UNIVERSITY OF CALIFORNIA

Los Angeles

Scalable and Robust Statistical Inference Algorithms for Linking Genotypes to Phenotypes

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Ali Pazokitoroudi

2023

ABSTRACT OF THE DISSERTATION

Scalable and Robust Statistical Inference Algorithms for Linking Genotypes to Phenotypes

by

Ali Pazokitoroudi
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2023
Professor Sriram Sankararaman, Chair

With the advancements in DNA sequencing technology and the decreasing cost of sequencing, there has been exponential growth in the amount of genomic data generated. This growth provides an unprecedented opportunity to access the genotypes of a large population, including millions of genetic variants, and to collect hundreds of thousands of phenotypic measurements from the same individuals. This opens doors to systematically studying the genetic architecture underlying complex traits and diseases. Genetic architecture refers broadly to a complete understanding of all genetic contributions to a given trait as well as to an awareness of the characteristics of this contribution.

During the past decade, variance components analysis has emerged as a robust statistical framework for investigating the genetic architectures of complex traits. To gain accurate and innovative insights into genetic architecture, applying flexible variance component models to large-scale datasets is crucial. However, fitting such models necessitates the use of scalable algorithms. Common approaches for estimating variance components involve searching for parameter values that maximize the likelihood or the restricted maximum likelihood (REML) [79]. Despite several algorithmic advancements [132, 63, 55, 61, 31, 70, 95], computing REML estimates of variance components on extensive datasets like the UK Biobank [4], which consists of approximately 500,000 genotyped individuals, millions of single nucleotide polymorphisms (SNPs), and hundreds of thousands of phenotypes, remains challenging. This thesis introduces a set of scalable and robust statistical inference algorithms rooted in variance component analysis. These algorithms are designed to estimate the variation in a trait that can be explained by linear and non-linear functions of the genotype, such as the interaction between alleles at a single genetic variant (dominance), the interaction between genetic variants (epistasis), and the interaction between environmental factors and genetic variants (GxE). Furthermore, these algorithms aim to estimate the distribution of these effects across the genome.

By applying our methods to the UK Biobank dataset, we uncover valuable insights into the genetic architecture of complex traits. Notable observations are as follows. First, we observe that both per-allele squared additive and GxE effect size increase with decreasing minor allele frequency (MAF) and linkage disequilibrium (LD). Second, testing whether GxE heritability is enriched around genes that are highly expressed in specific tissues, we find significant tissue-specific enrichments that include brain-specific enrichment for BMI and Basal Metabolic Rate in the context of smoking, adipose-specific enrichment for WHR in the context of sex, and cardiovascular tissue-specific enrichment for total cholesterol in the context of age. Third, we detect epistasis effects between SNPs located on the same chromosome and between SNPs located on different chromosomes. Fourth, our analyses indicate a limited contribution of dominance heritability to complex trait variation.

The dissertation of Ali Pazokitoroudi is approved.

Jason Ernst

Eleazar Eskin

Bogdan Pasaniuc

Sriram Sankararaman, Committee Chair

University of California, Los Angeles

2023

This thesis is dedicated to my parents and sister as a token of appreciation for their unconditional love, support, countless sacrifices, and numerous other things that, if detailed here, would fill more space than my entire Ph.D. work.

Table of Contents

List of Figures

Acknowledgments

Throughout my Ph.D. journey, I had the opportunity to interact with individuals whose contributions were essential for the completion of this thesis.

First and foremost, I extend my gratitude to my supervisor, Sriram Sankararaman, for his invaluable guidance and support, which played a pivotal role in shaping my research.

Secondly, I would like to thank my thesis committee members: Jason Ernst, Eleazar Eskin, and Bogdan Pasaniuc, for their valuable feedback. Additionally, I would like to acknowledge individuals who, while not officially part of my committee, provided helpful feedback and guidance, including Andrew Dahl, Eran Halperin, and Noah Zaitlen.

Thirdly, I am thankful to my collaborators and labmates for their input and feedback on various projects. Our collaborations were enjoyable, and I gained valuable insights from our joint works and discussions.

Finally, as the words of this thesis reach your eyes, I would like to thank you, the reader, for your interest in reading my thesis.

Curriculum Vitae

2009 – 2013	B.S. in Computer Science, University of Tehran, Tehran, Iran.
2013	Ranked 1st in the undergraduate class of 2013.
2015 – 2017	M.S. in Computer Science, Simon Fraser University, Vancouver, Canada.
2018– present	Ph.D. in Computer Science, University of California, Los Angeles, USA.
2023	Outstanding Graduate Student Research Award, Computer Science Department, UCLA.

Publications

A. Pazokitoroudi, Y. Wu, K. S. Burch, K. Hou, B. Pasaniuc, and S. Sankararaman, "Scalable multi-component linear mixed models with application to SNP heritability estimation", **RECOMB**, 2019.

A. Pazokitoroudi, Y. Wu, K. Burch, K. Hou, B. Pasaniuc, and S. Sankararaman, "Efficient variance components analysis across millions of genomes", **Nature Communications**, 2020.

A. Pazokitoroudi, A. M. Chiu, K. S. Burch, B. Pasaniuc, and S. Sankararaman, "Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data", **The American Journal of Human Genetics**, 2021.

A. Findley, A. Monziani, A. L. Richards, K. Rhodes, M. Ward, C. A. Kalita, A. Alazizi, **A. Pazokitoroudi**, S. Sankararaman, X. Wen, D. Lanfear, R. Pique-Regi, Y. Gilad, and F.

Luca, "Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions", **eLife**, 2021.

A. Pazokitoroudi, A. Dahl, N. Zaitlen, S. Rosset, and S. Sankararaman, "Biobank-scale estimation of the proportion of trait variance explained by gene-environment interactions", **RECOMB**, 2021.

S. Darabi, S. Fazeli, **A. Pazokitoroudi**, S. Sankararaman, and M. Sarrafzadeh, "Contrastive mixup: Self-and semi-supervised learning for tabular domain", **arXiv**, 2021.

W. Xinzhu*, R. Christopher*, **A. Pazokitoroudi**, A. Ganna, A. Gusev, A. Durvasula, S. Gazal, P.-R. Loh, D. Reich, and S. Sankararaman, "The lingering effects of Neanderthal introgression on human complex traits", **eLife**, 2023.

U. An, **A. Pazokitoroudi**, M. Alvarez, L. Huang, S. Bacanu, A. J. Schork, K. Kendler, P. Pajukanta, J. Flint, N. Zaitlen, N. Cai, A. Dahl, and S. Sankararaman, "Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries", **Nature Genetics**, 2023.

B. Fu, **A. Pazokitoroudi**, M. Sudarshan, L. Subramanian, and S. Sankararaman, "Fast kernel-based association testing of non-linear genetic effects for biobank-scale data", **Nature Communications**, 2023.

G. Kalantzis, **A. Pazokitoroudi**, H. Loya, H. Chen, S. Sankararaman, P. Palamara, "Scalable linear-mixed-model analysis of genotype and genealogical data", **Submitted to Bioinformatics**, 2023.

CHAPTER 1

Introduction

If we trace our biological history back in time, we can identify a point where we were a single cell called a zygote. This zygote was formed through the fusion of a father's sperm and a mother's egg cell at the moment of conception. The zygote carries a complete set of genetic material, including chromosomes inherited from both parents. Through subsequent divisions and cellular differentiation, the zygote develops into an embryo and gradually transforms into a fully formed organism. If we look at the inside of a zygote, it contains a total of 46 chromosomes. These chromosomes are organized into 23 pairs, with one set contributed by the father's sperm and the other set contributed by the mother's egg cell. The 23 pairs consist of one sex pair (either XX or XY) and 22 pairs of autosomes (non-sex chromosomes). The combination of chromosomes in the zygote determines the individual's genetic makeup and various inherited traits.

Each chromosome can be represented as a string of four letters, which correspond to the four nucleotide bases found in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). These letters represent the sequence of nucleotides along the DNA strand of a specific chromosome. The sequence of nucleotides along the DNA remains the same in all cells of an organism. This sequence contains the genetic information necessary for the synthesis of proteins and the regulation of various cellular processes. The differences between cells in different tissues or organs lie in the way they are spatially organized in three-dimensional space. This spatial arrangement helps determine which specific genes are expressed and active in a particular tissue or cell type, leading to the development and specialization of different cell types throughout the organism. Thus, while the genetic information remains constant, the regulation and expression of genes in specific tissues contribute to the diversity and functionality of different cell types in an organism.

When comparing the DNA of different individuals, there are regions within the genomes where there are variations or differences in the sequence of nucleotides. These variations are known as Single Nucleotide Polymorphisms (SNPs). SNPs represent a type of genetic variation where a single nucleotide (A, C, G, or T) at a specific position in the genome differs among individuals. Although SNPs are widespread in the human genome, they represent only a small portion of the entire genome. The human genome consists of billions of nucleotides, and SNPs occur at specific positions throughout this vast genetic code. While SNPs are relatively common, comprising millions of known variations, they still represent a small fraction of the entire genome. Understanding the distribution and effects of these SNPs is crucial for studying genetic diversity, disease susceptibility, and population genetics.

With the advancements in DNA sequencing technology and the decreasing cost of sequencing, there has been a significant and rapid expansion in the volume of genomic data being generated. This growth provides an opportunity to access the genotypes of a large population, encompassing millions of genetic variants, and to collect hundreds of thousands of phenotypic measurements from the same individuals.Researchers have tried to develop methods to link phenotypes to genotypes and environmental factors to understand the genetic architecture underlying complex traits and diseases. Genetic architecture refers broadly to a complete understanding of all genetic contributions to a given trait as well as to an awareness of the characteristics of this contribution and how it interacts with non-genetic factors.

Heritability is a fundamental concept used to assess the contribution of genetic variation to phenotypic variation. Broad-sense heritability encompasses all genetic factors that contribute to phenotypic differences, while Narrow-sense heritability specifically focuses on the additive genetic component, which refers to the genetic contribution that can be measured as a linear function of genotypes.

Once the heritability of a phenotype has been established, the next step is to understand how these genetic variations are distributed across the genome and how they contribute to the phenotype. This process is known as partitioning heritability. Partitioning heritability involves dissecting the overall genetic contribution to a trait or phenotype into different components, such as specific genes, genomic regions, pathways, or functional annotations. This analysis helps to identify which parts of the genome are responsible for the observed heritability and provides insights into the underlying genetic mechanisms. By partitioning heritability, researchers can gain a more detailed understanding of how genetic factors contribute to phenotypic variation, identify specific genes or genomic regions associated with the trait, and unravel the biological pathways and functional annotations that are relevant to the phenotype of interest.

The additive model assumes a linear relationship between genotypes and phenotypes. However, evidence suggests that genetic factors can interact with each other or with environmental factors, resulting in non-linear effects on phenotypes. To gain a deeper understanding of genetic architecture, it is crucial to extend our linear models to incorporate interaction terms. These interaction terms can include the interaction between alleles at a single genetic variant (referred to as dominance), the interaction between different genetic variants (referred to as epistasis), and the interaction between genetic variants and environmental factors (referred to as GxE). By incorporating these interaction terms into our models, we can better capture the complex interplay between genes and their environment, which can significantly influence phenotypic outcomes. Considering these non-linear effects and interactions allows for a more comprehensive understanding of the genetic architecture of complex traits and diseases.

Over the past decade, considerable progress has been made in developing computational and statistical tools to analyze genotype and phenotype data, enabling a deeper understanding of the genetic architecture of complex phenotypes. These tools encompass estimating heritability, partitioning heritability, quantifying the impact of non-linear effects (e.g., dominance, epistasis, GxE), and identifying causal SNPs. However, challenges arise from the nature of genetic data. Genetic data tend to be high-dimensional, containing a large number of SNPs that are often highly correlated. Moreover, the scale of the data can be substantial, requiring scalable and robust statistical methods to draw meaningful insights about the underlying models of phenotypes. To overcome these challenges, it is necessary to employ both scalable and robust statistical methods. Scalable approaches can efficiently handle the large-scale nature of genetics data, while robust statistical methods ensure reliable and accurate inferences from the data. By leveraging these methods, researchers can derive valuable insights and uncover the complex relationships between genetic variants and phenotypes.

Variance components analysis has emerged as a robust statistical framework for investigating the genetic architectures of complex traits. To gain accurate and innovative insights into genetic architecture, it is crucial to apply flexible variance component models to large-scale datasets. However, fitting such models necessitates the use of scalable algorithms. Common approaches for estimating variance components involve searching for parameter values that maximize the likelihood or the restricted maximum likelihood (REML) [79]. Despite several algorithmic advancements [132, 63, 55, 61, 31, 70, 95], computing REML estimates of variance components on extensive datasets like the UK Biobank [4], which consists of approximately 500,000 genotyped individuals, millions of single nucleotide polymorphisms (SNPs), and hundreds of thousands of phenotypes, remains challenging.

In this thesis, my collaborators and I focus on developing scalable and robust statistical inference algorithms within variance components analysis frameworks. We aim to address several important biological questions by analyzing large-scale genotype-phenotype data.

- Chapter 2: Our first objective is to quantify the amount of variation in phenotypes that can be explained by a linear function of genotypes. Specifically, we investigate the accuracy of the best linear predictor for phenotypes based on genotypes. Furthermore, we examine the distribution of this variation across the genomes in relation to functional annotations, genes, pathways, as well as minor allele frequency (MAF) and linkage disequilibrium (LD).

- Chapter 3: In this chapter, we explore whether the effect of an environmental variable on phenotype is influenced by the genetic background. We investigate the presence of a GxE (Gene-Environment interaction) effect. If such an effect is detected, we analyze how it is distributed across the genomes, considering tissue-specific genes, functional annotations, as well as MAF and LD.

- Chapter 4: Our focus in this chapter is to assess the contribution of the dominance effect, which is the interactions between alleles at a specific locus to phenotypic variations, particularly in comparison to additive effects.
- Chapter 5: In this chapter, we explore interactions between different SNPs that contribute to phenotypic variance. We investigate whether these SNPs interactions exist and find their locations across the genomes. Boyang Fu is primarily responsible for the real data analysis.

CHAPTER 2

Partitioning heritability across genome

2.1 Background

Variance components analysis [71] has emerged as a versatile tool in human complex trait genetics, enabling studies of the genetic contribution to variation in a trait [132] as well as its distribution across genomic loci [136, 63], allele frequencies [136], and functional annotations [136, 57, 34]. There is increasing interest in applying methods for variance components analysis to large-scale genetic datasets with the goal of uncovering novel insights into the genetic architecture of complex traits[21, 63]. A prominent example of the utility of these methods is in the estimation of SNP heritability (h^2_{SNP}) [132], the variance in a trait explained by a given set of genotyped SNPs. Variance components methods for estimating SNP heritability typically assume a genetic variance component that represents the fraction of phenotypic variation explained by the SNPs included in the study and a residual variance component. Recent studies have shown that these "single-component" methods yield biased estimates of SNP heritability due to the LD and MAF dependent architecture of complex traits [18, 29]. On the other hand, flexible models with multiple variance components [136, 63] that allows for SNP effects to vary with MAF and LD, have been shown to yield more accurate SNP heritability estimates [18, 29]. Recent work has shown that SNP heritability can be estimated with minimal assumptions about the genetic architecture [44]; however, this method cannot partition heritability across categories of SNPs of interest such as functional or population genomic annotations. Partitioning heritability requires fitting multiple variance components, thus creating the need for accurate and scalable methods that can fit tens or even hundreds of variance components to large-scale genomic data to obtain accurate and novel insights into genetic architecture.

While the ability to fit flexible variance component models to large-scale datasets is essential to obtain accurate and novel insights into genetic architecture, fitting such models requires scalable algorithms. Approaches for estimating variance components typically search for parameter values that maximize the likelihood or the restricted maximum likelihood (REML) [79]. Despite a number of algorithmic improvements [132, 63, 55, 61, 31, 70, 95], computing REML estimates of the variance components on data sets such as the UK Biobank [4] ($\approx 500,000$ individuals genotyped at nearly one million SNPs) remains challenging. The reason is that methods for computing these estimators typically perform repeated computations on the input genotypes.

We propose a method that can jointly estimate multiple variance components efficiently. Our proposed method, RHE-mc, is a randomized multi-component version of the classical Haseman-Elston regression for heritability estimation [37, 145]. RHE-mc builds on our previously proposed method, RHE-reg [129], which uses a randomized algorithm to estimate a single variance component. RHE-mc can simultaneously estimate multiple variance components as well as estimate variance components associated with continuous annotations and overlapping annotations. Unlike RHE-reg, RHE-mc uses a nonparametric block jackknife to estimate standard errors with little computational overhead. Further, unlike REML estimation algorithms, RHE-mc requires only a single pass over the input genotypes that results in a highly memory efficient implementation. The resulting computational efficiency permits RHE-mc to jointly fit 300 variance components in less than an hour on a dataset of about $300,000$ individuals and $500,000$ SNPs, about two orders of magnitude faster than state-of-the-art methods. On a dataset of one million individuals and one million SNPs, RHE-mc can fit 100 variance components in about 12 hours.

To demonstrate its utility, we first show that RHE-mc can accurately estimate genome-wide and partitioned SNP heritability under realistic genetic architectures (the functional dependence of SNP effect sizes on MAF and LD). We applied RHE-mc to 22 traits measured across $291,273$ individuals genotyped at $459,792$ common SNPs (MAF$> 1\%$) in the UK

Biobank to obtain estimates of genome-wide SNP heritability. We then used RHE-mc to partition heritability for the 22 traits across seven million imputed SNPs (MAF $> 0.1\%$) into 144 bins defined based on MAF and LD. We observe that the allelic effect size tends to increase with lower MAF and LD across the traits considered. Finally, we partitioned heritability for SNPs with MAF $> 0.1\%$ across 28 functional annotations. We recover previously reported enrichment of heritability in annotations corresponding to conserved regions [21] and also document enrichment of heritability in FANTOM5 enhancers in eczema, asthma, autoimmune disorders, and thyroid disorders.

2.2 Materials and Methods

We aim to fit a variance components model that relates phenotypes $\boldsymbol{y}$ measured across N individuals to their genotypes over M SNPs $\boldsymbol{X}$:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K &= \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{D}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N) \\
\boldsymbol{\beta}_k &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_k^2}{M_k} \boldsymbol{I}_{M_k}), k \in \{1, \ldots, K\}
\end{aligned}$$

where $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ is an arbitrary distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$. Each of the M SNPs is assigned to one of K non-overlapping categories so that $\boldsymbol{X}_k$ is the $N \times M_k$ matrix consisting of standardized genotypes of SNPs belonging to category k (note that the expected heritability is constant within categories when we use standardized genotypes). $\boldsymbol{\beta}_k$ denotes the effect sizes of SNPs assigned to category k which are drawn from a zero-mean normal distribution with variance parameter σ_k^2 (the variance component of category k) while σ_e^2 is the residual variance.

