > 0.03. After the aforementioned step, we retain 6,279 individuals and 465,011 variants.

Data imputation is done using a two-step approach where the genotype calls are pre-phased using Eagle v2.4.1 (Loh et al., 2016), and then imputation is done using Minimac4 (Das et al., 2016) with default options. Both steps use the 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium, 2015) as a reference panel. After discarding imputed variants with $R^2 < 0.3$ and MAF < 0.0003, we have a total of 25,667,109 imputed variants for the downstream analyses. Additionally, we extract the NHW samples filtered from the REAP approach earlier (n=4,329). With 11 out of 4,329 individuals have been removed during the pre-imputation QC steps, we proceed with the downstream analyses with 4,318 NHW individuals.

UKB. We use the imputed datasets released by UK Biobank. On the imputed datasets, we extract autosomal variants with imputation info score > 0.3 and remove multiallelic variants. Also excluded are variants with missing genotype rate > 0.05, Hardy-Weinberg equilibrium test $p < 1 \times 10^{-6}$, and MAF < 0.0001. After the filtering steps, we have a total of 33,932,888 variants.

B.2.4 GREML-LDMS

On the imputed datasets, we employ the GREML-LDMS method. For the GREML-LDMS-I approach, we follow the design laid out in Evans et al. (2018). First, we calculate segment-based LD scores using the default settings—200-kb block size with a 100-kb overlap—using the GCTA software and stratify SNPs into high LD and low LD score groups using the median as a threshold. In each LD group, SNPs are further partitioned into four MAF bins: common (MAF ≥ 0.05), uncommon ($0.01 \leq$ MAF < 0.05), rare ($0.0025 \leq$ MAF < 0.01),

and very rare ($0.0003 \leq \text{MAF} < 0.0025$). Then GRMs are computed using SNPs stratified into eight groups, hence creating eight GRMs. Finally, we run GREML analyses on each binary phenotype with fixed covariates.

ACCORD. After filtering steps, the ACCORD imputed dataset contains 4,318 NHW individuals and 15,349,988 variants. Covariates adjusted are sex, age at baseline, history of CVD at baseline, and the top five genetic principal components.

UKB. On the UKB imputed datasets, we adjust for sex, age in 2010, and the top ten genetic principal components.

B.2.5 GWAS

ACCORD. GWAS for complications are performed in 4,318 NHW participants. After filtration for variants with MAF 0.01, as done in Bulik-Sullivan et al. (2015), 8,480,081 SNPs form the GWAS panel. The association between each variant and each complication is tested by logistic regression in PLINK2.0 (Chang, 2020), assuming an additive genetic model and adjusting for sex, CVD history at baseline, age at baseline, and the top five genetic principal components. Manhattan and quantile-quantile (QQ) plots are provided in Figure B.4.

UKB. GWAS for complications is performed in 26,387 NHW samples. After MAF filtration (MAF 0.01), 8,949,996 variants form the GWAS panel. We adjust for sex, age in 2010, and the top ten genetic principal components. Manhattan and QQ plots are provided in Figure B.5.

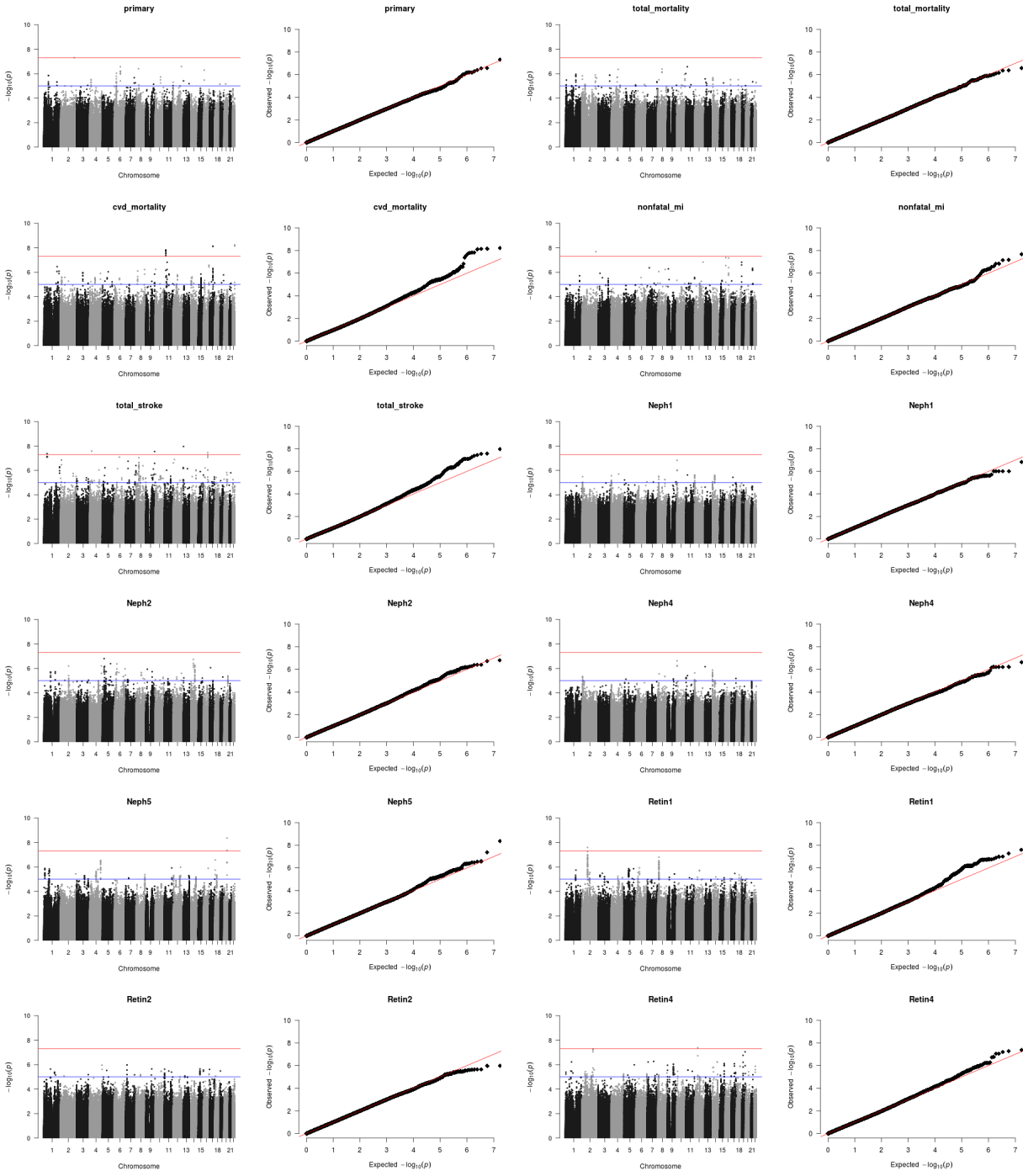

Figure B.4: Manhattan and QQ plots of GWAS p-values for the ACCORD phenotypes. Red line signifies genome-wide significance level ($p = 5 \times 10^{-8}$) while the blue line is a suggestive line ($p = 1 \times 10^{-5}$).