In this model, the genome-wide SNP heritability is defined as: $h_{SNP}^2 = \frac{\sum_{k=1}^{K} \sigma_k^2}{\sum_{k=1}^{K} \sigma_k^2 + \sigma_e^2}$ while the SNP heritability of category k is defined as: $h_k^2 = \frac{\sigma_k^2}{\sum_{k=1}^{K} \sigma_k^2 + \sigma_e^2}$. By choosing categories to represent genomic annotations of interest, *e.g.*, chromosomes, allele frequencies, and functional annotations, these models can be used to estimate the phenotypic variation that can be attributed to the relevant annotation.

The key inference problem in this model is the estimation of the variance components: $(\sigma_1^2, \dots, \sigma_K^2, \sigma_e^2)$. These parameters are typically estimated by maximizing the likelihood or the restricted likelihood. Instead, RHE-mc uses a scalable method-of-moments estimator, *i.e.*, finding values of the variance components such that the population moments match the sample moments [37, 30, 117, 33, 145]. RHE-mc uses a randomized algorithm that avoids explicitly computing $N \times N$ genetic relatedness matrices that are required by method-of-moments estimators. Instead, it operates on a smaller matrix formed by multiplying the input genotype matrix with a small number of random vectors (Methods). The application of a randomized algorithm for SNP heritability estimation using a single variance component was proposed in our previous work, RHE-reg [129]. RHE-mc extends our previous work in several directions. RHE-mc can efficiently fit multiple variance components (both non-overlapping and overlapping) and can also handle continuous annotations. The resulting algorithm has scalable runtime as it only requires operating on the genotype matrix one time. Further, RHE-mc uses a streaming implementation that does not require all the genotypes to be stored in memory leading to scalable memory requirements (Supplementary Notes). Finally, RHE-mc uses an efficient implementation of a block Jackknife to estimate standard errors with little computational overhead (Supplementary Notes).

2.2.1 Multi-component Linear mixed model

RHE-mc attempts to fit the following variance components model:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K &= \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{D}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N) \\ \boldsymbol{\beta}_k &\sim \mathcal{D}(\mathbf{0}, \frac{\sigma_k^2}{M_k} \boldsymbol{I}_{M_k}), k \in \{1, \dots, K\} \end{aligned} \tag{2.1}$$

Here $\boldsymbol{y}$ is a N-vector of centered phenotypes. Here each of the M SNPs is assigned to one of K non-overlapping categories. Each category k contains M_k SNPs, $k \in \{1, \dots, K\}$, $\sum_k M_k = M$. Let $\boldsymbol{X}_k$ be a $N \times M_k$ matrix where $x_{k,n,m}$ denotes the standardized genotype for individual n at SNP m in category k. We have $\sum_n x_{k,n,m} = 0$ and $\sum_n x_{k,n,m}^2 = N$ for $m \in \{1, 2, \ldots, M_k\}$. Let $\boldsymbol{\beta}_k$ be a M_k-vector of SNP effect sizes for the k-th category. In the above model, σ_e^2 is the residual variance, and σ_k^2 is the variance component of the k-th category. In this model, the total SNP heritability is defined as :

$$h_{SNP}^2 = \frac{\sum_{k=1}^K \sigma_k^2}{\sum_{k=1}^K \sigma_k^2 + \sigma_e^2} \tag{2.2}$$

The SNP heritability of category k is defined as:

$$h_k^2 = \frac{\sigma_k^2}{\sum_{k=1}^K \sigma_k^2 + \sigma_e^2}, k \in \{1, \ldots, K\} \tag{2.3}$$

Enrichment in bin k is defined as:

$$e_k = \frac{h_k^2/h_{SNP}^2}{M_k/M}, k \in \{1, \ldots, K\} \tag{2.4}$$

2.2.2 Method-of-moments for estimating multiple variance components

To estimate the variance components, RHE-mc uses a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments [41]. Since $\mathbb{E}[\boldsymbol{y}] = 0$, we derived the MoM estimates by equating the population covariance to the empirical covariance. The population covariance is given by:

$$cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sum_k \sigma_k^2 \boldsymbol{K}_k + \sigma_e^2 \boldsymbol{I}_N \tag{2.5}$$

Here $\boldsymbol{K}_k = \frac{\boldsymbol{X}_k \boldsymbol{X}_k^T}{M_k}$ is the genetic relatedness matrix (GRM) computed from all SNPs of k-th category. Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to find the variance components.

$$(\tilde{\sigma_1^2}, \ldots, \tilde{\sigma_K^2}, \tilde{\sigma_e^2}) = argmin_{(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2)} ||\boldsymbol{y}\boldsymbol{y}^T - (\sum_k \sigma_k^2 \boldsymbol{K}_k + \sigma_e^2 \boldsymbol{I})||_F^2 \tag{2.6}$$

The MoM estimator satisfies the following normal equations:

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{b} \\ \boldsymbol{b}^T & N \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\sigma_g^2}} \\ \tilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{y}^T \boldsymbol{y} \end{bmatrix} \tag{2.7}$$

Here $\tilde{\boldsymbol{\sigma}_{\boldsymbol{g}}^{\mathbf{2}}} = \begin{bmatrix} \tilde{\sigma_1^2} \\ \vdots \\ \tilde{\sigma_K^2} \end{bmatrix}$, $\boldsymbol{T}$ is a $K \times K$ matrix with entries $T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l), k, l \in \{1, \ldots, K\}$, $\boldsymbol{b}$ is a K-vector with entries $b_k = tr(\boldsymbol{K}_k) = N$ (because $\boldsymbol{X}_k$s is standardized), and $\boldsymbol{c}$ is a K-vector with entries $c_k = \boldsymbol{y}^T \boldsymbol{K}_k \boldsymbol{y}$. Each GRM $\boldsymbol{K}_k$ can be computed in time $\mathcal{O}(N^2 M_k)$ and $\mathcal{O}(N^2)$ memory. Given K GRMs, the quantities $T_{k,l}$, c_k, $k, l \in \{1, \ldots, K\}$, can be computed in $\mathcal{O}(K^2 N^2)$. Given the quantities $T_{k,l}$, c_k, the normal Equation (2.7) can be solved in $\mathcal{O}(K^3)$. Therefore, the total time complexity for estimating the variance components is $\mathcal{O}(N^2 M + K^2 N^2 + K^3)$.

2.2.3 Randomized estimator of multiple variance components

The key bottleneck in solving the normal Equation (2.7) is the computation of $T_{k,l}$, $k, l \in \{1, \ldots, K\}$ which takes $\mathcal{O}(N^2 M)$. Instead of computing the exact value of $T_{k,l}$, we use an unbiased estimator of the trace [48] based on the following identity: for a given $N \times N$ matrix $\boldsymbol{C}$, $\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z}$ is an unbiased estimator of $tr(\boldsymbol{C})$ ($E[\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z}] = tr[\boldsymbol{C}]$) where $\boldsymbol{z}$ be a random vector with mean zero and covariance $\boldsymbol{I}_N$. Hence, we can estimate the values $T_{k,l}$, $k, l \in \{1, \ldots, K\}$ as follows:

$$T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l) \approx \widehat{T_{k,l}} = \frac{1}{B} \frac{1}{M_k M_l} \sum_b \boldsymbol{z}_b^T \boldsymbol{X}_k \boldsymbol{X}_k^T \boldsymbol{X}_l \boldsymbol{X}_l^T \boldsymbol{z}_b \tag{2.8}$$

Here $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$ are B independent random vectors with zero mean and covariance $\boldsymbol{I}_N$. We draw these random vectors independently from a standard normal distribution. Computing $T_{k,l}$ using the unbiased estimator involves four multiplications of sub-matrices of the genotype matrix with a vector, repeated B times. Therefore, the total running time for estimating the matrix $\boldsymbol{T}$ is $\mathcal{O}(NMB + K^2 NB)$.

Moreover, we can leverage the structure of the genotype matrix which only contains entries in $\{0, 1, 2\}$. For a fixed genotype matrix $\boldsymbol{X}_k$, we can improve the per iteration time complexity of matrix-vector multiplication from $\mathcal{O}(NM)$ to $\mathcal{O}(\frac{NM}{max(\log_3 N, \log_3 M)})$ by using the Mailman algorithm [59]. Solving the normal equations takes $\mathcal{O}(K^3)$ time so that the overall time complexity of our algorithm is $\mathcal{O}(\frac{NMB}{\max(\log_3(N), \log_3(M))} + K^2(K + NB))$.

RHE-mc uses a block Jackknife to estimate standard errors. In Supplementary Notes, we show how the block Jackknife estimates can be computed with little additional computational overhead. Further, we also show how covariates can be efficiently included in the model (Supplementary Notes).

2.2.4 Multi-component LMM with overlapping annotations

RHE-mc can also be applied in the setting where annotations overlap. Following [21], the heritability of SNPs belong to annotation k is defined as:

$$h_k^2 = \frac{\sum_{i \in S_k} \sum_{j : i \in S_j} \frac{\sigma_j^2}{M_j}}{\sum_{k=1}^{K} \sigma_k^2 + \sigma_e^2}, k \in \{1, \ldots, K\} \tag{2.9}$$

where S_k is the set of SNPs in k-th annotation and $M_k = |S_k|$. Enrichment in bin k is defined as $e_k = \frac{h_k^2/h_{SNP}^2}{M_k/M}$.

2.2.5 Multi-component LMM with continuous annotations

We have described the derivation of RHE-mc using binary annotations. Following [28], we can extend RHE-mc to support continuous-value annotations as follows :

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K &= \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{D}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N) \\
\boldsymbol{\beta}_k &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_k^2}{M_k} diag(\boldsymbol{a_k})), k \in \{1, \ldots, K\}
\end{aligned} \tag{2.10}$$

this model is totally similar to the model in Equation (4.1) except that here we assume that the variances of effect sizes depend on continuous-valued annotation. Let $\boldsymbol{a_k}$ be a M_k-vector where $\boldsymbol{a_{k,i}}$ is the value of k-th annotation at SNP i (the elements of $\boldsymbol{a_k}$ must be non-negative). Let S_k be the set of SNPs belong to annotation k. In this model, the SNP heritability of annotation k is defined as :

$$h_k^2 = \frac{\sum_{i \in S_k} \frac{\sigma_k^2}{M_k} a_{k,i}}{\sum_{k=1}^{K} \sum_{i \in S_k} \frac{\sigma_k^2}{M_k} a_{k,i} + \sigma_e^2}, k \in \{1, \ldots, K\} \tag{2.11}$$

To estimate the variance components of this new model, we only need to replace $\boldsymbol{X_k}$ with $\boldsymbol{X}_k diag(\sqrt{\boldsymbol{a_k}})$ in the Equation (2.5) for every annotation k. We assessed the accuracy of RHE-mc in estimating variance components with continuous annotation in Supplementary Notes.

2.3 Results

2.3.1 Simulations

We performed simulations to compare the performance of RHE-mc with several state-of-the-art methods for heritability estimation that cover the spectrum of methods that have been proposed.

We considered two simulation settings. In the large-scale simulation setting, we simulated phenotypes for the full set of UK Biobank genotypes consisting of $M = 593,300$ array SNPs and $N = 337,205$ individuals. We obtained the individuals by keeping unrelated white British individuals which are $> 3^{rd}$ degree relatives (defined as pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$)[4], and removing individuals with putative sex chromosome aneuploidy. The small-scale setting was designed so that we could compare the accuracies of RHE-mc to REML methods. In this setting, we simulated phenotypes from a subsampled set of genotypes from the UK Biobank data genotypes used in large scale simulation [112]. Specifically, we chose randomly a subset of $N = 10,000$ individuals from the large scale data. Therefore, in small scale, we have $M = 593,300$ array SNPs and $N = 10,000$ individuals. We simulated phenotypes from genotypes using the following model which is used in [44, 18]:

$$\begin{aligned}
\sigma_m^2 &= Sc_m w_m^b [f_m(1-f_m)]^a \\
(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, .., \boldsymbol{\beta}_m)^T &\sim \mathcal{N}(\boldsymbol{0}, diag(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2)) \\
y|\boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{X\beta}, (1-h^2)\boldsymbol{I}_N) \qquad (2.12)
\end{aligned}$$

where S is a normalizing constant chosen so that $\sum_{m=1}^{M} \sigma_m^2 = h^2$. Here $h^2 \in [0,1]$, $a \in \{0, 0.75\}$,$b \in \{0,1\}$, β_m, f_m and w_m are the effect size, the minor allele frequency and

LDAK score of m^{th} SNP respectively. Let $c_m \in \{0, 1\}$ be an indicator variable for the causal status of SNP m .The LD score of a SNP is defined to be the sum of the squared correlation of the SNP with all other SNPs that lie within a specific distance, and the LDAK score of a SNP is computed based on local levels of LD such that the LDAK score tends to be higher for SNPs in regions of low LD [109]. The above models relating genotype to phenotype are commonly used in methods for estimating SNP heritability: the GCTA Model (when $a = b = 0$ in Equation 2.12), which is used by the software GCTA [134] and LD Score regression (LDSC) [3], and the LDAK Model (where $a = 0.75, b = 1$ in Equation 2.12) used by software LDAK [109]. Moreover, under each model, we varied the proportion and minor allele frequency (MAF) of causal variants (CVs). Proportion of causal variants were set to be either 100% or 1%, and MAF of causal variants drawn uniformly from $[0, 0.5]$ or $[0.01, 0.05]$ or $[0.05, 0.5]$ to consider genetic architectures that are either infinitesimal or sparse as well genetic architectures that include a mixture of common and rare SNPs as well as one that includes only common SNPs. The true heritability were chosen from $\{0.1, 0.25, 0.5, 0.8\}$

We generated 100 sets of simulated phenotypes for each setting of parameters and report accuracies averaged over these 100 sets.

2.3.2 Comparisons

For the large-scale simulations, we compared RHE-mc to methods that rely on summary statistics for estimating heritability. Among the summary statistic methods, LD score regression (LDSC) [3] uses the slope from the GWAS χ^2 statistics regressed on the LD scores to estimate heritability. Stratified LD score regression (S-LDSC) [21] is an extension of LDSC for partitioning heritability from summary statistics. SumHer is the summary statistic analog of LDAK [107]. We ran S-LDSC with 10 binary MAF bin annotations defined such that each bin contains exactly 10% of the typed SNPs; this is intended to mirror the 10 MAF bin annotations in the S-LDSC "baseline-LD model" [28] (see Supplementary Table 5). To run SumHer, we used the LDAK software to compute the default "LDAK weights" using in-sample LD [109, 108, 107]. We then computed "LD tagging" using 1-Mb windows centered on each SNP as recommended [107]. To do a fair comparison we computed LD scores for LDSC, S-LDSC, GRE, and SumHer by using in-sample LD among the M SNPs, and in all simulations we aim to estimate the SNP-heritability explained by the same set of M SNP. We described the parameter settings of summary statistic methods in Supplementary Notes.

For the small-scale simulations, we compared RHE-mc to GCTA-mc and HE-mc [134]. GCTA-mc and HE-mc are the extensions of GCTA and HE to a multi-component LMM respectively where the variance components are typically defined by binning SNPs according to their MAF as well as local LD [18]. We ran both GCTA-mc and RHE-mc using 24 bins formed by the combination of 6 bins based on MAF (MAF$\leq$ 0.01,0.01 <MAF$\leq$ 0.02,0.02 <MAF$\leq$ 0.03,0.03 <MAF$\leq$ 0.4,0.04 <MAF$\leq$ 0.05,MAF$>$ 0.05) as well as 4 bins based on quartiles of the LDAK score of a SNP. We ran both GCTA-mc and RHE-mc allowing for estimates of a variance component to be negative.

For comparisons of runtime, we compared RHE-mc to GCTA [134] and BOLT-REML [63] which is a computationally efficient approximate method to compute the REML estimator. We ran all methods with 22 components (one for each chromosome). We also ran RHE-mc with $\approx$ 300 components (corresponding to 10 Mb bins) on the UK Biobank genotype (Supplementary Figure 10). To create our largest dataset, we replicate individuals from the UK Biobank and a subset of the imputed SNPs to obtain a dataset with one million individuals and SNPs. We use the latest versions of BOLT-REML (Version 2.3.2) and GCTA (Version 1.92.1) in our comparison. All comparisons are performed on an Intel(R) Xeon(R) CPU 2.10 GHz server with 128 GB RAM.

2.3.3 Heritability estimates in the UK Biobank

We estimated SNP-heritability for 22 real complex traits (6 quantitative, 16 binary) in the UK Biobank [4]. In this study, we restricted our analysis to SNPs that were present in the UK Biobank Axiom array used to genotype the UK Biobank. SNPs with greater than 1% missingness and minor allele frequency smaller than 1% were removed. Moreover, SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7} were removed. We restricted our study to self-reported British white ancestry individuals which are $> 3^{rd}$ degree relatives that is defined as pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$ [4]. Furthermore, we removed individuals who are outliers for genotype heterozygosity and/or missingness. Finally we obtained a set of $N = 291,273$ individuals and $M = 459,792$ SNPs to use in the real data analyses. We included age, sex, and the top 20 genetic principal components (PCs) as covariates in our analysis for all traits. We used PCs precomputed by the UK Biobank from a superset of $488,295$ individuals. Additional covariates were used for waist-to-hip ratio (adjusted for BMI) and diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy, and oral contraceptives).

2.3.4 Heritability partitioning

In our initial analysis, we removed SNPs with greater than 1% missingness and minor allele frequency smaller than 1%. Moreover, we removed SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7} as well as SNPs that lie within the MHC region (Chr6: 25–35 Mb) to obtain $4,824,392$ SNPs. We restricted our study to individuals self-reported British white ancestry individuals which are $> 3^{rd}$ degree relatives that is defined as pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$ [4]. Furthermore, we removed individuals who are outliers for genotype heterozygosity and/or missingness. Finally, we obtained $291,273$ individuals. We partitioned SNPs into eight bins based on two MAF bins (MAF≤ 0.05, MAF> 0.05) and quartiles of the LD-scores. For each bin k, we computed the heritability enrichment as the ratio of the percentage of heritability explained by SNPs in bin k to the the percentage of SNPs in bin k.

We considered an additional analysis in which we included SNPs with MAF $> 0.1\%$ resulting in $N = 291,273$ unrelated white British individuals and $M = 7,774,235$ imputed SNPs (MAF $> 0.1\%$). We defined 144 bins based on 4 LD bins and 36 MAF bins. The four LD bins are defined based on quartile of LD-scores, and 36 MAF bins are defined based on 9-quantile of the following four intervals: $0.001 \leq$MAF≤ 0.01, $0.01 <$MAF≤ 0.05,

$0.05 \leq$MAF≤ 0.10 , $0.10 <$MAF≤ 0.50.