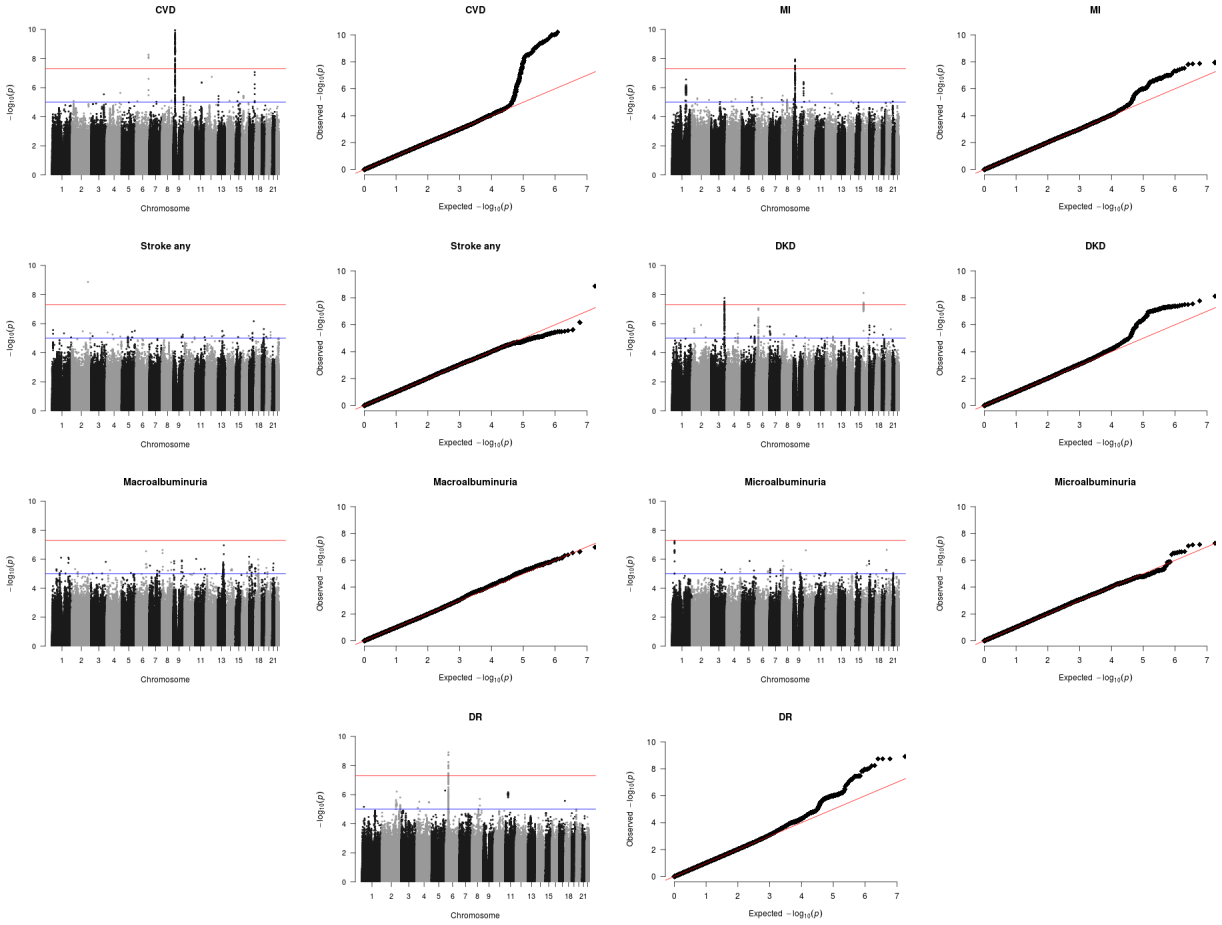

Figure B.5: Manhattan and QQ plots of GWAS p-values for the UKB phenotypes. Red line signifies genome-wide significance level ($p = 5 \times 10^{-8}$) while the blue line is a suggestive line ($p = 1 \times 10^{-5}$).

B.2.6 Stratified LD score regression (S-LDSC)

Here we partition SNP heritability, applying S-LDSC to GWAS summary statistics for the trait of interest. In none cell type-specific analyses, we use the 'full baseline model' generated by Finucane et al. (2015). The full baseline model is comprised of 53 overlapping functional categories (including coding, promoter, enhancer, and conserved regions) and is not specific to any cell type. In tissue-type specific analyses, we use the 53 specifically expressed gene annotations curated from the Genotype-Tissue Expression (GTEx) project (GTEx Consor-

tium, 2015) by Finucane et al. (2018). For all S-LDSC analyses, we use 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium, 2015) European population SNPs as an LD reference panel. All annotations and reference panel data were obtained from Alkes Price's group data repository (see URLs in Appendix B.4).

B.3 Additional tables and figures

	DKD	Microalbuminuria	DR
CVD	0.25 (0.28)	-0.11 (0.24)	0.26 (0.25)
DKD		0.36 (0.27)	0.35 (0.35)
Microalbuminuria			0.07 (0.25)

Table B.1: Genetic correlation estimates and the standard errors between selected phenotypes using the UKB genotype data. Adjusted for sex, age in 2010, and the top ten genetic principal components.