2.3.5 Accuracy of genome-wide SNP heritability estimates in simulations

We assessed the accuracy of RHE-mc in estimating genome-wide SNP heritability as previous attempts at estimating SNP heritability have been shown to be sensitive to assumptions about how SNP effect size varies with MAF and LD[18]. Starting with genotypes of $M = 593,300$ array SNPs over $N = 337,205$ unrelated white British individuals in the UK Biobank, we simulated phenotypes according to 64 MAF and LD-dependent architectures by varying the SNP heritability, the proportion of variants that have non-zero effects (causal variants or CVs), the distribution of causal variants across minor allele frequencies (CVs distributed across all minor allele frequency bins or CVs restricted to either common or low-frequency bins), and the form of coupling between the SNP effect size and MAF as well as LD. For RHE-mc, we partitioned the SNPs into 24 variance components based on 6 MAF bins as well as 4 LD bins . The key parameter in applying RHE-mc is the number of random vectors B which we set to 10. RHE-mc estimates were relatively insensitive when we increased the number of random vectors B to 100 (Supplementary Figures 1 and 2, Supplementary Table 1)). Across these 64 architectures, RHE-mc is relatively unbiased (a two-sided t-test of the hypothesis of no bias is not rejected across any of the architectures at a p-value < 0.05) with the largest relative bias observed to be 0.5% of the true SNP heritability (Supplementary Figure 3). We used a block Jackknife (number of blocks = 100) to estimate the standard errors of RHE-mc and confirmed that the estimated standard errors are close to the true SE (Supplementary Table 2).

We compared the accuracy of RHE-mc to state-of-the-art methods for heritability estimation that can be applied to large datasets (across architectures where the true SNP heritability was fixed at 0.5). These methods, LDSC [3], SumHer [107], S-LDSC [22], and GRE [44], all leverage summary statistics while RHE-mc requires individual genotype data. We found that estimates from the summary-statistic methods tend to be sensitive to the underlying genetic architecture: across 16 architecture relative biases range from -31% to

27% for LDSC, -27% to 5% for S-LDSC, and -5% to 9% for SumHer (Figure 2.1). We also compared to a recently proposed method (GRE [44]) that only estimates genome-wide SNP heritability (without partitioning by MAF/LD) and observed that relative biases ranged from 1% to 1.4% for GRE and from -1.5% to 0.5% for RHE-mc. We also considered architectures in which only rare variants are causal and found RHE-mc is accurate relative to other methods (Supplementary Figure 4). These results further emphasize that RHE-mc can accurately estimate SNP-heritability through fitting multiple variance components.

We compared RHE-mc to the state-of-the-art REML-based variance component estimation method, GCTA-mc (multi-component GREML [134, 18, 131]) and to exact multi-component Haseman-Elston Regression (HE-mc) as implemented in GCTA[134]. We ran each of these methods by partitioning SNPs into 24 variance components (6 MAF bins by 4 LD bins, see Methods). To make these experiments computationally feasible, we simulated phenotypes starting from a smaller set of genotypes ($M = 593,300$ array SNPs and $N = 10,000$ white British individuals). Across 16 architectures where the true SNP heritability was fixed at 0.25, the relative biases for RHE-mc range from -3.2% to 3.6%, and from -3.2% to 5% for GCTA-mc (Figure 2.2). On average, RHE-mc has standard errors that are 1.1 times larger than GCTA-mc (which range from 0.97 to 1.24) and 1.08 times larger than HE-mc (which range from 1.00 to 1.21).

2.3.6 Accuracy of heritability partitioning in simulations

We also evaluated the accuracy of RHE-mc in partitioning SNP heritability in both small-scale ($M = 593,300$ SNPs, $N = 10,000$ individuals) (Supplementary Figure 5) and large-scale settings ($M = 593,300$ SNPs, $N = 337,205$ individuals) (see Supplementary Figure 6). For these experiments, we restrict our attention to architectures for which the causal variants (CVs) are chosen to lie within a narrow range of MAF. Since the variance components correspond to bins of MAF and LD, a subset of the variance components would have no causal SNPs and hence have a heritability of zero. We assess the accuracy of estimates of heritability aggregated over these components (termed the non-causal bin) as well as the heritability aggregated over the remaining genetic components (termed the causal bin). For example, variance components that correspond to MAF $\in [0.01, 0.05]$ would be included in the causal bin for an architecture that restricts the MAF of CVs to lie in the range $[0.01, 0.05]$. For the small-scale simulations, we compared RHE-mc to GCTA-mc. We ran both methods by partitioning the SNPs into 24 variance components based on 6 MAF bins as well as 4 LD bins defined by quartiles of the measure of LDAK weight at a SNP . Across the genetic architectures tested, estimates of heritability within each of the causal and non-causal bins are highly concordant between RHE-mc and GCTA-mc (Supplementary Figure 5, Supplementary Table 3): for the causal bin, the relative bias ranges from -4% to 0.4% for RHE-mc and -3.6% to 2% for GCTA-mc while, for the non-causal bin, the bias ranges from 0 to 0.7% for RHE-mc and 0 to 1.4% for GCTA-mc (Supplementary Table 3). For the large-scale settings, RHE-mc remains accurate: the relative bias ranges from -2.6% to 3.2% (causal bin) and the bias ranges from -0.5% to 0.2% (non-causal bin) over the genetic architectures considered (Supplementary Figure 6, Supplementary Table 4).

Heritability partitioning has been used to estimate heritability attributed to functional genomic annotations [21]. However, some of these annotations (such as FANTOM5 enhancers) are quite small covering $< 1\%$ of the genome. We explored the ability of RHE-mc to accurately estimate heritability as a function of the size of the annotation. To this end, we performed simulations using $N = 291,273$ unrelated white British individuals and $M = 459,792$ common SNPs. We defined 8 annotations (4 MAF bins and 2 LD bins) in which we fixed the enrichment of a selected bin and varied the proportion of SNPs in the selected category. RHE-mc obtained accurate estimates of enrichment even when the selected bin only contained 0.4% of the genome-wide SNPs (comparable to the size of FANTOM5 enhancers). RHE-mc estimates are well-calibrated: when the bin has zero enrichment, RHE-mc rejected the null hypothesis of no enrichment in 5% of the simulations while attaining high power to reject the null hypothesis even when the bin contained $< 1\%$ of the SNPs (Supplementary Notes).

2.3.7 Computational efficiency

We benchmarked the runtime and memory usage of RHE-mc as a function of number of individuals, SNPs and variance components (Figure 2.3, Table 2.1). We ran RHE-mc with $B = 10$ random vectors and 22 variance components where each chromosome forms a distinct component. On a dataset of $\approx 300,000$ individuals and $\approx 500,000$ SNPs, RHE-mc can fit 22 variance components in less than an hour and ≈ 300 variance components (corresponding to bins of size 10 Mb) with little increase in its runtime. On a dataset of one million individuals and one million SNPs, RHE-mc can fit 100 variance components in a few hours. Further, due to its use of a streaming implementation that only requires the genotypes to be operated on once, the memory requirement of RHE-mc is modest: all experiments required less than 60 GB. We compared the run time and memory usage of RHE-mc with REML-based methods (GCTA [134] and BOLT-REML [63]) on the UK Biobank genotypes consisting of around $500,000$ SNPs over varying sample sizes and observed that RHE-mc achieves several orders-of-magnitude reduction in runtime. Summary-statistic methods such as S-LDSC requires pre-computed inputs which depend on the runtimes of other softwares making a direct comparison of speed difficult. Thus, we have restricted our comparison to individual-level methods where the benchmarking can be done in a comparable manner.

2.3.8 Estimating total SNP heritability in the UK Biobank

We applied RHE-mc to estimate genome-wide SNP heritability for 22 complex traits (6 quantitative and 16 binary traits) measured in the UK Biobank . We analyzed $N = 291,273$ unrelated white British individuals and $M = 459,792$ SNPs genotyped on the UK Biobank Axiom array (Methods). We ran RHE-mc with $B = 10$ and with SNPs divided into eight bins based on two MAF bins ($0.01 \leq$MAF< 0.05, MAF≥ 0.05) and quartiles of the LD-scores. We compared the estimates from RHE-mc to those from LDSC, S-LDSC, SumHer, and GRE. Restricting our analysis to 18 traits for which the point estimate of genome-wide SNP heritability from RHE-mc is > 0.05, the estimates from S-LDSC, GRE, SumHer and LDSC were on average 2.5%, 10%, 25%, and 67% higher than RHE-mc (Figure 2.4). Relative to the simulation results, the estimates from S-LDSC are generally consistent with those from RHE-mc. This is likely due to the fact that, in simulations, our application of S-LDSC used only MAF bins. On the other hand, in real data, we used S-LDSC with the recommended baseline-LD annotations (including functional annotations).

We then applied RHE-mc to estimate genome-wide heritability attributable to imputed variants. The genome-wide estimates of SNP heritability from RHE-mc on imputed SNPs (MAF$>$ 1%) are concordant with the estimates from array SNPs (2.8% higher on average). We then analyzed $M = 7,774,235$ imputed genotypes with MAF $> 0.1\%$ using 144 bins formed by 4 LD bins and 36 MAF bins (Methods). Genome-wide SNP heritability estimates from RHE-mc on imputed SNPs (MAF$>$ 0.1%) are 11.4% higher than RHE-mc on imputed SNPs (MAF$>$ 1%). (Figure 2.4, Supplementary Figure 7). Following previous work [44], we have removed the MHC region to enable a systematic comparison since the estimation of LD in the MHC region can be challenging; it would be of interest to compare methods when the MHC is included.

2.3.9 Partitioning SNP heritability across allele frequency and LD bins

We used RHE-mc to partition SNP heritability of 22 complex traits across MAF and LD bins. We analyzed $M = 7,774,235$ imputed SNPs with MAF $> 0.1\%$. We used 144 bins formed by 4 LD bins and 36 MAF bins . We compute the allelic effect size of SNPs in bin k as $\frac{h_k^2}{2f_k(1-f_k)M_k}$ where h_k^2 is the heritability estimated in bin k, f_k is the mean MAF in bin k, and M_k is the number of SNPs in bin k. We observe that allelic effect size increases with lower MAF and LD. For height, in the lowest quartile of LD scores, SNPs with MAF $\approx 0.1\%$ have allelic effect sizes $\approx 27x \pm 8$ larger than SNPs with MAF $\approx 50\%$. Similarly, among SNPs with MAF $\approx 50\%$, SNPs in the lowest quartile of LD scores have allelic effect sizes $\approx 5x \pm 1$ larger than SNPs in the highest quartile (Figure 2.5 for height; other traits in Supplementary Figure 9). While these trends have been observed in previous studies [28, 29, 120], the ability of RHE-mc to jointly fit multiple variance components allows us to estimate effect sizes at SNPs with MAF as low as 0.1%. We caution that negative heritability estimates in bins of lowest MAF and high LD score could be arise due to one or more of the following factors: low number of SNPs in this bin (we did not constrain our variance components estimates to be non-negative), the inadequacy of the assumed heritability model, and errors in the imputed genotypes used for the analysis.

2.3.10 Partitioning heritability by functional annotations

The ability of RHE-mc to estimate variance components associated with a large number of overlapping annotations enables us to explore the contribution a variety of functional genomic annotations to trait heritability using individual-level data in the UK Biobank. We applied RHE-mc to jointly partition heritability of 22 complex traits across 28 functional annotations as defined in [21]. We restricted our analysis to $N = 291,273$ unrelated white British individuals and $M = 5,670,959$ imputed SNPs (we restrict to SNPs with MAF $> 0.1\%$ which are also present in 1000 Genomes Project). We grouped the traits into five categories (autoimmune, diabetes, respiratory, anthropometric, cardiovascular); for a representative trait from each category, we report enrichment of each of the 28 functional annotations in Figure 2.6 (see Methods; for all traits see Supplementary Figure 8). Our results are largely concordant with previous studies [21, 29]: we observe enrichment of heritability across traits in conserved regions (Z-score > 3 in 15 traits). We also observe an enrichment of heritability at FANTOM5 enhancers (labeled Enhancer_Andersson in Figure 2.6) in asthma, eczema, autoimmune disorders (broad), hypothyroidism, and thyroid disorders (Z-score > 3) even though these annotations cover only 0.4% of the analyzed SNPs .

2.4 Discussion

We have presented RHE-mc, an algorithm that can efficiently estimate multiple variance components on large-scale genotype data. In light of increasing evidence for SNP effect sizes that vary as a function of covariates such as MAF and LD and the bias associated with methods that fit only a single variance component [18], the ability to define flexible models endowed with multiple variance components is important to obtain unbiased estimates of fundamental quantities such as SNP heritability. We confirm that RHE-mc yields accurate genome-wide SNP heritability estimates under diverse genetic architectures. In applications to 22 complex traits in the UK Biobank, RHE-mc yields heritability estimates on array SNPs that are lower on average relative to S-LDSC and SumHer. We have explored the utility of RHE-mc in heritability partitioning analyses. These analyses show that allelic effect sizes tend to increase with a decrease in MAF and LD consistent with previous studies [29]. We also partitioned heritability across functional annotations to reveal enrichment of heritability at FANTOM5 enhancers in specific traits such as asthma and eczema.

We discuss several limitations of RHE-mc as well as directions for future work. First, the method-of-moments estimator underlying RHE-mc tends to yield slightly larger standard errors, on average, relative to REML estimators. The relative performance of the two methods likely depends on a number of aspects of the study design such as sample size, number of SNPs, the LD structure, relatedness patterns, and the underlying genetic architecture. Nevertheless, our method is designed to be applicable to massive datasets for which the heritability estimates are relatively precise. Developing scalable variance components estimators that are as efficient as REML-based methods is an important direction for future work. Second, this work has primarily explored the partitioning of heritability across discrete annotations. While we have shown how the methodology can be extended to continuous-valued annotations, it would be of interest to explore variation in trait heritability as a function of the value of an annotation. On the other hand, the ability of RHE-mc to fit many annotations allows the annotation to be divided into a sufficiently large number of bins. Third, we have applied RHE-mc to binary traits available in the UK Biobank treating these traits as continuous. Methods that explicitly model binary traits as well as the underlying ascertainment involved in case-control studies are likely to lead to more accurate heritability estimates [33, 125]. For example, the PCGC method [33] is an extension of HE regression and it would be of interest to develop a scalable randomized PCGC estimator. Fourth, RHE-mc requires access to individual-level genotype and phenotype data.

Methods that only require summary statistic data (GRE [44], LDSC [3], and SumHer [107]) have the advantage of being applicable to datasets where acquiring access to individual-level data can be challenging [44]. Finally, our method could potentially lead to improvements in association testing, trait prediction, and understanding of polygenic selection.

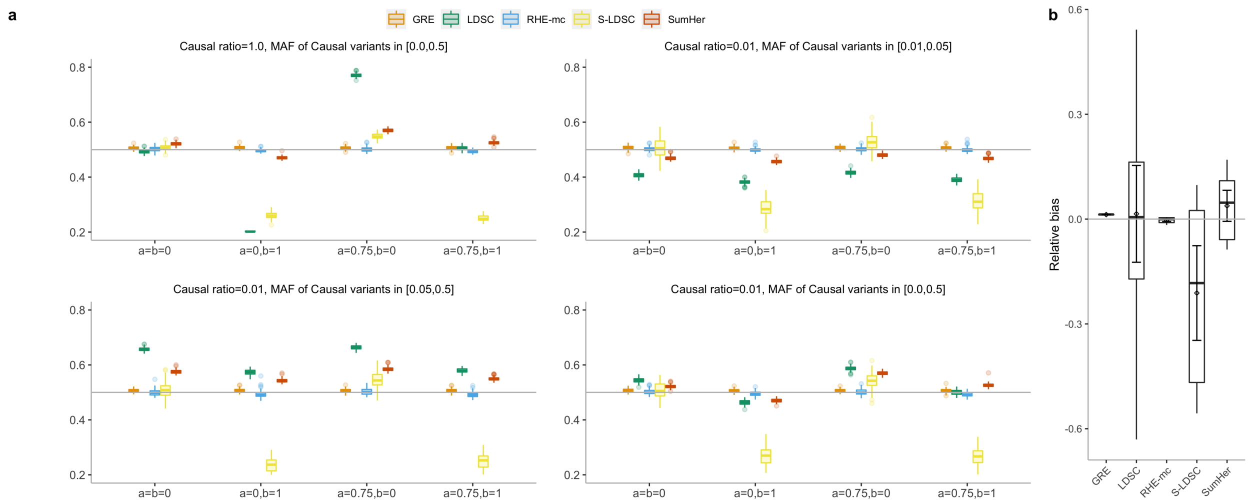


Figure 2.1: **Comparison of estimates of genome-wide SNP heritability from RHE-mc with LDSC, GRE , S-LDSC , and SumHer in large-scale simulations ($N = 337,205$ unrelated individuals, $M = 593,300$ array SNPs).** **a**: We compared methods for heritability estimation under 16 different genetic architectures. We set true heritability to 0.5 and varied the MAF range of causal variants (MAF of CV), the coupling of MAF with effect size ($a = 0$ indicates no coupling of MAF and $a = 0.75$ indicates coupling of MAF), and the effect of local LD on effect size ($b = 0$ indicates no LDAK weights and $b = 1$ indicates LDAK weights) . Each boxplot represents estimates from 100 simulations. **b**: Relative bias of each method (as a percentage of the true h^2) across 16 distinct MAF- and LD-dependent architectures. Each boxplot contains 16 points; each point is the relative bias estimated from 100 simulations under a single genetic architecture.Points and error bars represent the mean and ±2 SE. In **a** and **b**, boxplot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles, respectively . Here, we run RHE-mc using 24 bins formed by the combination of 6 bins based on MAF as well as 4 bins based on quartiles of the LDAK score of a SNP . We run S-LDSC with only 10 MAF bins (see Supplementary Table 5). To do a fair comparison, for every method, we computed LD scores and LDAK weights by using in-sample LD, and in all simulations we aim to estimate the SNP-heritability explained by the same set of M SNPs.