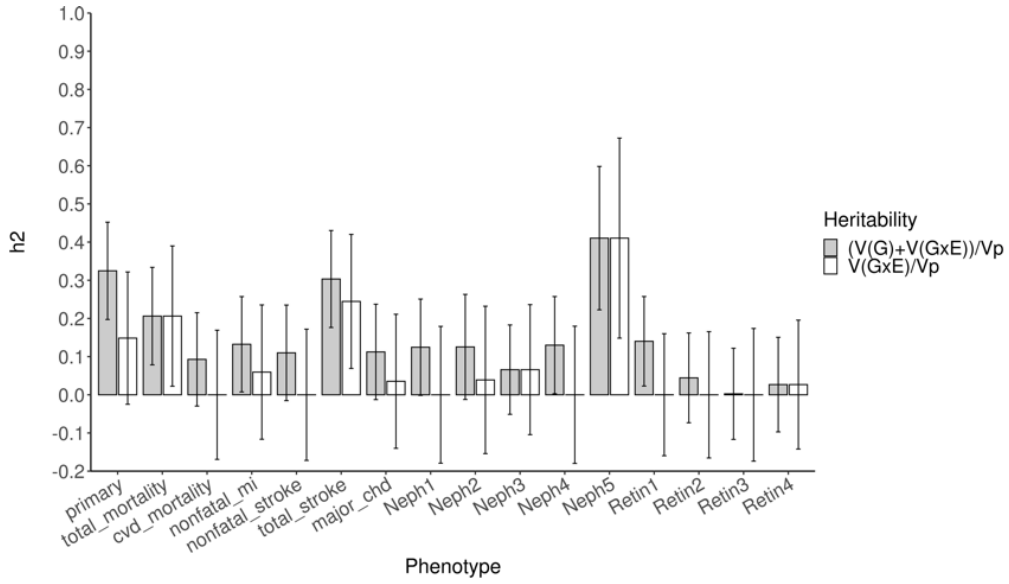

Figure B.6: Heritability estimates and standard errors of diabetes complication outcomes using the ACCORD genotype data and incorporating interaction with intensive glycemic treatment. The grey bar represents the genetic plus interaction components, while the white bar signifies the interaction component.

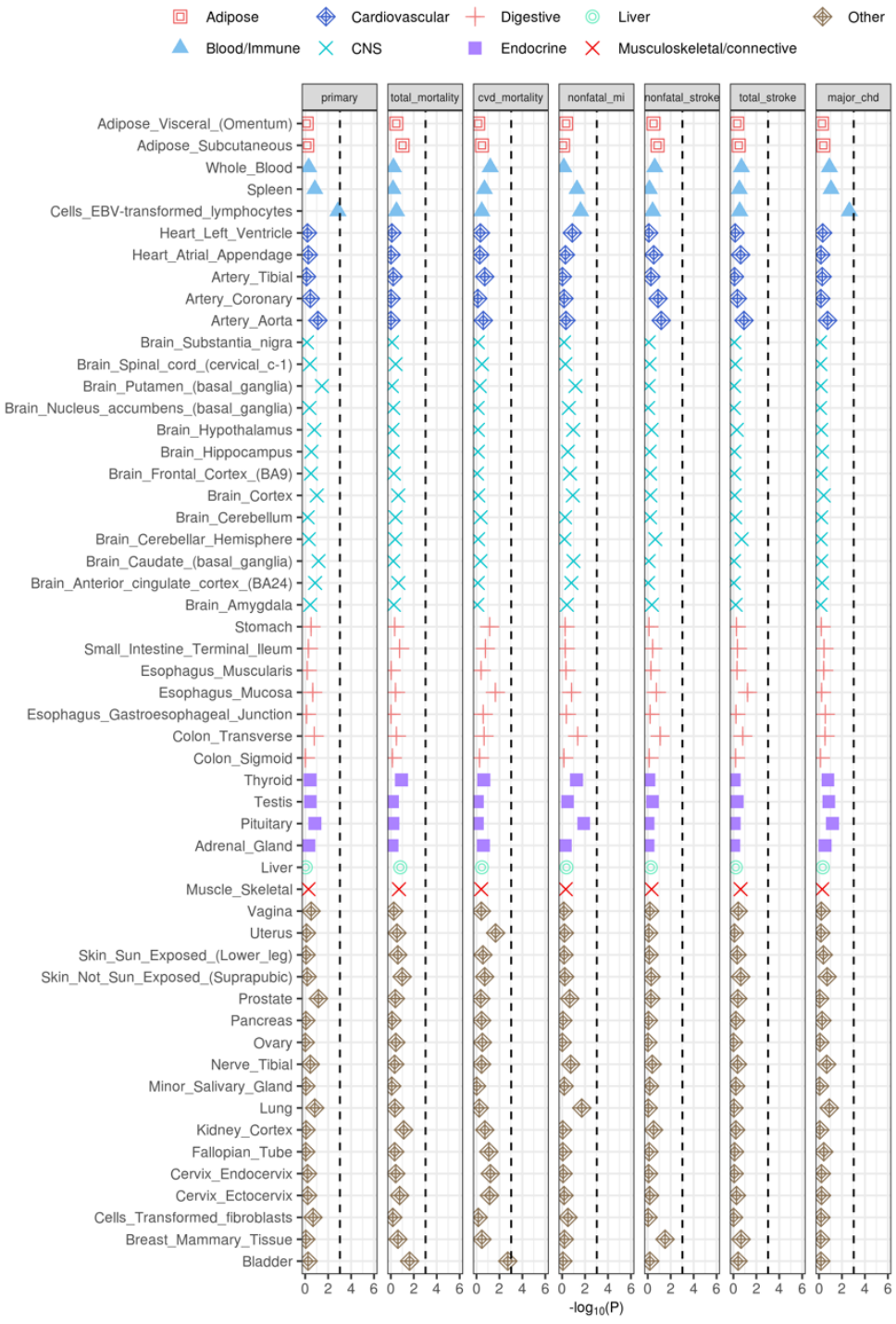

Figure B.7: Enrichment of the ACCORD macrovascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (2018). The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).