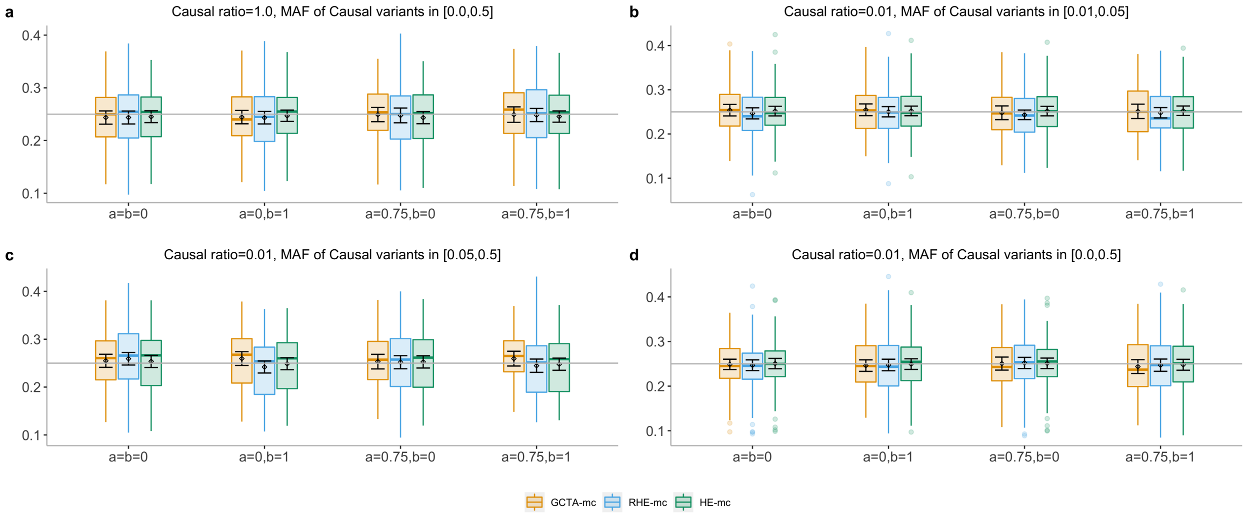


Figure 2.2: **Comparison of SNP heritability estimates from RHE-mc with GCTA-mc (GCTA with multiple variance components) and HE-mc(HE with multiple variance components)** ($N = 10,000$ **unrelated individuals,** $M = 593,300$ **array SNPs)**. In **a**, **b**, **c** and **d**: We compared heritability estimates from these methods under 16 different genetic architectures. We varied the MAF range of causal variants (MAF of CV), the coupling of MAF with effect size (a), and the effect of local LD on effect size ($b = 0$ indicates no LDAK weights and $b = 1$ indicates LDAK weights . We ran 100 replicates where the true heritability of the phenotype is 0.25. We run RHE-mc, HE-mc and GCTA-mc using 24 bins formed by the combination of 6 bins based on MAF as well as 4 bins based on quartiles of the LDAK score of a SNP . Across all different genetic architectures, the relative biases range from −3.2% to 3.6% for RHE-mc, and from −3.2% to 5% for GCTA-mc, and from −2.6% to 1.45% for HE-mc . On average, RHE-mc has SEs that are 1.1 and 1.08 times larger than GCTA-mc and HE-mc respectively. Black points and error bars represent the mean and ±2 SE. Each boxplot represents estimates from 100 simulations. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles, respectively. The SE's are computed from 100 simulations (Note that GCTA-mc did not run successfully on all 100 simulations).

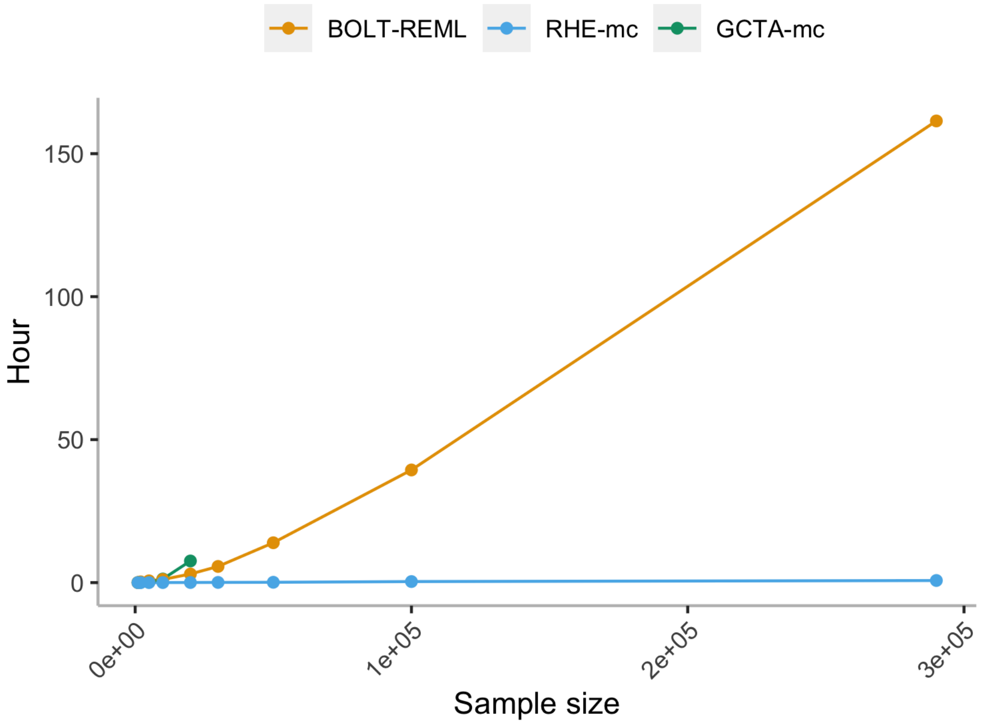


Figure 2.3: **Comparison of running time of RHE-mc, GCTA-mc, and BOLT-REML**. We compared runtime of RHE-mc, GCTA-mc, and BOLT-REML with increasing sample size N (for a fixed number of SNPs $M = 459,792$ and components $K = 22$).

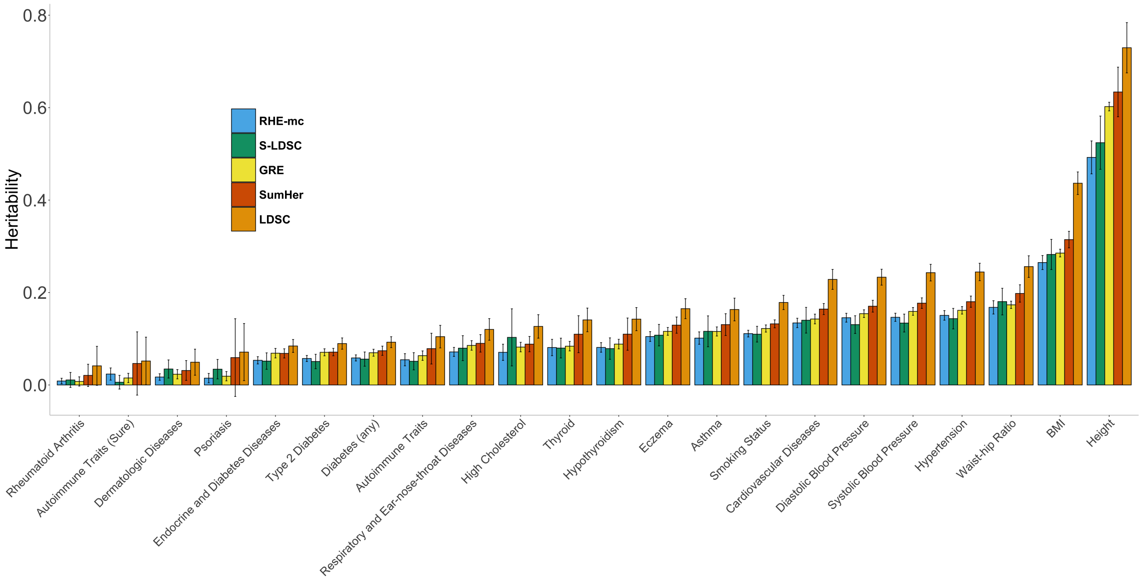


Figure 2.4: **Estimates of genome-wide SNP heritability from RHE-mc, LDSC, S-LDSC, GRE, and SumHer for 22 complex traits and diseases in the UK Biobank**. We restricted our analysis to $N = 291,273$ unrelated white British individuals. We applied all methods to $M = 459,792$ array SNPs (MAF$>$ 1%). We ran S-LDSC with baseline-LD model. For every method, LD scores or LDAK weights are computed using in-sample LD among the SNPs, and we aim to estimate the SNP-heritability explained by the same set of SNPs. RHE-mc was applied to array SNPs with 8 MAF/LD bins. Black error bars mark ±2 standard errors centered on the estimated heritability. We used a block Jackknife (number of blocks = 100) to estimate the standard errors. In supplementary Figure 7, we also report RHE-mc estimates of genome-wide SNP heritability on $M = 4,824,392$ imputed SNPs (MAF $>$ 1%) with 8 MAF/LD bins and $M = 7,774,235$ imputed SNPs (MAF $>$ 0.1%) with 144 MAF/LD bins .

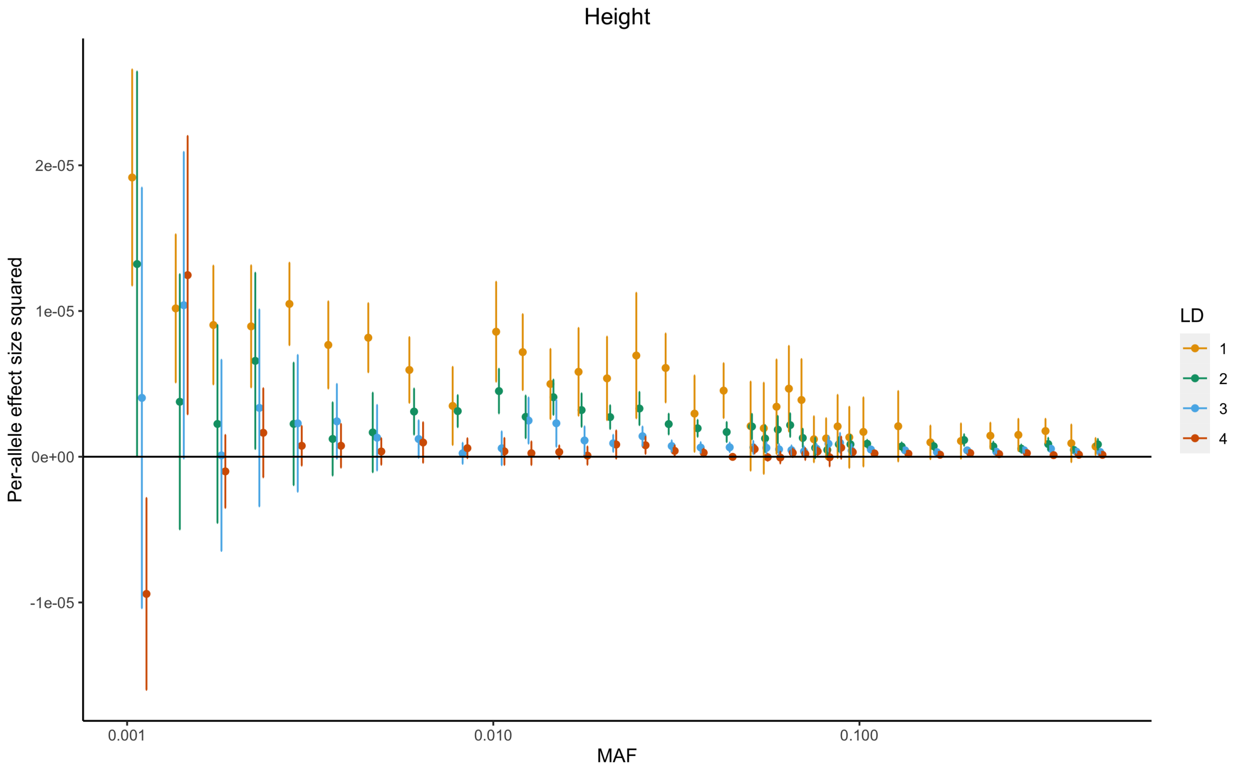


Figure 2.5: **Per-allele squared effect size of height as a function of MAF**: We applied RHE-mc to $N = 291,273$ unrelated white British individuals and $M = 7,774,235$ imputed SNPs. SNPs were partitioned into 144 bins based on LD score (4 bins based on quartiles of the LD score with i denoting the i^{th} quartile) and MAF (36 MAF bins) . Per-allele effect size squared for bin k is defined as $\frac{h_k^2}{M_k * 2f_k * (1-f_k)}$ where h_k^2 is the heritability attributed to bin k, M_k is the number of SNPs in bin k, and f_k is the average MAF in bin k. Each point represents the estimated allelic effect size. Bars mark ±2 standard errors centered on the estimated allelic effect size. See Supplementary Figures for results on all 22 traits.

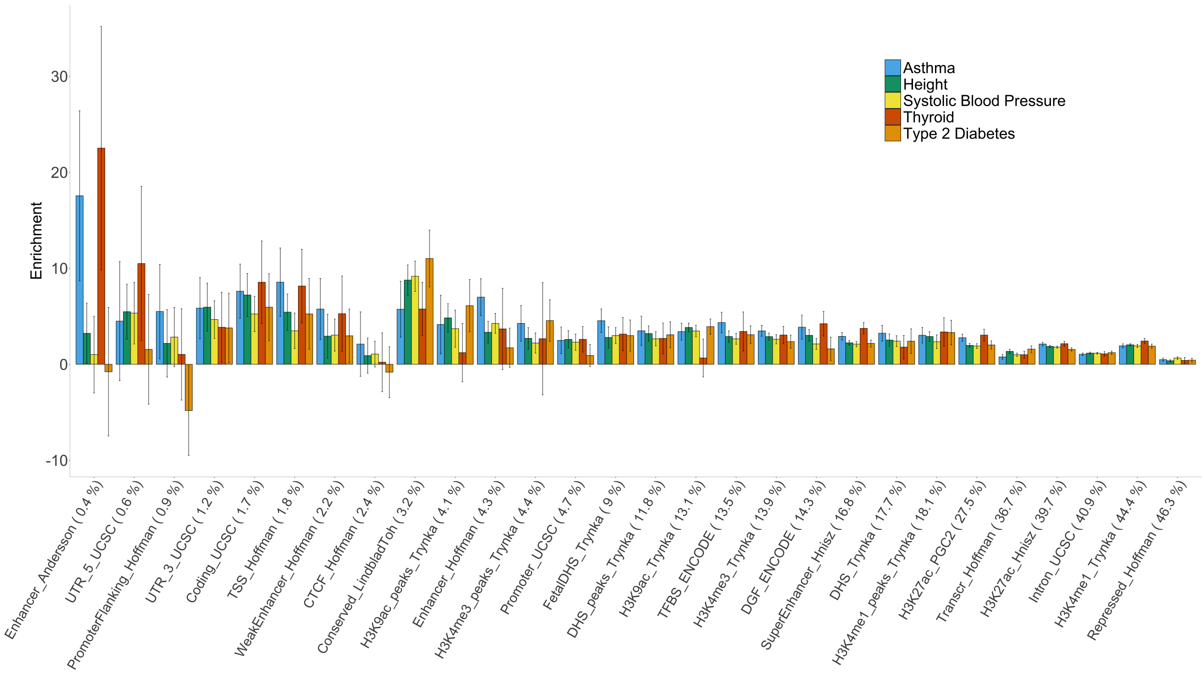


Figure 2.6: **Enrichment of heritability across 28 functional annotations**: We applied RHE-mc to $N = 291,273$ unrelated white British individuals and $M = 5,670,959$ imputed SNPs (MAF $> 0.1\%$ and present in 1000 Genomes Project). SNPs were partitioned based on 28 functional annotations that were defined in a previous study [21]. We grouped 22 traits in the UK Biobank into five categories (autoimmune, diabetes, respiratory, anthropometric, cardiovascular). Here we plot enrichment of five traits (one representative trait per category). Each bar represents the estimated enrichment. Black error bars mark ±2 standard errors centered on the estimated enrichment. Annotations are ordered by the proportions of SNPs in that annotation (given in parentheses). See Supplementary Figure 8 for results on all 22 traits.

Parameters			Running time (hour)		
N	M	K	RHE-mc	GCTA-mc	BOLT-REML
10,000	459,792	22	< 1	1.3	1
100,000	459,792	22	< 1	-	40
291,273	459,792	22	< 1	-	162
291,273	459,792	300	< 1	-	-
291,273	4,824,392	8	3.2	-	-
$1,000,000$	$1,000,000$	8	3	-	-
$1,000,000$	$1,000,000$	100	12.4	-	-

Table 2.1: **Comparison of running time of RHE-mc, GCTA-mc, and BOLT-REML**. Here M, N and K are the number of SNPs, individuals and variance components respectively. RHE-mc can run efficiently even on datasets with one million individuals and SNPs as well as efficiently computing hundreds of variance components. All comparisons were performed on an Intel(R) Xeon(R) CPU 2.10 GHz server with 128 GB RAM.

2.5 Supplementary Notes

2.5.1 Computing the standard errors of the estimates

We obtain standard errors for RHE-mc using a block jackknife [56]. A jackknife subsample is created by leaving out a subset of observations from a dataset. The jackknife estimate of a parameter can be found by estimating the parameter for each subsample, omitting the i-th jackknife block. A naive way to compute jackknife estimate requires computing the estimator of the parameters for every sub-sample. For instance, in our problem, if we define J jackknife blocks, then we need to run RHE-mc for every sub-sample which takes $\mathcal{O}(J(\frac{NMB}{\max(\log_3(N),\log_3(M))}+K^2(K+NB)))$. We propose an efficient way to compute the jackknife estimate in time $\mathcal{O}(\frac{NMB}{\max(\log_3(N),\log_3(M))}+JK^2(K+NB))$.

Let $\boldsymbol{X}$ be a $N \times M$ matrix of standardized genotypes where N and M are the numbers of individuals and SNPs, respectively. To generate J jackknife subsamples, we partition X into J non-overlapping blocks $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(J)}$ such that $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(J)}]$. Note that for every j, $\boldsymbol{X}^{(j)}$ is a $N \times M_j$ matrix where M_j is the number of SNPs in the j-th block.

We create the j-th jackknife subsample by removing the j-th block $\boldsymbol{X}^{(j)}$ from $\boldsymbol{X}$. To estimate the variance components of the j-th jackknife subsample, we need to compute the corresponding quantities of the jth subsample in the normal equations (Methods). Let $\boldsymbol{K}_k^{(-j)}$ be the GRM of the k-th partition which is created by removing the j-th block $\boldsymbol{X}^{(j)}$ from $\boldsymbol{X}$ where $k \in \{1, \ldots, K\}$, $j \in \{1, \ldots, J\}$. In Algorithm 1, we show how we can compute $tr(\widehat{\boldsymbol{K}_k^{(-j)} \boldsymbol{K}_l^{(-j)}})$ and $\boldsymbol{y}^T \boldsymbol{K}_i^{(-j)} \boldsymbol{y}$, for all $k, l \in \{1, \ldots, K\}$, $j \in \{1, \ldots, J\}$ efficiently.

2.5.2 Including covariates

We can extend the LMM to include covariates as follows:

$$\boldsymbol{y} | \boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k = \boldsymbol{W}\boldsymbol{\alpha} + \sum_k \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon} \tag{2.13}$$

Here $\boldsymbol{W}$ is a $N \times C$ matrix of covariates while $\boldsymbol{\alpha}$ is a C-vector of fixed effects.

It is easy to see that the matrix $\boldsymbol{V} = \boldsymbol{I}_N - \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$ is symmetric and idempotent ($\boldsymbol{V}^2 = \boldsymbol{V}$) of rank $N - C$. Therefore, we consider the eigendecomposition of $\boldsymbol{V} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$, where $\boldsymbol{D}$ is a diagonal matrix with $N - C$ ones and C zeros on the diagonal (we can assume that first $N - C$ elements are one). Now let the matrix $\boldsymbol{U}_{N \times (N-C)}$ represent the first $N - C$ columns of $\boldsymbol{E}$. It is not hard to see that $\boldsymbol{U}$ satisfies $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{N-C}$, $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{V}$, $\boldsymbol{U}^T\boldsymbol{W} = 0$. Now we multiplying by $\boldsymbol{U}^T$ on both sides of the above equation:

$$\boldsymbol{U}^T\boldsymbol{y} = \boldsymbol{U}^T \sum_k \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{U}^T\boldsymbol{\epsilon} \tag{2.14}$$

$$cov(\boldsymbol{U}^T\boldsymbol{y}) = E[\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T] - E[\boldsymbol{U}^T\boldsymbol{y}]E[\boldsymbol{U}^T\boldsymbol{y}] \tag{2.15}$$

The matrix $\boldsymbol{U}^T$ is constant and the vector $\boldsymbol{y}$ is random. Therefore, we have $E[\boldsymbol{U}^T\boldsymbol{y}] = \boldsymbol{U}^T E[\boldsymbol{y}]$.

$$\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T = (\boldsymbol{U}^T \sum_k \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{U}^T\boldsymbol{\epsilon})(\boldsymbol{U}^T \sum_k \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{U}^T\boldsymbol{\epsilon})^T = \tag{2.16}$$

$$\sum_i\sum_j \boldsymbol{U}^T\boldsymbol{X}_i\boldsymbol{\beta}_i(\boldsymbol{U}^T\boldsymbol{X}_j\boldsymbol{\beta}_j)^T + (\boldsymbol{U}^T\boldsymbol{\epsilon})\sum_i(\boldsymbol{U}^T\boldsymbol{X}_i\boldsymbol{\beta}_i)^T + \sum_i \boldsymbol{U}^T\boldsymbol{X}_i\boldsymbol{\beta}_i(\boldsymbol{U}^T\boldsymbol{\epsilon})^T + \boldsymbol{U}^T\boldsymbol{\epsilon}(\boldsymbol{U}^T\boldsymbol{\epsilon})^T$$

Hence

$$E[\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T] = \sum_k \frac{\sigma_{g_k}^2}{M_k}(\boldsymbol{U}^T\boldsymbol{X}_k)(\boldsymbol{U}^T\boldsymbol{X}_k)^T + \sigma_\epsilon^2\boldsymbol{U}^T\boldsymbol{U} \tag{2.17}$$

Using $\boldsymbol{K}_k = \frac{\boldsymbol{X}_k\boldsymbol{X}_k^T}{M_k}$, we have:

$$cov(\boldsymbol{U}^T\boldsymbol{y}) = \boldsymbol{U}^T(\sum_k \sigma_{g_k}^2\boldsymbol{K}_k)\boldsymbol{U} + \sigma_\epsilon^2\boldsymbol{I}_{N-C} \tag{2.18}$$

The MoM estimator is obtained by solving the following ordinary least squares problem:

$$(\tilde{\sigma_1^2}, \ldots, \tilde{\sigma_K^2}, \tilde{\sigma_e^2}) = argmin_{(\sigma_1^2,\ldots,\sigma_K^2,\sigma_e^2)}||\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T - \boldsymbol{U}^T(\sum_k \sigma_k^2\boldsymbol{K}_k)\boldsymbol{U} - \sigma_\epsilon^2\boldsymbol{I}_{N-C}||_F^2 \tag{2.19}$$

We need to solve the following normal equations to estimate the variance components.

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{b} \\ \boldsymbol{b}^T & N-C \end{bmatrix} \begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_k^2 \\ \sigma_e^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{y}^T\boldsymbol{V}\boldsymbol{y} \end{bmatrix} \tag{2.20}$$

Here $\boldsymbol{V} = \boldsymbol{I}_N - \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$ and $\boldsymbol{T}$ is a $K \times K$ matrix where $T_{k,l} = tr(\boldsymbol{K}_k\boldsymbol{V}\boldsymbol{K}_l\boldsymbol{V})$, and b is a $K-$vector where $b_k = tr(\boldsymbol{V}\boldsymbol{K}_k)$, and $\boldsymbol{c}$ is a K- vector where $c_k = \boldsymbol{y}^T\boldsymbol{V}\boldsymbol{K}_k\boldsymbol{V}\boldsymbol{y}$. Commonly, the number of covariates C is small (tens to hundreds) so that including covariates does not significantly affect the computational cost. The cost of computing the elements of the normal equations 2.20 includes the cost of inverting $\boldsymbol{W}^T\boldsymbol{W}$ which is a $C \times C$ matrix and multiplying $\boldsymbol{W}$ by a real-valued N-vector which can be done in $\mathcal{O}(C^3 + NC)$.

2.5.3 Streaming version

Here we describe the streaming version of RHE-mc algorithm. In the Methods section, we showed that our MoM estimator satisfies the following normal equation.

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{b} \\ \boldsymbol{b}^T & N \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\sigma}_{\boldsymbol{g}}^{\boldsymbol{2}}} \\ \tilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix} \tag{2.21}$$

Here $\tilde{\boldsymbol{\sigma}_{\boldsymbol{g}}^{\mathbf{2}}} = \begin{bmatrix} \tilde{\sigma_1^2} \\ \vdots \\ \tilde{\sigma_K^2} \end{bmatrix}$, $\boldsymbol{T}$ is a $K \times K$ matrix with entries $T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l), k, l \in \{1, \ldots, K\}$, $\boldsymbol{b}$ is a K-vector with entries $b_k = tr(\boldsymbol{K}_k) = N$ (because $\boldsymbol{X}_k$s is standardized), and $\boldsymbol{c}$ is a K-vector with entries $c_k = \boldsymbol{y}^T \boldsymbol{K}_k \boldsymbol{y}$. Here we estimate $T_{k,l}$ as follows :

$$T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l) \approx \widehat{T_{k,l}} = \frac{1}{B} \frac{1}{M_k M_l} \sum_b \boldsymbol{z}_b^T \boldsymbol{X}_k \boldsymbol{X}_k^T \boldsymbol{X}_l \boldsymbol{X}_l^T \boldsymbol{z}_b \tag{2.22}$$

Here $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$ are B independent random vectors with zero mean and covariance $\boldsymbol{I}_N$.

We read genotype matrix $\boldsymbol{X}_k$ for every $k \in \{1, \ldots, K\}$ block by block. We define J blocks over $\boldsymbol{X}_k$ by partitioning the columns of $\boldsymbol{X}_k$ to J groups such that $\boldsymbol{X_k} = [\boldsymbol{X_k}^{(1)} \ldots \boldsymbol{X_k}^{(J)}]$.

Algorithm 1 Streaming version of RHE-mc

```
for $k = 1$ to $K$ do
    for $j = 1$ to $J$ do
        Read $\boldsymbol{X}^j_k$
        for $b = 1$ to $B$ do
            $Z_{(k,j,b)} = \boldsymbol{X}^j_k {\boldsymbol{X}^j_k}^T z_b$
        end for
        $v = {\boldsymbol{X}^j_k}^T y$
        $H_{(k,j)} = v^T v$
        Release the memory allocated to $\boldsymbol{X}^j_k$
    end for
end for
Let $U_{k,b,0} = \sum_j Z_{(k,j,b)}$, and
Let $U_{k,b,j} = U_{k,b,0} - Z_{(k,j,b)}$, for every $k, b, j$.
Let $V_{k,0} = \sum_j H_{k,j}$,
Let $V_{k,j} = V_{k,0} - H_{k,j}$ for every $k, j$
for $j = 0$ to $J$ do
    for every pair of genotype matrices $k$ and $l$ do
        $\widehat{T_{k,l}} = \frac{1}{B} \frac{1}{M_k M_l} \sum_b U^T_{(k,j,b)} U_{(l,j,b)}$
    end for
    for every genotype matrix $k$ do
        $c_k = \frac{1}{M_k} V^2_{(k,j)}$
    end for
    Solve the normal equation for $j^{th}$ sub-sample ($j = 0$ corresponds to the original genotype matrix
  used for computing the point estimates)
end for
Compute the jackknife SE from the point estimates of $J$ sub-samples.
```

In the above algorithm, the 3-D matrices Z and U need $\mathcal{O}(JKBN)$ memory, the 2-D matrices V and H need $\mathcal{O}(JN)$ memory. So the total space complexity will be $\mathcal{O}(JKBN)$. The total running time of this implementation is $\mathcal{O}(\frac{NMB}{\max(\log_3(N), \log_3(M))} + JK^2(K + NB)))$. For simplicity, we assume that the streaming blocks are the same as jackknife blocks. However, we can set the size of the streaming blocks to be different from the jackknife blocks to make the algorithm more efficient in terms of memory usage.

2.5.4 Parameter settings for summary statistics methods

For running LDSC we computed the LD score of each SNP within 2-Mb windows centered on the SNP. We ran LD score regression with an unconstrained intercept and with regression weights that account for correlations between association statistics at SNPs in LD and heteroscedasticity [3]. To prevent the LDSC software from dropping high-effect SNPs we used the following flags –not-M-5-50 and –chisq-max 99999.
In simulations, we ran S-LDSC with 10 binary MAF bins, which are defined such that each bin contains 10% of the typed SNPs; this is done to reflect the 10 MAF bin annotations in the S-LDSC baseline-LD model [28] (see Table 2.6 for the details of MAF bins). In analyzing the 22 real complex traits, we run S-LDSC with baseline-LD model[28].
To run SumHer, first we computed the default LDAK weights using in-sample LD [109]. After that we computed LD tagging using 1-Mb windows centered on each SNP and setting $\alpha = -0.25$ as recommended [107]. We used default values for the other parameter settings for running SumHer.

To do a direct comparison among LDSC, S-LDSC, and SumHer, we ran an in-sample LD version of each method meaning that we used same set of SNPs to compute LD scores and LDAK weights, perform the regression, and estimate SNP-heritability.

2.5.5 Continuous annotations

We assessed the accuracy of RHE-mc in estimating variance components with continuous annotation. We simulated a phenotype with true heritability 0.5 from $9K$ individuals and $15k$ SNPs under the GCTA model. We ran RHE-mc with single component, no annotations, and standardized genotypes. We next ran RHE-mc with single component, non-standardized genotypes, where we added a continuous annotation defined as $1/var(i)$ for SNP i where $var(i)$ is the variance of SNP i across individuals. We obtain a concordant estimate of genome-wide SNP heritability 0.45 ± 0.03 in the first case and 0.46 ± 0.03 in the second case.

2.5.6 Power as a function of annotation size

To quantify the power of RHE-mc as a function of the size of an annotation, we performed simulations using $N = 291,273$ unrelated white British individuals and $M = 459,792$ common SNPs. We defined 8 annotations (4 MAF bins and 2 LD bins) in which we fixed the heritability of a selected bin and varied the proportion of SNPs in the selected category. We then plotted the probability of rejection; the results are displayed in Supplementary Figure 11 . Furthermore, we simulated phenotypes in which we fixed the enrichment of a selected bin and varied the size of the selected bin, the results are displayed in Supplementary Table 6.

2.6 Supplementary Figures

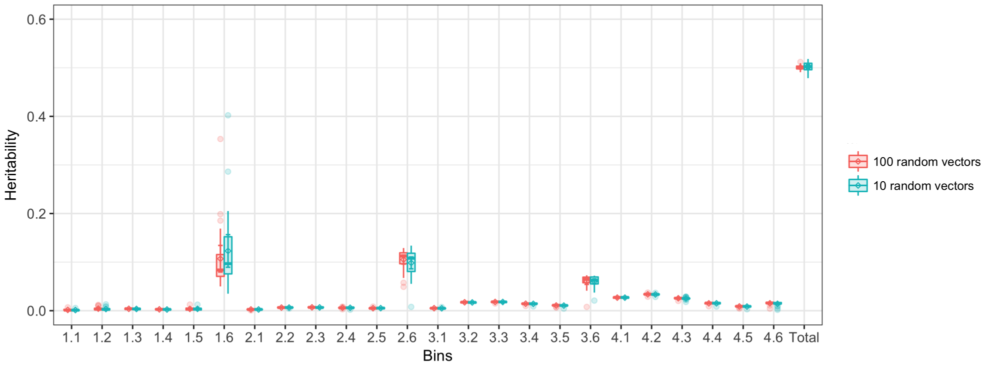


Figure 2.7: **Comparison of RHE-mc heritability estimates with $B = 10$ and $B = 100$ random vectors on large-scale simulated data (M=590K array SNPs and N=337K individuals)**: We ran RHE-mc with 24 bins(based on 6 MAF bins and 4 LDAK bins, see Methods). Here x-axis represents the bins ($i.j$ denotes the bin defined based on i-th ldak bin and j-th MAF bin) and y-axis represents the heritability. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles, respectively. Each box plot represents estimates from 100 simulations. Diamond points and error bars represent the mean and ±2 SE centered on estimated heritability, respectively. Mean and standard errors (SE's) are computed from 100 replicates.

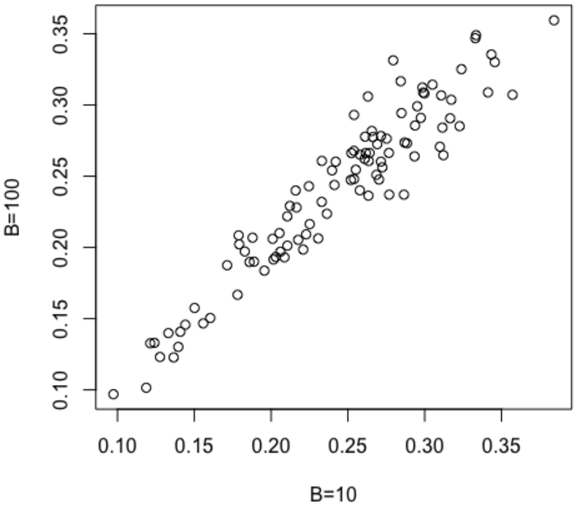


Figure 2.8: Comparison of RHE-mc estimates with B=10 and B=100 on small scale data (M=590K array SNPs and N=10k individuals): We simulated 100 phenotypes such that the true total heritability is 0.25. Here x-axis represents the RHE-mc estimates when $B = 10$, and y-axis represents RHE-mc estimates when $B = 100$.

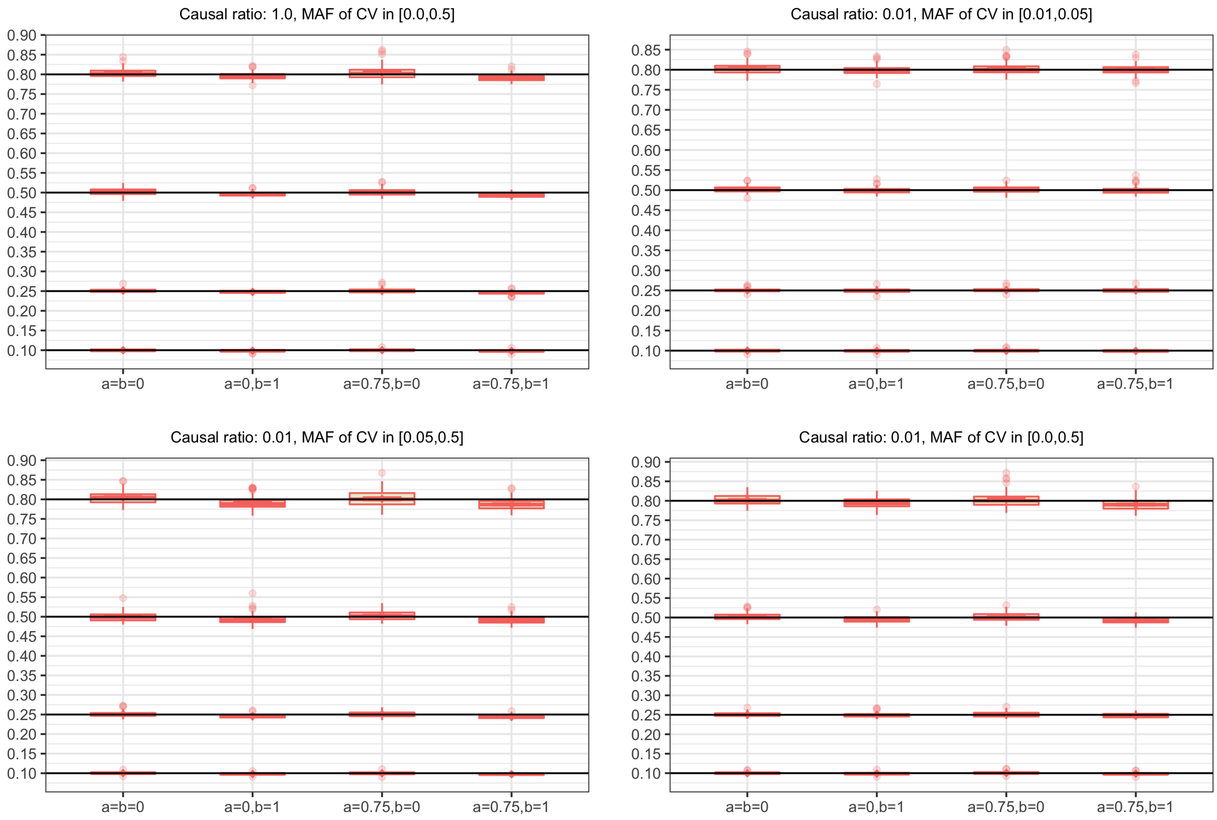


Figure 2.9: **Accuracy of genome-wide SNP heritability estimated by RHE-mc across** 64 **distinct MAF- and LD-dependent architectures in genome-wide simulations (**$N = 337,205$ **unrelated individuals,** $M = 593,300$ **array SNPs)**. For simulating the phenotypes, we chose true heritability from $\{0.1, 0.25, 0.5, 0.8\}$, varied the ratio of causal variants (causal ratio $\in \{0.01, 1.0\}$), varied the MAF range of causal variants (MAF of CV), the coupling of MAF with effect size ($a = 0$ indicates no coupling of MAF and $a = 0.75$ indicates coupling of MAF), and the effect of local LD on effect size ($b = 0$ indicates no LDAK weights and $b = 1$ indicates LDAK weights) . We ran RHE-mc using 24 bins formed by the combination of 6 bins based on MAF as well as 4 bins based on quartiles of the LDAK score of a SNP . Boxplot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles, respectively. Each box plot represents RHE-mc estimates from 100 simulations.