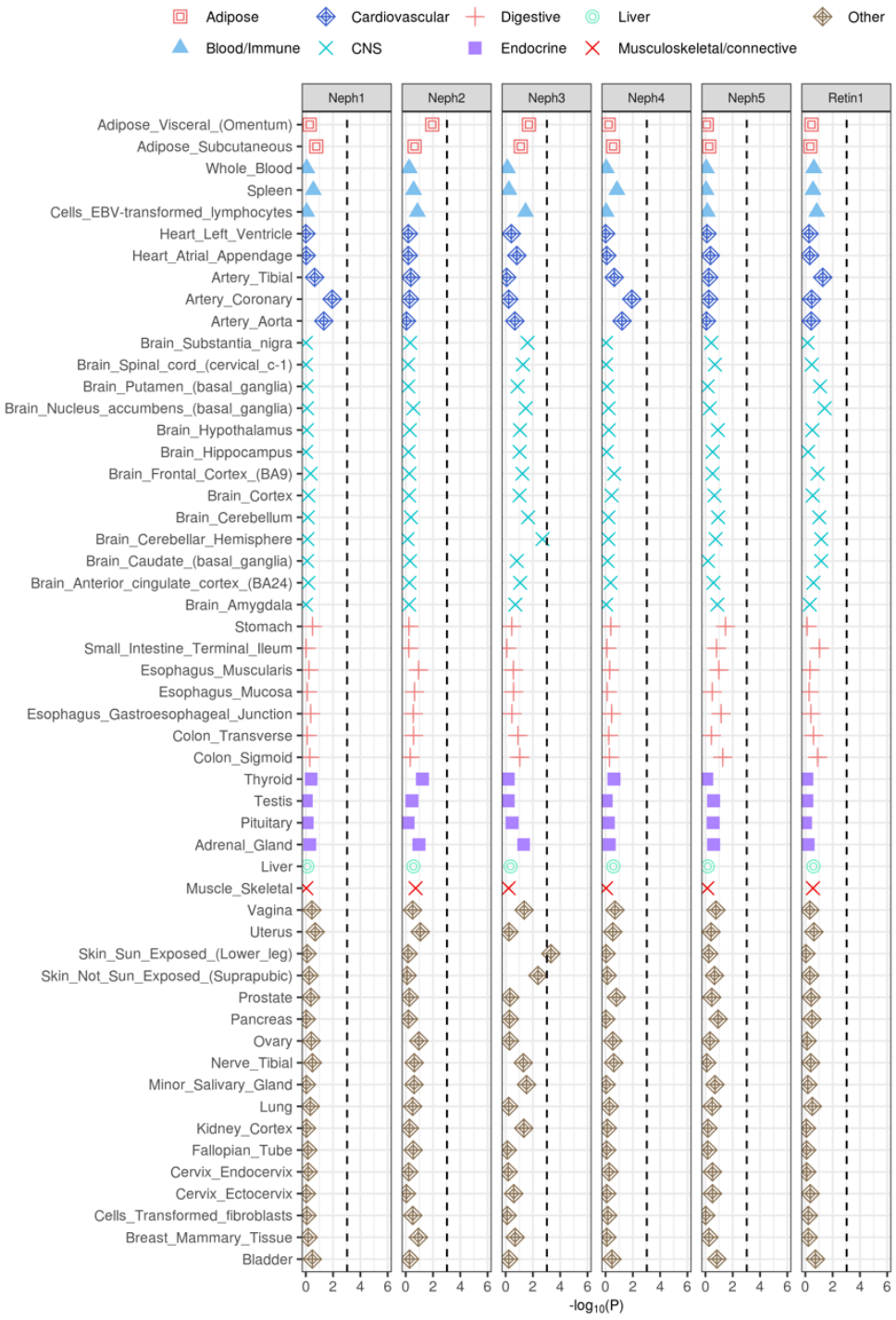

Figure B.8: Enrichment of the ACCORD microvascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (2018). The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).

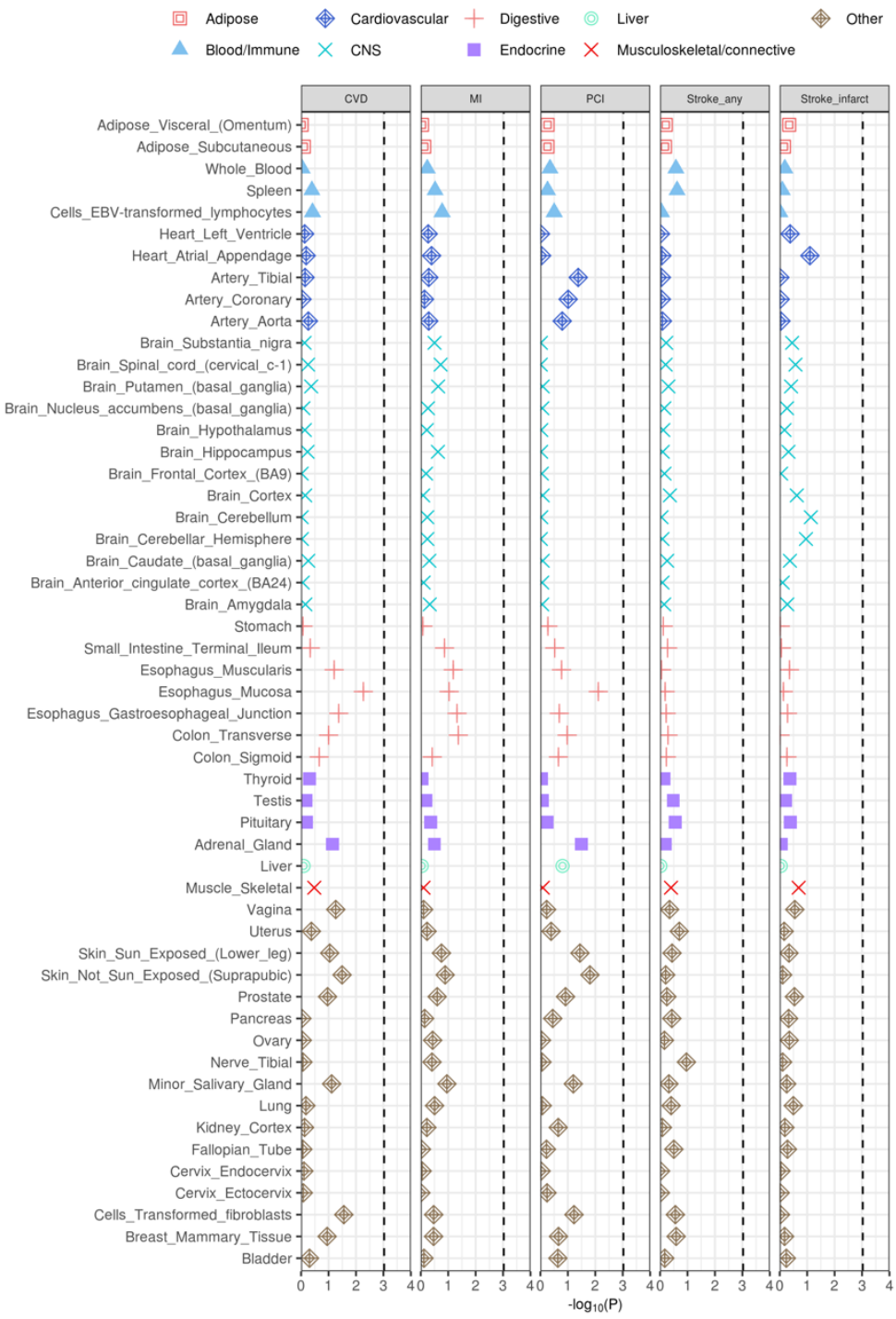

Figure B.9: Enrichment of the UKB macrovascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (2018). The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).