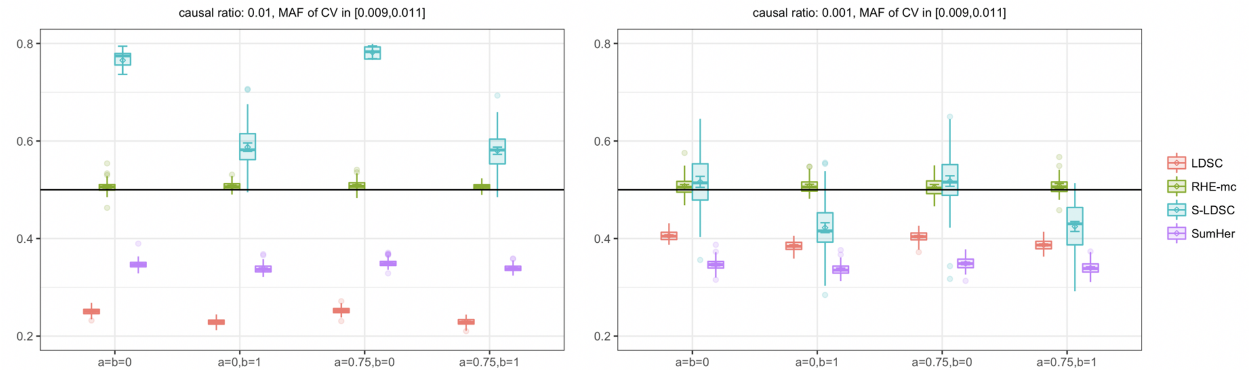


Figure 2.10: **Comparison of estimates of genome-wide SNP heritability from RHE-mc with LDSC, S-LDSC , and SumHer when only rare variants are causal in large-scale simulations (**$N = 337,205$ **unrelated individuals,** $M = 593,300$ **array SNPs)**. We compared methods for heritability estimation under different genetic architectures when only rare variants are causal. We set true heritability to 0.5, the MAF range of causal variants (MAF of CV) to be between $[0.009, 0.011]$ and varied the coupling of MAF with effect size ($a = 0$ indicates no coupling of MAF and $a = 0.75$ indicates coupling of MAF), and the effect of local LD on effect size ($b = 0$ indicates no LDAK weights and $b = 1$ indicates LDAK weights) . Here, we run RHE-mc using 24 bins formed by the combination of 6 bins based on MAF as well as 4 bins based on quartiles of the LDAK score of a SNP . We run S-LDSC with 10 MAF bins (see Supplementary Table S5). To do a fair comparison, for every method, we computed LD scores and LDAK weights by using in-sample LD, and in all simulations we aim to estimate the SNP-heritability explained by the same set of M SNPs. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles, respectively. Each box plot represents estimates from 100 simulations. Diamond points and error bars represent the mean and ±2 SE centered on estimated heritability respectively. Mean and standard errors (SE's) are computed from 100 simulations.

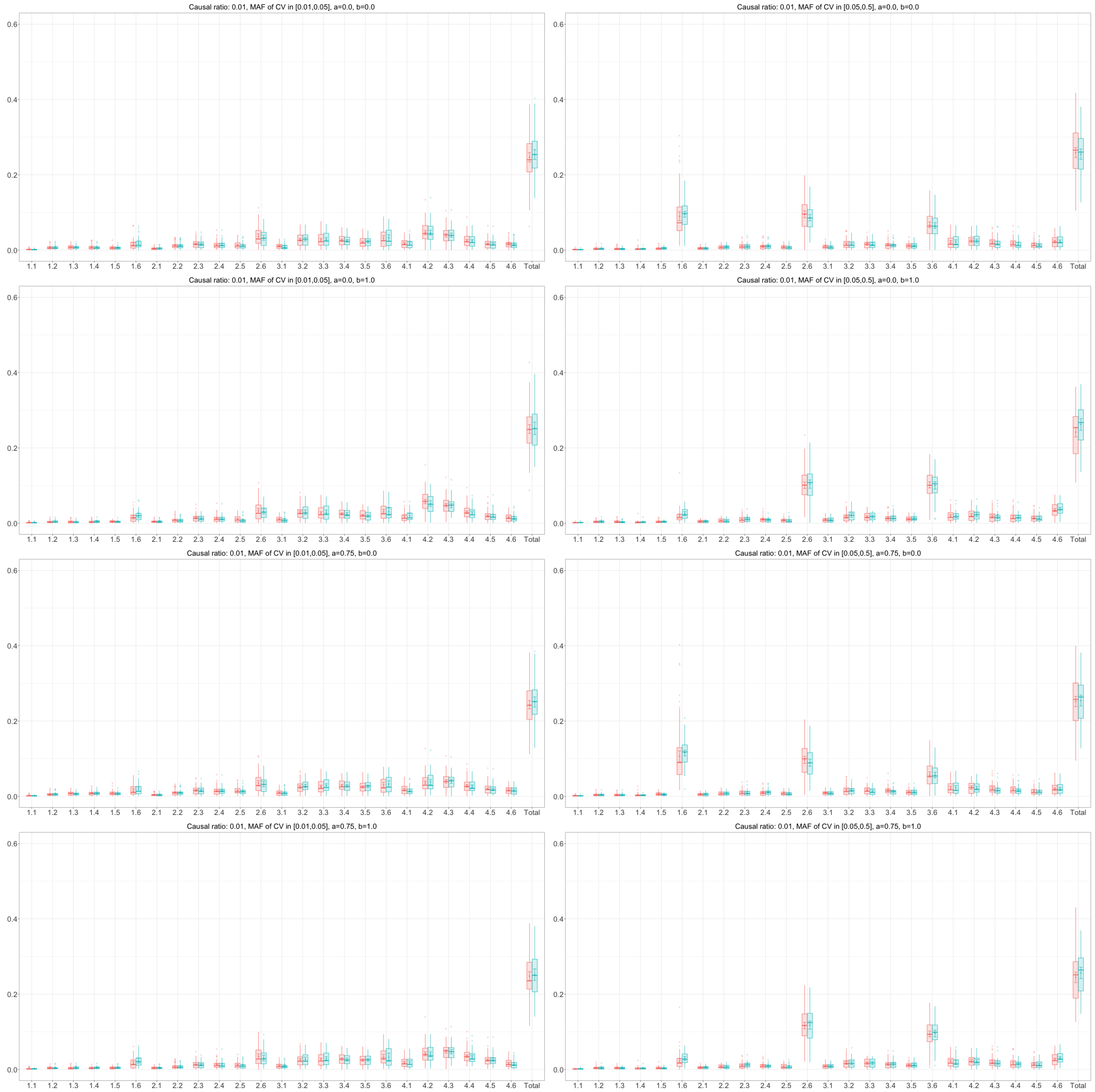


Figure 2.11: **Comparison of RHE-mc (red color) with GCTA-mc(blue color) in estimating partitioned heritability under 8 different genetic architectures on small-scale simulated data (M = 590k array SNPs and N = 10k individuals)**: We partition SNPs into 24 bins based on 6 MAF bins and 4 LDAK bins (Methods). True total SNP heritability is 0.25. Here x-axis represents the partitions ($i.j$ denotes the bin defined based on i-th ldak bin and j-th MAF bin. The lower bin number denotes the lower MAF (LDAK weights). For example, bin 1.6 contains SNPs which are in the first quartile of LDAK weights and MAF$>$ 0.05). y-axis represents the heritability. Each boxplot shows the distribution of estimates from 100 simulations. Note that GCTA-mc did not run successfully on all 100 simulations. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles, respectively. Each box plot represents estimates from 100 simulations. Diamond points and error bars represent the mean and ±2 SE centered on estimated heritability respectively. Mean and standard errors (SE's) are computed from 100 replicates.

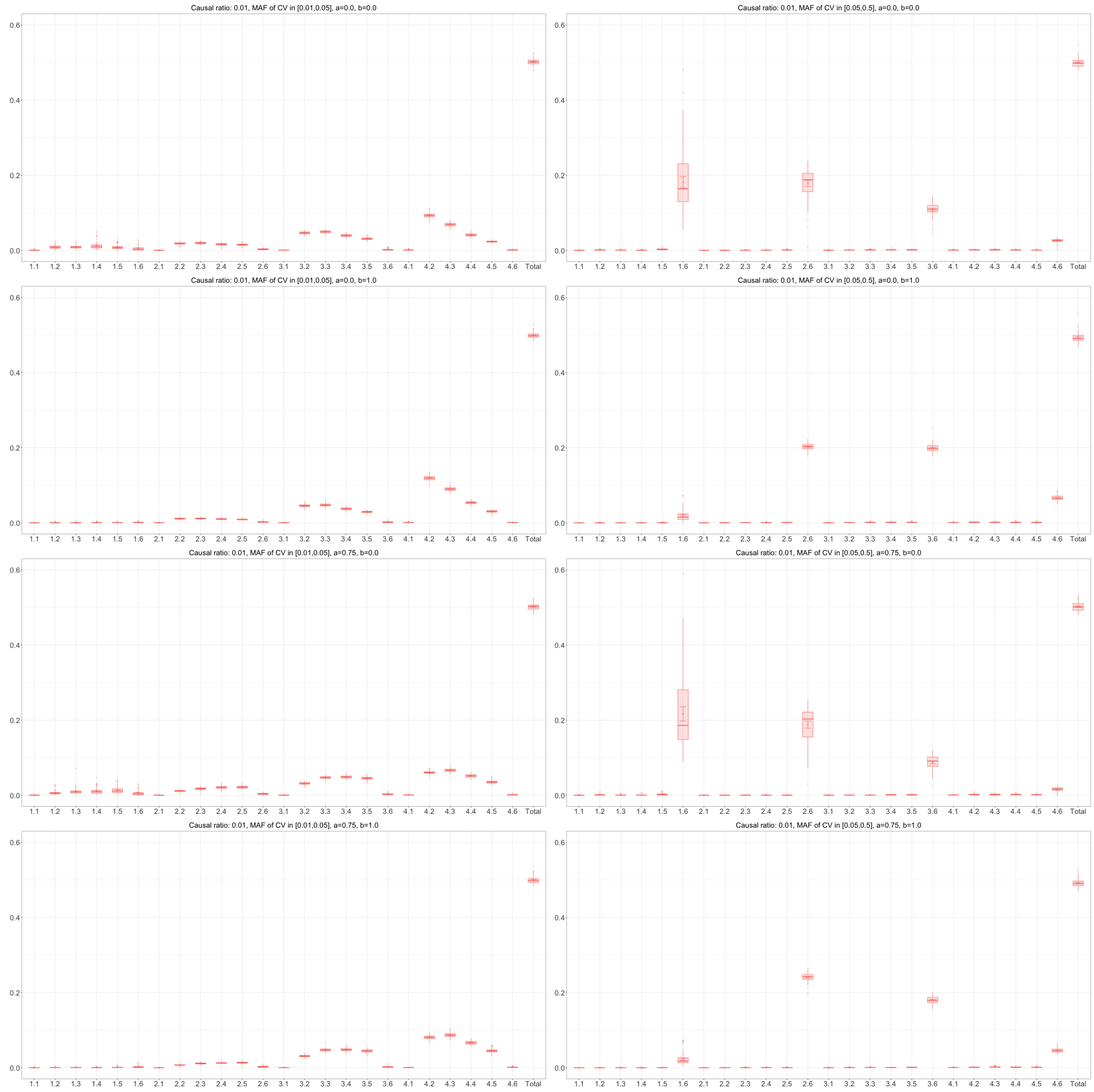


Figure 2.12: **Partitioned heritability estimates from RHE-mc on large-scale simulated data ($M = 590K$ array SNPs and $N = 337K$ individuals)**: We ran RHE-mc with 24 bins based on 6 MAF bins and 4 LDAK bins (Methods) over 8 different genetic architectures. Here x-axis represents the partitions ($i.j$ denotes the bin defined based on i-th ldak bin and j-th MAF bin) and y-axis represents the heritability. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles, respectively. Each box plot represents estimates from 100 simulations. Diamond points and error bars represent the mean and ±2 SE centered on estimated heritability respectively. Mean and standard errors (SE's) are computed from 100 replicates.

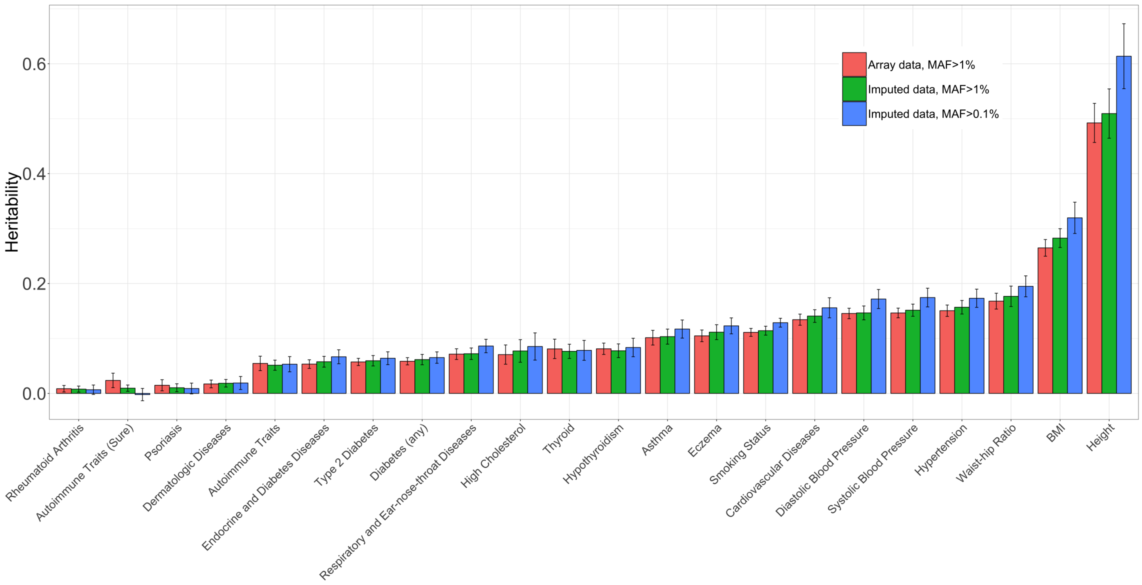
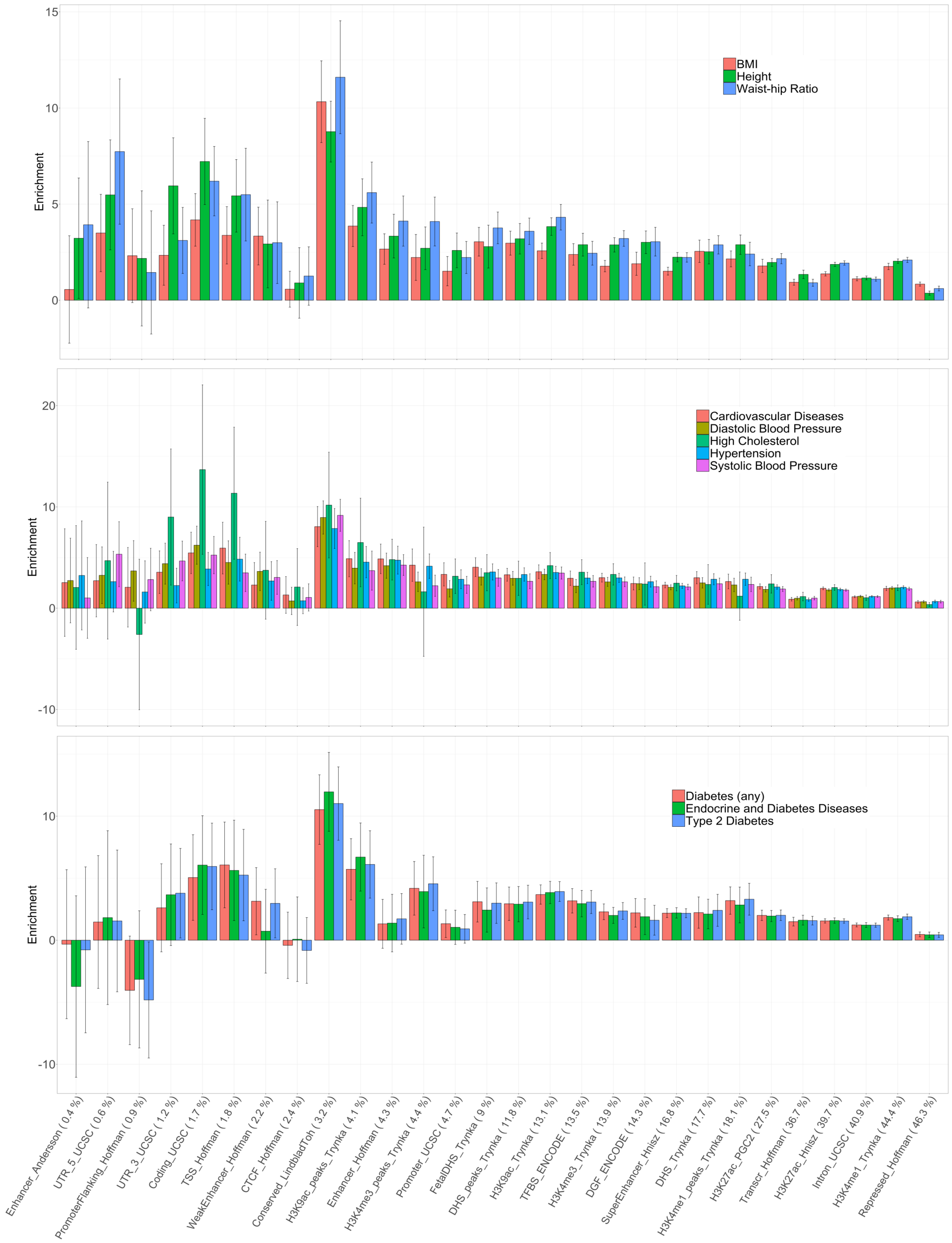


Figure 2.13: **Estimates of genome-wide SNP heritability from RHE-mc for 22 complex traits and diseases in the UK Biobank**: We restricted our analysis to $N = 291,273$ unrelated white British individuals. First, we applied RHE-mc to $M = 459,792$ array SNPs (MAF$>$ 1%) with 8 MAF/LD bins. Second, we applied RHE-mc to $M = 4,824,392$ imputed SNPs (MAF $>$ 1%) with 8 MAF/LD bins(Methods). Third, we applied RHE-mc to $M = 7,774,235$ imputed SNPs (MAF $>$ 0.1%) with 144 MAF/LD bins (Methods). Black bars mark ±2 standard errors centered on estimated heritability.


15
10
5
0
Enrichment
BMI
Height
Waist-hip Ratio
20
10
0
-10
Enrichment
Cardiovascular Diseases
Diastolic Blood Pressure
High Cholesterol
Hypertension
Systolic Blood Pressure
10
0
-10
Enrichment
Diabetes (any)
Endocrine and Diabetes Diseases
Type 2 Diabetes
Enhancer_Andersson (0.4 %)
UTR_5_UCSC (0.6 %)
PromoterFlanking_Hoffman (0.9 %)
UTR_3_UCSC (1.2 %)
Coding_UCSC (1.7 %)
TSS_Hoffman (1.8 %)
WeakEnhancer_Hoffman (2.2 %)
CTCF_Hoffman (2.4 %)
Conserved_LindbladToh (3.2 %)
H3K9ac_peaks_Trynka (4.1 %)
Enhancer_Hoffman (4.3 %)
H3K4me3_peaks_Trynka (4.4 %)
Promoter_UCSC (4.7 %)
FetalDHS_Trynka (9 %)
DHS_peaks_Trynka (11.8 %)
H3K9ac_Trynka (13.1 %)
TFBS_ENCODE (13.5 %)
H3K4me3_Trynka (13.9 %)
DGF_ENCODE (14.3 %)
SuperEnhancer_Hnisz (16.8 %)
DHS_Trynka (17.7 %)
H3K4me1_peaks_Trynka (18.1 %)
H3K27ac_PGC2 (27.5 %)
Transcr_Hoffman (36.7 %)
H3K27ac_Hnisz (39.7 %)
Intron_UCSC (40.9 %)
H3K4me1_Trynka (44.4 %)
Repressed_Hoffman (46.3 %)

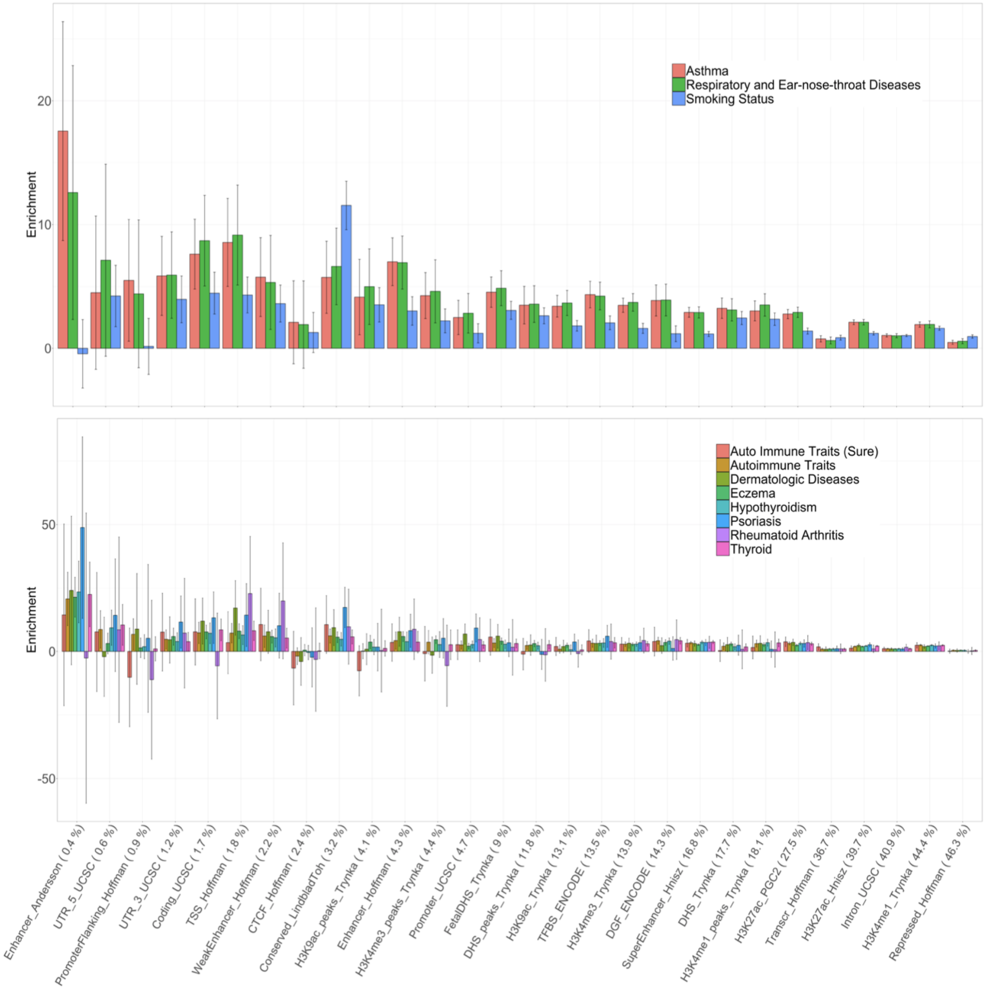
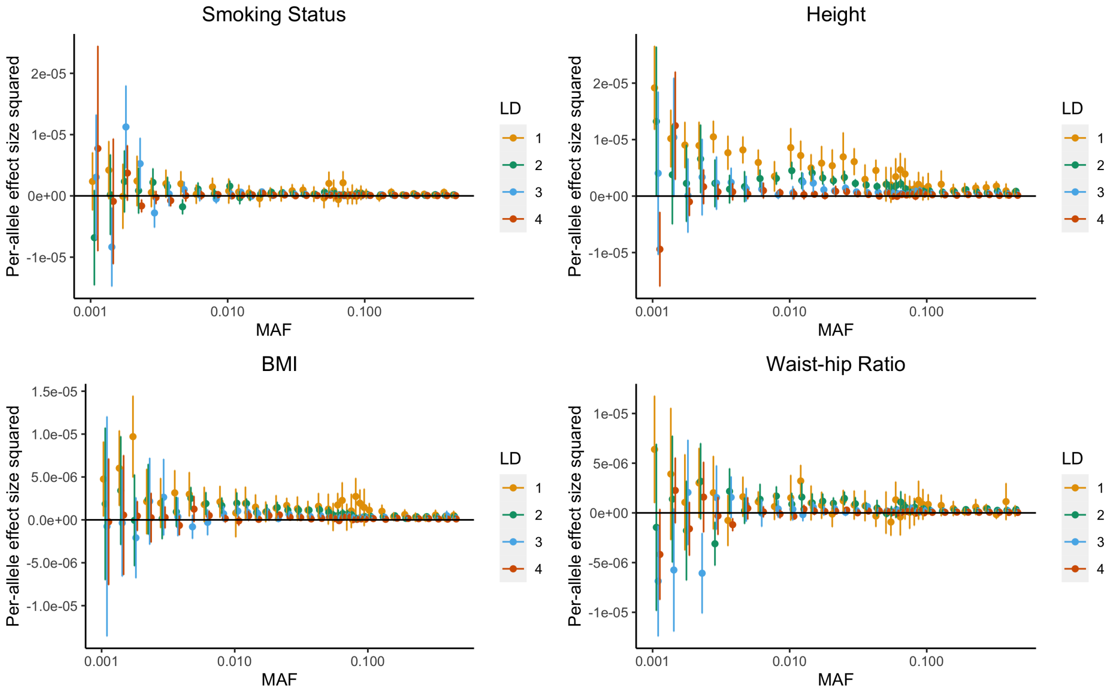
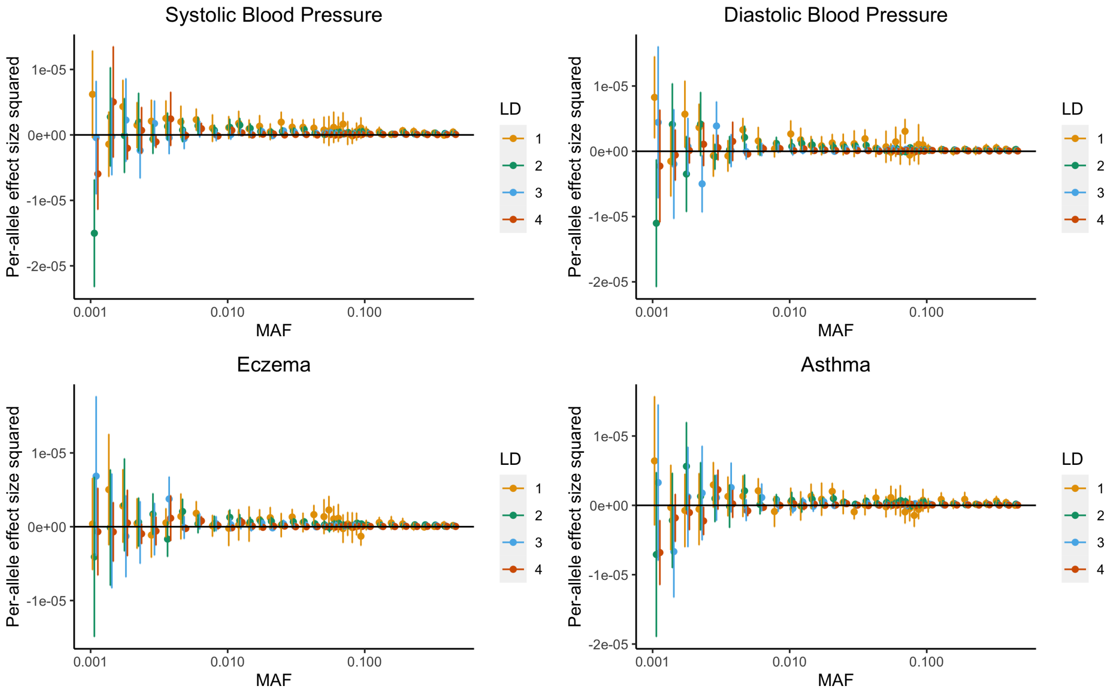
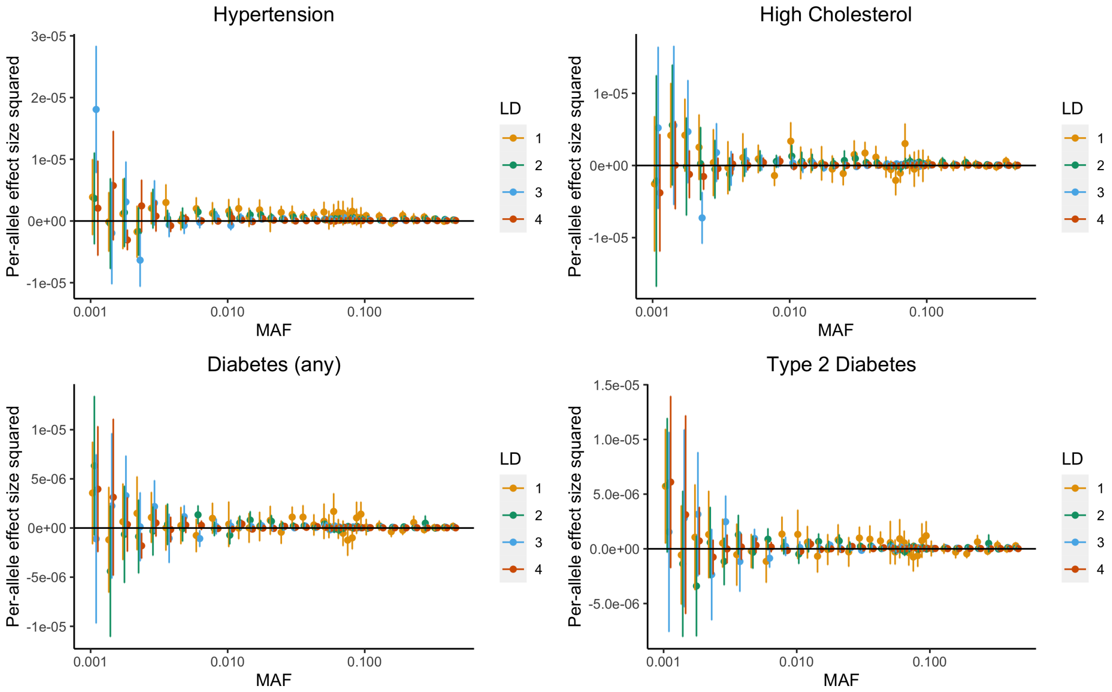
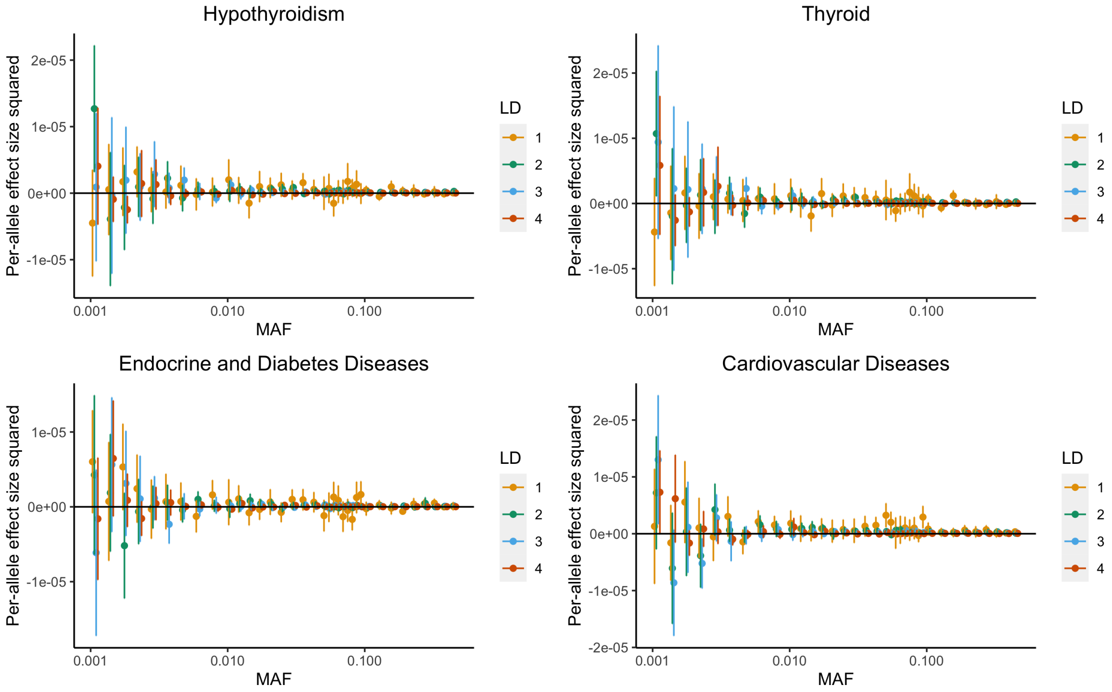


Figure 2.14: **Enrichment of heritability across 28 functional annotations**: We applied RHE-mc to $N = 291,273$ unrelated white British individuals and $M = 5,670,959$ imputed SNPs (MAF $> 0.1\%$ and present in 1000 Genomes Project). SNPs were partitioned based on 28 functional annotations that were defined in a previous study [21]. We grouped 22 traits in the UK Biobank into five categories (autoimmune, diabetes, respiratory, anthropometric, and cardiovascular). Black bars mark ±2 standard errors centered on estimated enrichment. Annotations are ordered by the proportions of SNPs in that annotation (given in parentheses)


Smoking Status
Per-allele effect size squared
2e-05
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Height
Per-allele effect size squared
2e-05
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
BMI
Per-allele effect size squared
1.5e-05
1.0e-05
5.0e-06
0.0e+00
-5.0e-06
-1.0e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Waist-hip Ratio
Per-allele effect size squared
1e-05
5e-06
0e+00
-5e-06
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4


Systolic Blood Pressure
Per-allele effect size squared
1e-05
0e+00
-1e-05
-2e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Diastolic Blood Pressure
Per-allele effect size squared
1e-05
0e+00
-1e-05
-2e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Eczema
Per-allele effect size squared
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Asthma
Per-allele effect size squared
1e-05
0e+00
-1e-05
-2e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4


Hypertension
Per-allele effect size squared
3e-05
2e-05
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
High Cholesterol
Per-allele effect size squared
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Diabetes (any)
Per-allele effect size squared
1e-05
5e-06
0e+00
-5e-06
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Type 2 Diabetes
Per-allele effect size squared
1.5e-05
1.0e-05
5.0e-06
0.0e+00
-5.0e-06
0.001
0.010
0.100
MAF
LD
1
2
3
4


Hypothyroidism
Per-allele effect size squared
2e-05
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Thyroid
Per-allele effect size squared
2e-05
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Endocrine and Diabetes Diseases
Per-allele effect size squared
1e-05
0e+00
-1e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4
Cardiovascular Diseases
Per-allele effect size squared
2e-05
1e-05
0e+00
-1e-05
-2e-05
0.001
0.010
0.100
MAF
LD
1
2
3
4

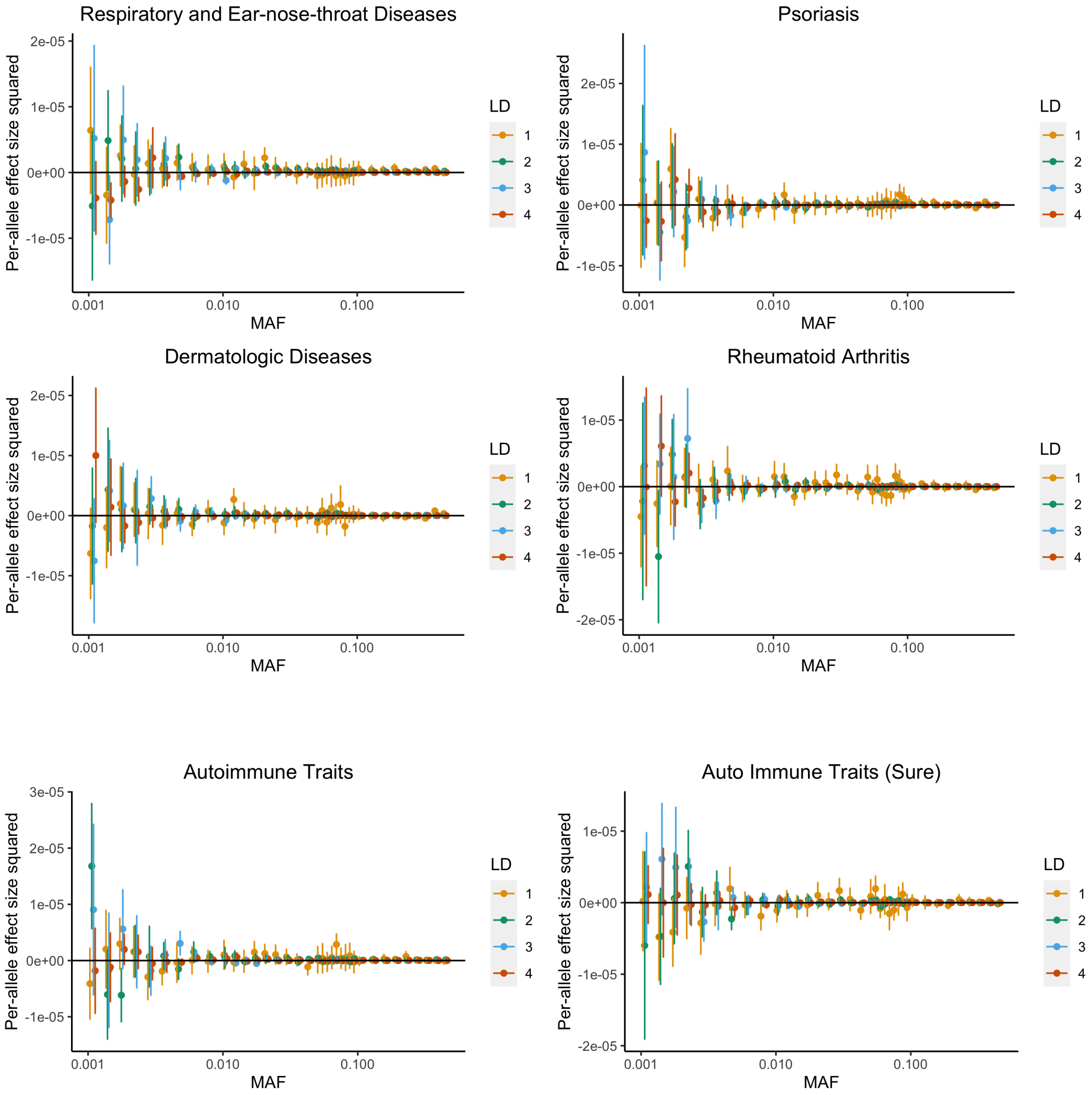


Figure 2.15: **Per-allele effect size squared of 22 traits as a function of MAF**: We applied RHE-mc to $N = 291,273$ unrelated white British individuals and $M = 7,774,235$ imputed SNPs. SNPs were partitioned into 144 bins based on LD score (4 bins based on quartiles of the LD score with i denoting the i^{th} quartile) and MAF (36 MAF bins) . Per allele heritability for bin k is defined as $\frac{h_k^2}{M_k * 2f_k * (1-f_k)}$ where h_k^2 is the heritability attributed to bin k, M_k is the number of SNPs in bin k, and f_k is the average MAF in bin k. Points represent estimated per-allele heritability. Bars mark ±2 standard errors centered on estimated per-allele heritability.