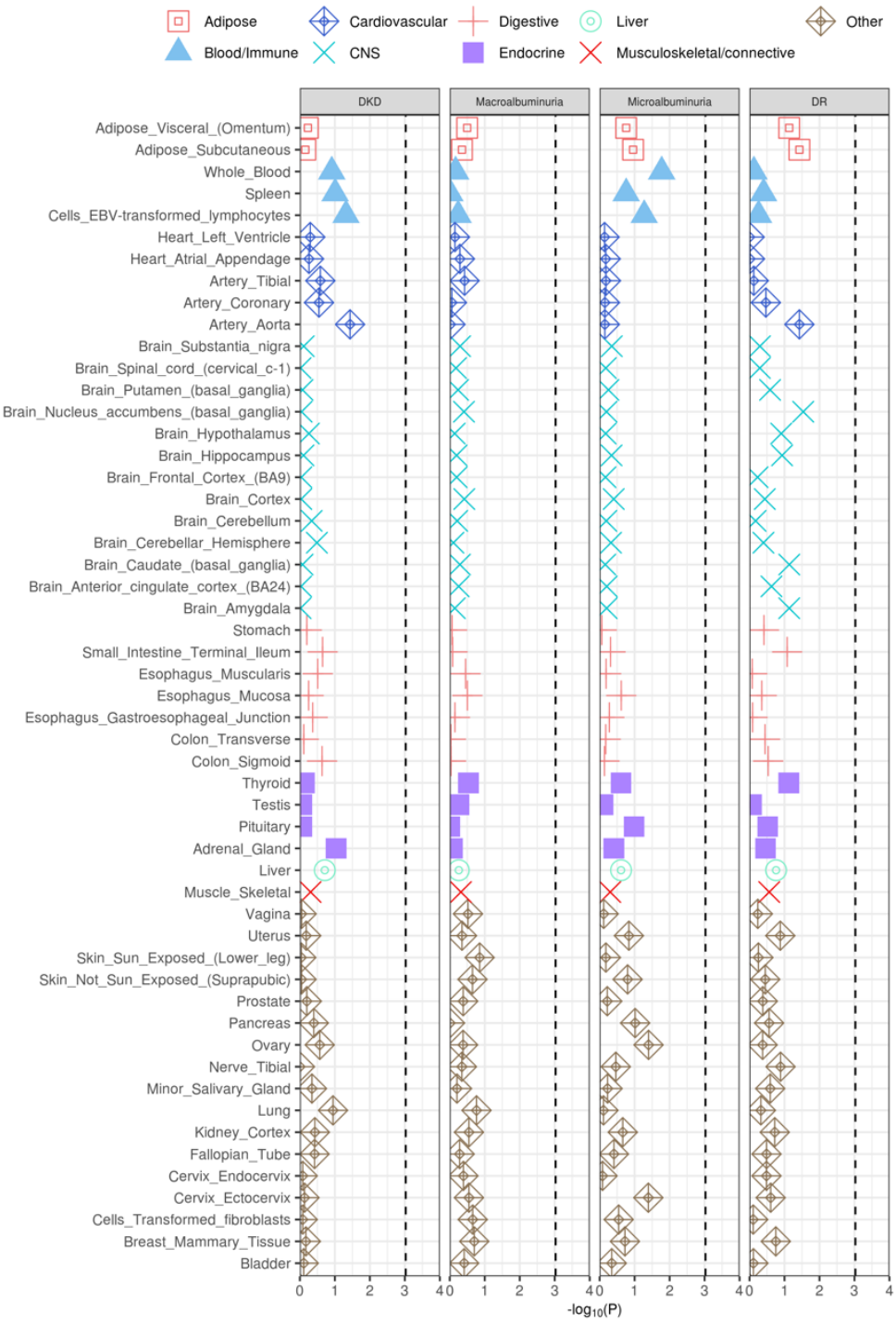

Figure B.10: Enrichment of the UKB microvascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (17). The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).

B.4 URLs

Baseline LDSC annotations, https://data.broadinstitute.org/alkesgroup/LDSCORE/. Finucane GTEx annotations, https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores/. LDSC, https://github.com/bulik/ldsc/wiki.

Bibliography

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Abifadel, M., Rabès, J.-P., Devillers, M., Munnich, A., Erlich, D., Junien, C., Varret, M., and Boileau, C. (2009). Mutations and polymorphisms in the proprotein convertase subtilisin kexin 9 (PCSK9) gene in cholesterol metabolism and disease. *Human Mutation*, 30(4):520–529.

Action to Control Cardiovascular Risk in Diabetes Study Group (2008). Effects of intensive glucose lowering in type 2 diabetes. *New England Journal of Medicine*, 358(24):2545–2559.

Afkarian, M., Zelnick, L. R., Hall, Y. N., Heagerty, P. J., Tuttle, K., Weiss, N. S., and De Boer, I. H. (2016). Clinical manifestations of kidney disease among us adults with diabetes, 1988-2014. *The Journal of the American Medical Association*, 316(6):602–610.

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12(1):246.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664.

American Diabetes Association (1998). Implications of the united kingdom prospective diabetes study. *Diabetes Care*, 21(12):2180–2184.

American Diabetes Association (2018). 9. cardiovascular disease and risk management: standards of medical care in diabetes—2018. *Diabetes Care*, 41(Supplement 1):S86–S104.

Arar, N. H., Freedman, B. I., Adler, S. G., Iyengar, S. K., Chew, E. Y., Davis, M. D., Satko, S. G., Bowden, D. W., Duggirala, R., Elston, R. C., et al. (2008). Heritability of the severity of diabetic retinopathy: the find-eye study. *Investigative Ophthalmology & Visual Science*, 49(9):3839–3845.

Bakin, S. (1999). *Adaptive Regression and Model Selection in Data Mining Problems*. PhD thesis.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773.

Bates, D. M. and Pinheiro, J. C. (1998). Computational methods for multilevel modelling. *University of Wisconsin, Madison, WI*, pages 1–29.

Benn, M., Nordestgaard, B. G., Jensen, J. S., Grande, P., Sillesen, H., and Tybjærg-Hansen, A. (2005). Polymorphism in apob associated with increased low-density lipoprotein levels in both genders in the general population. *The Journal of Clinical Endocrinology & Metabolism*, 90(10):5797–5803.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, 40(6):695.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077.

Borch-Johnsen, K., Nørgaard, K., Hommel, E., Mathiesen, E. R., Jensen, J. S., Deckert, T., and Parving, H.-H. (1992). Is diabetic nephropathy an inherited complication? *Kidney International*, 41(4):719–722.

Bowden, D. W. (2002). Genetics of diabetes complications. *Current Diabetes Reports*, 2(2):191–200.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3):369.

Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., Bielak, L. F., Zhao, W., Smith, J. A., Peyser, P. A., et al. (2016). A statistical approach

for testing cross-phenotype effects of rare variants. *The American Journal of Human Genetics*, 98(3):525–540.

Broadaway, K. A., Duncan, R., Conneely, K. N., Almli, L. M., Bradley, B., Ressler, K. J., and Epstein, M. P. (2015). Kernel approach for modeling interaction effects in genetic association studies of complex quantitative traits. *Genetic Epidemiology*, 39(5):366–375.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., and Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

Cannon, C. P., Blazing, M. A., Giugliano, R. P., McCagg, A., White, J. A., Theroux, P., Darius, H., Lewis, B. S., Ophuis, T. O., Jukema, J. W., et al. (2015). Ezetimibe added to statin therapy after acute coronary syndromes. *New England Journal of Medicine*, 372(25):2387–2397.

Centers for Disease Control and Prevention (2020). *National Diabetes Statistics Report, 2020*. Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services, Altlanta, GA.

Chang, C. C. (2020). Data management and summary statistics with plink. *Statistical Population Genomics*, page 49.

Chen, H., Meigs, J. B., and Dupuis, J. (2014). Incorporating gene-environment interaction in testing for association with rare genetic variants. *Human Heredity*, 78(2):81–90.

Chen, J. and Chen, Z. (2008). Extended bayesian information criterion for model selection with large model space. *Biometrika*, 95:759–771.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769.

Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364.

Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S., Lu, X., and Ruden, D. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., and Hobbs, H. H. (2005). Low ldl cholesterol in individuals of african descent resulting from frequent nonsense mutations in pcsk9. *Nature Genetics*, 37(2):161–165.

Craven, M. and Bockhorst, J. (2005). Markov networks for detecting overalpping elements in sequence data. In *Advances in Neural Information Processing Systems*, pages 193–200.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287.

Davis, J., Burnside, E. S., de Castro Dutra, I., Page, D., Ramakrishnan, R., Costa, V. S., and Shavlik, J. W. (2005). View learning for statistical relational learning: With an application to mammography. In *IJCAI*, pages 677–683. Citeseer.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38.

Dering, C., Hemmelmann, C., Pugh, E., and Ziegler, A. (2011). Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genetic Epidemiology*, 35(S1):S12–S17.

Deshmukh, H. A., Colhoun, H. M., Johnson, T., McKeigue, P. M., Betteridge, D. J., Durrington, P. N., Fuller, J. H., Livingstone, S., Charlton-Menys, V., Neil, A., et al. (2012). Genome-wide association study of genetic determinants of LDL-c response to atorvastatin therapy: importance of lp (a). *Journal of Lipid Research*, 53(5):1000–1011.

Diabetes Control and Complications Trial Research Group (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine*, 329(14):977–986.

Diabetes Control and Complications Trial Research Group (1997). Clustering of long-term complications in families with diabetes in the diabetes control and complications trial. *Diabetes*, 46(11):1829–1839.

Dutta, D., Scott, L., Boehnke, M., and Lee, S. (2019). Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genetic Epidemiology*, 43(1):4–23.

Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G. R., Das, S., Gazal, S., Bjelland, D. W., De Candia, T. R., Goddard, M. E., Neale, B. M., et al. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, 50(5):737–745.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *Annals of Statistics*, 40(4):2043.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228.

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics*, 50(4):621–629.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9):1026–1034.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135.

Goadrich, M., Oliphant, L., and Shavlik, J. (2004). Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction. In *International Conference on Inductive Logic Programming*, pages 98–115. Springer.

GTEx Consortium (2015). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Gudmundsson, J., Sulem, P., Gudbjartsson, D. F., Masson, G., Agnarsson, B. A., Benediktsdottir, K. R., Sigurdsson, A., Magnusson, O. T., Gudjonsson, S. A., Magnusdottir, D. N., et al. (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics*, 44(12):1326–1329.

Hackinger, S. and Zeggini, E. (2017). Statistical methods to detect pleiotropy in human complex traits. *Open Biology*, 7(11):170125.

Hallman, D. M., Huber, J. C., Gonzalez, V. H., Klein, B. E., Klein, R., and Hanis, C. L. (2005). Familial aggregation of severity of diabetic retinopathy in mexican americans from starr county, texas. *Diabetes Care*, 28(5):1163–1168.

Hao, X., Zeng, P., Zhang, S., and Zhou, X. (2018). Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS genetics*, 14(1):e1007186.

Harjutsalo, V., Katoh, S., Sarti, C., Tajima, N., and Tuomilehto, J. (2004). Population-based assessment of familial clustering of diabetic nephropathy in type 1 diabetes. *Diabetes*, 53(9):2449–2454.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

He, M. and Allen, A. (2010). Testing gene–treatment interactions in pharmacogenetic studies. *Journal of Biopharmaceutical Statistics*, 20(2):301–314.

Heid, I. M., Boes, E., Müller, M., Kollerits, B., Lamina, C., Coassin, S., Gieger, C., Döring, A., Klopp, N., Frikke-Schmidt, R., et al. (2008). Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based kora study sheds new light on intergenic regions. *Circulation: Cardiovascular Genetics*, 1(1):10–20.

Helgeland, Ø., Hertel, J. K., Molven, A., Ræder, H., Platou, C. G., Midthjell, K., Hveem, K., Nygård, O., Njølstad, P. R., and Johansson, S. (2015). The chromosome 9p21 cvd-and t2d-associated regions in a norwegian population (the hunt2 survey). *International journal of endocrinology*, 2015.

Hietala, K., Forsblom, C., Summanen, P., and Groop, P.-H. (2008). Heritability of proliferative diabetic retinopathy. *Diabetes*, 57(8):2176–2180.

Hoffmann, T. J., Theusch, E., Haldar, T., Ranatunga, D. K., Jorgenson, E., Medina, M. W., Kvale, M. N., Kwok, P.-Y., Schaefer, C., Krauss, R. M., et al. (2018). A large electronic-health-record-based genome-wide study of serum lipids. *Nature Genetics*, 50(3):401–413.

Holmen, O. L., Zhang, H., Fan, Y., Hovelson, D. H., Schmidt, E. M., Zhou, W., Guo, Y., Zhang, J., Langhammer, A., Løchen, M.-L., et al. (2014). Systematic evaluation of coding variation identifies a candidate causal variant in tm6sf2 influencing total cholesterol and myocardial infarction risk. *Nature genetics*, 46(4):345–351.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2):339–355.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37.

Ismail-Beigi, F., Craven, T., Banerji, M. A., Basile, J., Calles, J., Cohen, R. M., Cuddihy, R., Cushman, W. C., Genuth, S., Grimm Jr, R. H., et al. (2010). Effect of intensive treatment of hyperglycaemia on microvascular outcomes in type 2 diabetes: an analysis of the ACCORD randomised trial. *The Lancet*, 376(9739):419–430.

Keller, M. C. and Coventry, W. L. (2005). Quantifying and addressing parameter indeterminacy in the classical twin design. *Twin Research and Human Genetics*, 8(3):201–213.

Khuri, A. I. and Sahai, H. (1985). Variance components analysis: a selective literature survey. *International Statistical Review/Revue Internationale de Statistique*, pages 279–300.

Kok, S. and Domingos, P. (2005). Learning the structure of Markov logic networks. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 441–448. ACM.

Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, 82(397):97–105.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lange, K. (2016). *MM Optimization Algorithms*. SIAM.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20.

Lange, K., Papp, J. C., Sinsheimer, J. S., Sripracha, R., Zhou, H., and Sobel, E. M. (2013). Mendel: the swiss army knife of genetic analysis programs. *Bioinformatics*, 29(12):1568–1570.

Lange, L. A., Bowden, D. W., Langefeld, C. D., Wagenknecht, L. E., Carr, J. J., Rich, S. S., Riley, W. A., and Freedman, B. I. (2002). Heritability of carotid artery intima-medial thickness in type 2 diabetes. *Stroke*, 33(7):1876–1881.

Lange, L. A., Burdon, K., Langefeld, C. D., Liu, Y., Beck, S. R., Rich, S. S., Freedman, B. I., Brosnihan, K. B., Herrington, D. M., Wagenknecht, L. E., et al. (2006). Heritability and expression of c-reactive protein in type 2 diabetes in the diabetes heart study. *Annals of Human Genetics*, 70(6):717–725.

Lange, L. A., Hu, Y., Zhang, H., Xue, C., Schmidt, E. M., Tang, Z.-Z., Bizon, C., Lange, E. M., Smith, J. D., Turner, E. H., et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with ldl cholesterol. *The American Journal of Human Genetics*, 94(2):233–245.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.

Lee, S., Miropolsky, L., and Wu, M. (2017a). `SKAT`*: SNP-Set (Sequence) Kernel Association Test*. R package version 1.3.2.1.

Lee, S., Won, S., Kim, Y. J., Kim, Y., Consortium, T.-G., Kim, B.-J., and Park, T. (2017b). Rare variant association test with multiple phenotypes. *Genetic Epidemiology*, 41(3):198–209.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305.

Leslie, R. and Pyke, D. (1982). Diabetic retinopathy in identical twins. *Diabetes*, 31(1):19–21.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.

Lim, E., Chen, H., Dupuis, J., and Liu, C.-T. (2020). A unified method for rare variant analysis of gene-environment interactions. *Statistics in medicine*, 39(6):801–813.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, 72(1):156–164.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., et al. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics*, 48(11):1443.

Looker, H. C., Nelson, R. G., Chew, E., Klein, R., Klein, B. E., Knowler, W. C., and Hanson, R. L. (2007). Genome-wide linkage analyses to identify loci for diabetic retinopathy. *Diabetes*, 56(4):1160–1166.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384.

Maity, A., Sullivan, P. F., and Tzeng, J.-i. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 36(7):686–695.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747.

Marvel, S. W., Rotroff, D. M., Wagner, M. J., Buse, J. B., Havener, T. M., McLeod, H. L., and Motsinger-Reif, A. A. (2017). Common and rare genetic markers of lipid variation in subjects with type 2 diabetes from the accord clinical trial. *PeerJ*, 5:e3187.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195.

McPherson, R. and Tybjaerg-Hansen, A. (2016). Genetics of coronary artery disease. *Circulation Research*, 118(4):564–578.

Meng, W., Shah, K. P., Pollack, S., Toppila, I., Hebert, H. L., McCarthy, M. I., Groop, L., Ahlqvist, E., Lyssenko, V., Agardh, E., et al. (2018). A genome-wide association study suggests new evidence for an association of the nadph oxidase 4 (nox 4) gene with severe diabetic retinopathy in type 2 diabetes. *Acta Ophthalmologica*, 96(7):e811–e819.

Noble, J. A. and Erlich, H. A. (2012). Genetics of type 1 diabetes. *Cold Spring Harbor perspectives in medicine*, 2(1):a007732.

Paila, U., Chapman, B. A., Kirchner, R., and Quinlan, A. R. (2013). Gemini: integrative

exploration of genetic variation and genome annotations. *PLoS Computational Biology*, 9(7).

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, 109:109–129.

Postmus, I., Trompet, S., Deshmukh, H. A., Barnes, M. R., Li, X., Warren, H. R., Chasman, D. I., Zhou, K., Arsenault, B. J., Donnelly, L. A., et al. (2014). Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nature Communications*, 5:5068.

Quinn, M., Angelico, M., Warram, J., and Krolewski, A. (1996). Familial factors determine the development of diabetic nephropathy in patients with iddm. *Diabetologia*, 39(8):940–945.

Raghavan, V., Bollmann, P., and Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229.

Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics*, 43(11):1066–1073.

Rivas, M. A. and Moutsianas, L. (2015). Power of rare variant aggregate tests. In *Assessing Rare Variation in Complex Traits*, pages 185–199. Springer.

Robinson, D. L. (1987). Estimation and use of variance components. *The Statistician*, pages 3–14.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432.

Salem, R. M., Todd, J. N., Sandholm, N., Cole, J. B., Chen, W.-M., Andrews, D., Pezzolesi, M. G., McKeigue, P. M., Hiraki, L. T., Qiu, C., et al. (2019). Genome-wide association study of diabetic kidney disease highlights biology involved in glomerular basement membrane collagen. *Journal of the American Society of Nephrology*, 30(10):2000–2016.

Sandholm, N., Van Zuydam, N., Ahlqvist, E., Juliusdottir, T., Deshmukh, H. A., Rayner, N. W., Di Camillo, B., Forsblom, C., Fadista, J., Ziemek, D., et al. (2017). The genetic landscape of renal complications in type 1 diabetes. *Journal of the American Society of Nephrology*, 28(2):557–574.

Schaid, D. J., Sinnwell, J. P., Larson, N. B., and Chen, J. (2020). Penalized variance components for association of multiple genes with traits. *Genetic Epidemiology*, n/a(n/a).

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.

Shah, H. S., Gao, H., Morieri, M. L., Skupien, J., Marvel, S., Paré, G., Mannino, G. C., Buranasupkajorn, P., Mendonca, C., Hastings, T., et al. (2016). Genetic predictors of cardiovascular mortality during intensive glycemic control in type 2 diabetes: findings from the accord clinical trial. *Diabetes Care*, 39(11):1915–1924.

Shlyakhter, I., Sabeti, P. C., and Schaffner, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Simonson, M. A., Wills, A. G., Keller, M. C., and McQueen, M. B. (2011). Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. *BMC Medical Genetics*, 12(1):1–9.

Singla, P. and Domingos, P. (2005). Discriminative training of markov logic networks. In *Proceedings of the 20th National Conference on Artificial Intelligene (AAAI)*, volume 5, pages 868–873. AAAI Press.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483.

Southam, L., Gilly, A., Süveges, D., Farmaki, A.-E., Schwartzentruber, J., Tachmazidou, I., Matchan, A., Rayner, N. W., Tsafantakis, E., Karaleftheri, M., et al. (2017). Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nature communications*, 8(1):1–11.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). 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).

Suo, C., Toulopoulou, T., Bramon, E., Walshe, M., Picchioni, M., Murray, R., and Ott, J. (2013). Analysis of multiple phenotypes in genome-wide genetic mapping studies. *BMC Bioinformatics*, 14(1):151.

Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.-P., Mahajan, A., Lagou, V., Marullo, L.,

Ferreira, T., Miraglio, B., Timonen, S., et al. (2015). The impact of low-frequency and rare variants on lipid levels. *Nature Genetics*, 47(6):589–597.

Tachmazidou, I., Dedoussis, G., Southam, L., Farmaki, A.-E., Ritchie, G. R., Xifara, D. K., Matchan, A., Hatzikotoulas, K., Rayner, N. W., Chen, Y., et al. (2013). A rare functional cardioprotective apoc3 variant has risen in frequency in distinct population isolates. *Nature communications*, 4(1):1–6.

Tenesa, A. and Haley, C. S. (2013). The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics*, 14(2):139–149.

Thompson, W. A. et al. (1962). The problem of negative estimates of variance components. *The Annals of Mathematical Statistics*, 33(1):273–289.

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., and Risch, N. (2012). Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130.

U.S. Food & Drug Administration (2018). Preventable adverse drug reactions: A focus on drug interactions. `https://www.fda.gov/drugs/drug-interactions-labeling/preventable-adverse-drug-reactions-focus-drug-interactions`. Accessed: 2019-09-07.

Van Zuydam, N. R., Ahlqvist, E., Sandholm, N., Deshmukh, H., Rayner, N. W., Abdalla, M., Ladenvall, C., Ziemek, D., Fauman, E., Robertson, N. R., et al. (2018). A genome-wide

association study of diabetic kidney disease in subjects with type 2 diabetes. *Diabetes*, 67(7):1414–1427.

Wagenknecht, L. E., Langefeld, C. D., Bowden, D. W., Carr, J. J., Freedman, B. I., and Rich, S. S. (2001). Familial aggregation of coronary artery calcium in families with type 2 diabetes. *Diabetes*, 50(4):861–866.

Wallace, C., Newhouse, S. J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R. J., Marçano, A. C. B., Hajat, C., et al. (2008). Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *The American Journal of Human genetics*, 82(1):139–149.

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518.

Wu, B. and Pankow, J. S. (2016). Sequence kernel association test of multiple continuous phenotypes. *Genetic Epidemiology*, 40(2):91–100.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Xu, J., Wei, W. B., Yuan, M. X., Yuan, S. Y., Wan, G., Zheng, Y. Y., Li, Y. B., Wang, S., Xu, L., Fu, H. J., et al. (2012). Prevalence and risk factors for diabetic retinopathy: the beijing communities diabetes study 6. *Retina*, 32(2):322–329.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10):1114.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., De Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., et al. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43(6):519.

Yang, T., Chen, H., Tang, H., Li, D., and Wei, P. (2019). A powerful and data-adaptive test for rare-variant–based gene-environment interaction analysis. *Statistics in Medicine*, 38(7):1230–1244.

Yang, Y. and Zou, H. (2014). `gglasso`*: Group Lasso Penalized Learning Using A Unified BMD Algorithm*. R package version 1.3.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zdravkovic, S., Wienke, A., Pedersen, N., Marenberg, M., Yashin, A., and De Faire, U. (2002). Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 swedish twins. *Journal of Internal Medicine*, 252(3):247–254.

Zdravkovic, S., Wienke, A., Pedersen, N. L., and de Faire, U. (2007). Genetic influences on angina pectoris and its impact on coronary heart disease. *European Journal of Human Genetics*, 15(8):872–877.

Zhai, J., Kim, J., Knox, K. S., Twigg III, H. L., Zhou, H., and Zhou, J. J. (2018). Variance component selection with applications to microbiome taxonomic data. *Frontiers in Microbiology*, 9:509.

Zhan, X., Zhao, N., Plantinga, A., Thornton, T. A., Conneely, K. N., Epstein, M. P., and Wu, M. C. (2017). Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*, 206(4):1779–1790.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, H., Zhao, N., Mehrotra, D. V., and Shen, J. (2020). Composite Kernel Association Test (CKAT) for SNP-set Joint Assessment of Genotype and Genotype-by-treatment Interaction in Pharmacogenetics Studies. *Bioinformatics*. btaa125.

Zhao, N., Zhang, H., Clark, J. J., Maity, A., and Wu, M. C. (2019). Composite kernel machine regression based on likelihood ratio test for joint testing of genetic and gene–environment interaction effect. *Biometrics*, 75(2):625–637.

Zhou, H., Hu, L., Zhou, J., and Lange, K. (2019). MM algorithms for variance components models. *Journal of Computational and Graphical Statistics*, 28(2):350–361.

Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375.

Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The Annals of Applied Statistics*, 11(4):2027.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zoungas, S., Arima, H., Gerstein, H. C., Holman, R. R., Woodward, M., Reaven, P., Hayward, R. A., Craven, T., Coleman, R. L., Chalmers, J., et al. (2017). Effects of intensive glucose control on microvascular outcomes in patients with type 2 diabetes: a meta-analysis of individual participant data from randomised controlled trials. *The Lancet Diabetes & Endocrinology*, 5(6):431–437.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464.