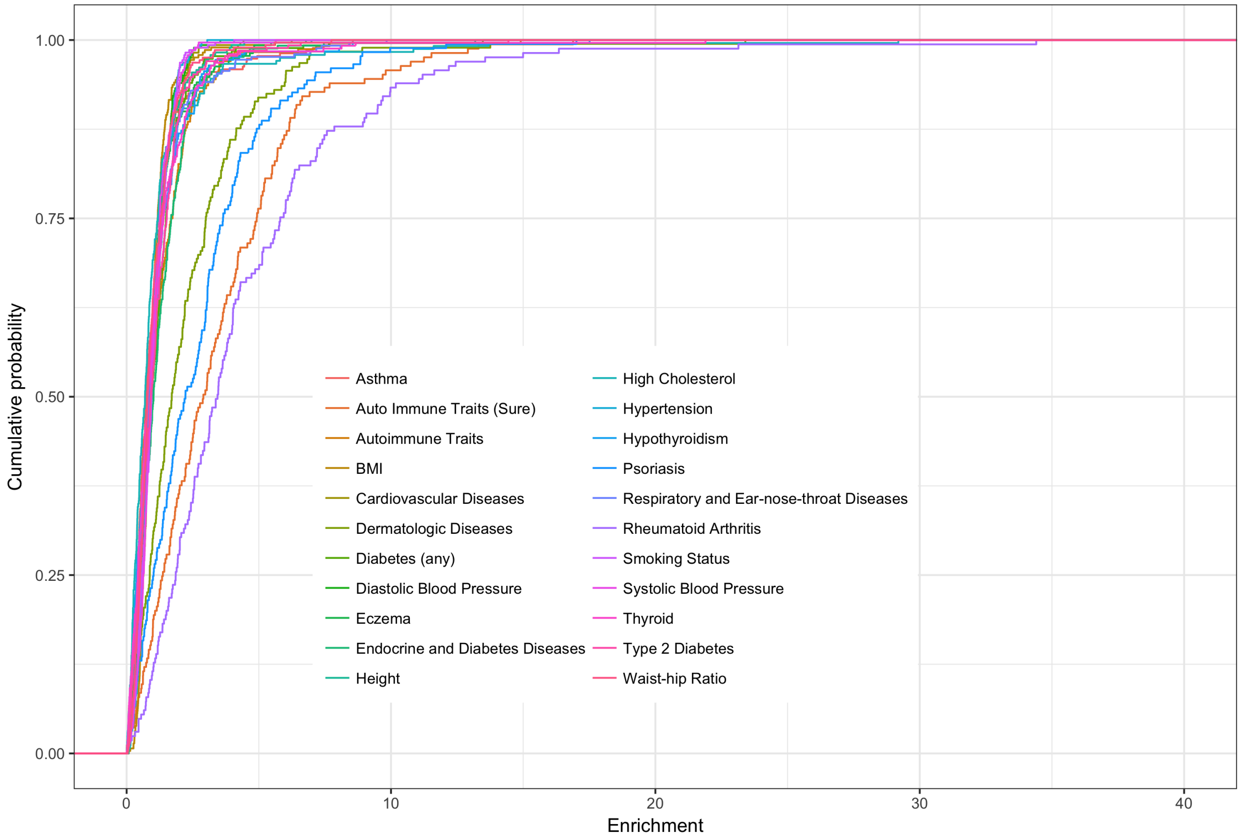


Figure 2.16: **Partitioning of genome-wide SNP heritability from RHE-mc for 22 complex traits and diseases in the UK Biobank ($N = 291,273$ unrelated white British individuals, $M = 459,792$ common SNPs with respect to 300 bins defined based on 10Mb base pairs**. Here we plot the empirical cumulative probability respectively of the enrichment.

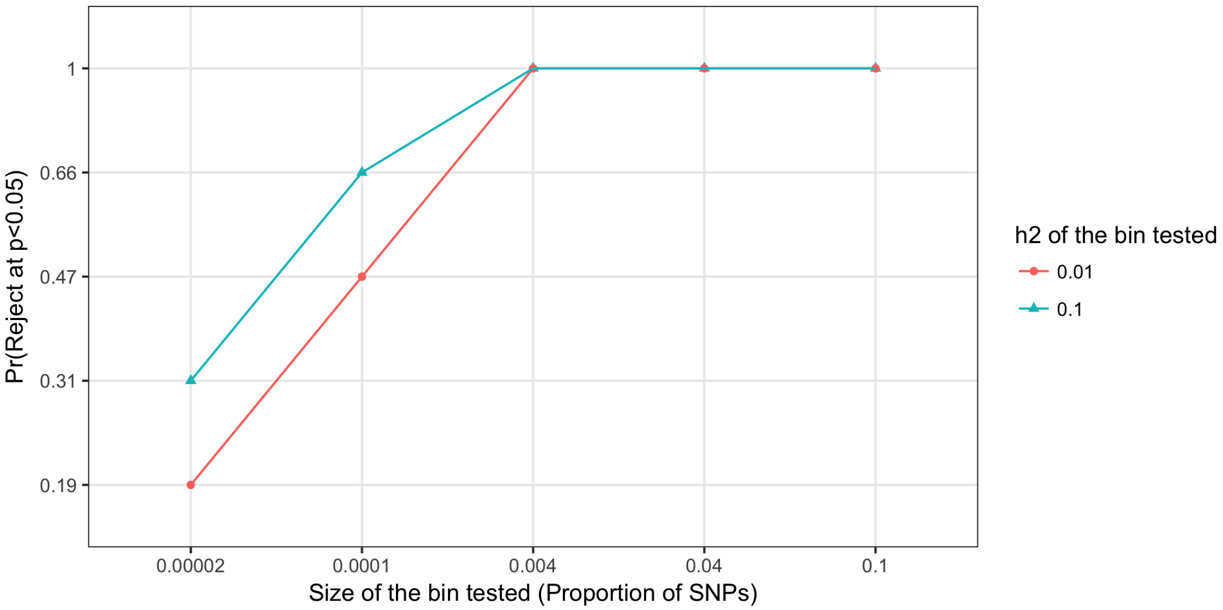


Figure 2.17: **Power as a function of annotation size**. Each point represents a rejection probability over 100 simulations. All simulations have $h^2_{total} = 0.7$, $N = 291,273$, $M = 459,792$, $p_{causal} = 0.05$.

2.7 Supplementary Tables

Number of random vectors	Point estimate (true SE)	Point estimate (SE of the estimator due to randomization)
10	0.24 (0.06)	0.24 (0.02)
100	0.24 (0.05)	0.25 (0.001)

Table 2.2: **Comparison of RHE-mc estimates with B=10 and B=100 on a small scale (M=590K array SNPs and N=10k individuals).** Here, we quantify the contribution of randomization to the SE of the estimator. The true total heritability is 0.25. We first computed the SE of RHE-mc for $B = 10$ and $B = 100$ from 100 simulation replicates (second column). We then computed the SE of the estimates (due to the randomization) for a single replicate. For $B = 10$, randomization contributes about a third of the total SE ($\frac{0.02}{0.06}$).

Genetic architecture			SE	
Percentage of causal SNPs	MAF of causal SNPs	MAF-LD coupling	True SE	Jackknife SE
0.01	[0.01,0.05]	a=b=0	0.012	0.013
0.01	[0.01,0.05]	a=0,b=1	0.018	0.015
0.01	[0.01,0.05]	a=0.75,b=0	0.016	0.015
0.01	[0.01,0.05]	a=0.75,b=1	0.013	0.013
0.01	[0.0,0.5]	a=b=0	0.009	0.013
0.01	[0.0,0.5]	a=0,b=1	0.016	0.014
0.01	[0.0,0.5]	a=0.75,b=0	0.018	0.019
0.01	[0.0,0.5]	a=0.75,b=1	0.012	0.015
0.01	[0.05,0.5]	a=b=0	0.012	0.015
0.01	[0.05,0.5]	a=0,b=1	0.021	0.017
0.01	[0.05,0.5]	a=0.75,b=0	0.014	0.014
0.01	[0.05,0.5]	a=0.75,b=1	0.015	0.017
1.0	[0.0,0.5]	a=b=0	0.007	0.007
1.0	[0.0,0.5]	a=0,b=1	0.007	0.007
1.0	[0.0,0.5]	a=0.75,b=0	0.006	0.006
1.0	[0.0,0.5]	a=0.75,b=1	0.007	0.008

Table 2.3: **Comparison of true SE with jackknife SE under 16 different genetic architectures**: We defined 100 blocks over SNPs to estimate block jackknife SE. We ran RHE-mc with 24 bins based on 6 MAF bins and 4 LDAK score bins. Jackknife SE yields estimates of true SE with relative bias -3% on average over 16 genetic architectures. True h^2 is fixed to 0.5 across all settings.

Genetic architecture		Heritability (GCTA-mc)		Heritability (RHE-mc)	
MAF of causal SNPs	MAF/LD coupling	Causal bin	Non-causal bin	Causal bin	Non-causal bin
$[0.01, 0.05]$	$a = b = 0$	0.244±0.061	0.009±0.051	0.242±0.064	0.004±0.052
$[0.01, 0.05]$	$a = 0, b = 1$	0.243±0.062	0.008±0.047	0.247±0.060	0.003±0.051
$[0.01, 0.05]$	$a = 0.75, b = 0$	0.241±0.061	0.009±0.050	0.240±0.062	0.002±0.051
$[0.01, 0.05]$	$a = 0.75, b = 1$	0.247±0.056	0.004±0.048	0.244±0.06	0.003±0.051
$[0.05, 0.5]$	$a = b = 0$	0.251±0.048	0.012±0.003	0.251±0.052	0.007±0.058
$[0.05, 0.5]$	$a = 0, b = 1$	0.248±0.052	0.014±0.054	0.240±0.049	0.001±0.055
$[0.05, 0.5]$	$a = 0.75, b = 0$	0.255±0.047	0.000±0.060	0.251±0.052	0.000±0.060
$[0.05, 0.5]$	$a = 0.75, b = 1$	0.250±0.048	0.005±0.05	0.241±0.050	0.002±0.058

Table 2.4: **Heritability contribution of causal bin vs non-causal bins on small-scale simulated data (**$M = 590k$ **array SNPs and** $N = 10k$ **individuals)**: We ran both RHE-mc and GCTA-mc with 24 bins based on 6 MAF bins and 4 LDAK bins for eight genetic architectures (Methods). In all simulations, the proportion of causal variants is 0.01, and the true total heritability is 0.25. The causal SNPs are restricted to lie within a specific range of MAF, *i.e.*, within $[0.01, 0.05]$ for the first four rows and $[0.05, 0.5]$ for the last four. Non-causal bins refer to those bins where none of the SNPs is causal, *i.e.*, in each of the first four genetic architectures. These would correspond to bins with MAF $\notin [0.01, 0.05]$. Causal bins refer to all remaining bins. Standard errors are computed from 100 replicates.

Genetic architecture		Heritability	
MAF of causal SNPs	MAF/LD coupling	Causal bin	Non-causal bin
$[0.01, 0.05]$	$a = b = 0$	0.501 ± 0.006	0.000 ± 0.004
$[0.01, 0.05]$	$a = 0, b = 1$	0.498 ± 0.007	0.000 ± 0.003
$[0.01, 0.05]$	$a = 0.75, b = 0$	0.500 ± 0.008	0.002 ± 0.004
$[0.01, 0.05]$	$a = 0.75, b = 1$	0.490 ± 0.007	0.001 ± 0.003
$[0.05, 0.5]$	$a = b = 0$	0.501 ± 0.036	-0.001 ± 0.030
$[0.05, 0.5]$	$a = 0, b = 1$	0.487 ± 0.012	0.005 ± 0.005
$[0.05, 0.5]$	$a = 0.75, b = 0$	0.508 ± 0.026	-0.005 ± 0.023
$[0.05, 0.5]$	$a = 0.75, b = 1$	0.490 ± 0.009	0.000 ± 0.005

Table 2.5: **Heritability contribution of causal vs non-causal bins on large-scale simulated data ($M = 590K$ array SNPs and $N = 337K$ individuals)**: We ran RHE-mc with 24 bins based on 6 MAF bins and 4 LDAK bins (Methods). Standard errors are computed from 100 replicates. The percentage of causal SNPs and true h^2 are fixed to 0.01 and 0.5, respectively, across all the settings.

MAF bin	Range	Number of SNPs
1	[0, 0.0126)	59330
2	[0.0126, 0.020)	59330
3	[0.020, 0.029)	59330
4	[0.029, 0.0433)	59330
5	[0.043, 0.0658)	59330
6	[0.065, 0.106)	59330
7	[0.106, 0.170)	59330
8	[0.170, 0.260)	59330
9	[0.260, 0.373)	59330
10	[0.373, 0.5)	59330

Table 2.6: **MAF bins which are used in running S-LDSC over the large scale simulated data.**

True enrichment	Proportion of SNPs	point estimate	SE	Pr(rejection at p< 0.05)
2	0.4%	2.06	0.4	100%
1	0.4%	1.02	0.14	100%
0	0.4%	0.0	0.02	0.5%
2	0.01%	2.18	1.07	30%

Table 2.7: **Power as a function of annotation size**. SE, point estimate, and probability of rejections are computed from 100 replicates. All simulations have $h^2_{total} = 0.7$, $N = 291,273$, $M = 459,792$, $p_{causal} = 0.05$.

Trait	Heritability		
	Chromosome	MAF/LD	10Mb
Autoimmune Traits	0.064 ± 0.005	0.054 ± 0.006	0.070 ± 0.004
Auto Immune Traits (Sure)	0.011 ± 0.002	0.023 ± 0.006	0.029 ± 0.001
Dermatologic Diseases	0.020 ± 0.003	0.0172 ± 0.003	0.021 ± 0.001
Psoriasis	0.017 ± 0.002	0.014 ± 0.005	0.022 ± 0.003
Rheumatoid Arthritis	0.008 ± 0.002	0.008 ± 0.002	0.010 ± 0.003
Eczema	0.124 ± 0.007	0.104 ± 0.005	0.13 ± 0.006
Hypothyroidism	0.097 ± 0.008	0.081 ± 0.005	0.11 ± 0.007
Thyroid	0.095 ± 0.009	0.081 ± 0.008	0.109 ± 0.008
Diastolic Blood Pressure	0.170 ± 0.005	0.145 ± 0.004	0.173 ± 0.003
Systolic Blood Pressure	0.172 ± 0.006	0.146 ± 0.004	0.171 ± 0.004
Cardiovascular Diseases	0.165 ± 0.006	0.134 ± 0.005	0.17 ± 0.004
Hypertension	0.179 ± 0.006	0.150 ± 0.005	0.183 ± 0.006
High Cholesterol	0.099 ± 0.015	0.070 ± 0.008	0.102 ± 0.003
Diabetes (any)	0.069 ± 0.004	0.058 ± 0.003	0.072 ± 0.003
Endocrine and Diabetes Diseases	0.064 ± 0.004	0.053 ± 0.003	0.065 ± 0.003
Type 2 Diabetes	0.068 ± 0.004	0.057 ± 0.003	0.069 ± 0.005
BMI	0.330 ± 0.014	0.264 ± 0.007	0.328 ± 0.013
Height	0.583 ± 0.026	0.492 ± 0.017	0.59 ± 0.021
Waist-hip Ratio	0.196 ± 0.009	0.167 ± 0.007	0.2 ± 0.005
Asthma	0.122 ± 0.009	0.101 ± 0.006	0.127 ± 0.007
Smoking Status	0.130 ± 0.004	0.111 ± 0.003	0.132 ± 0.002
Respiratory and Ear-nose-throat Diseases	0.086 ± 0.007	0.071 ± 0.004	0.091 ± 0.004

Table 2.8: **Estimates of genome-wide SNP heritability from RHE-mc for 22 complex traits and diseases in the UK Biobank ($N = 291,273$ unrelated white British individuals, $M = 459,792$ common SNPs)**. We run RHE-mc with 8 bins defined based on two MAF bins (MAF≤ 0.05, MAF> 0.05) and quartiles of the LD-scores. Furthermore, we run RHE-mc with 22 bins defined based on chromosome number. On average, partitioning based on chromosome numbers leads 21% higher estimates of genome-wide SNP heritability for 22 traits than partitioning based on MAF and LD. For instance, it leads 18% and 13% higher estimates of heritability for height and BMI respectively. We also partitioned SNP based on 10 Mb genomic regions (300 variance components).

CHAPTER 3

Gene-by-Environment interactions effects

3.1 Background

Understanding the contribution of additive and non-additive genetic effects to complex trait variation is a central question in human genetics. A possible source of non-additive effects arises from interactions between genetic and environmental factors. These gene-environment interactions (GxE) have been investigated at the level of individual genetic variants in a number of studies. In functional genomics, the effect of variants that on gene expression have been found to be modulated by environmental factors such as age, tissue, cell type, or other genetic variants[1, 32, 53]. In the context of complex traits, GxE has been observed for specific variants and exposures like lifestyle factors [138, 135, 94, 73], air pollution [19], and microbe exposure[49]. While these studies have provided insights into novel mechanisms and pathways underlying trait variation, the small effects of individual genetic variants do not allow us to quantify the overall contribution of GxE to variation to a complex trait.

GxE are also important in understanding sources of heritability, *i.e.*, the maximal proportion of variation in a trait that can be explained by genetic variation [118]. While there has been substantial attention focused on estimating narrow-sense heritability from genome-wide SNP genotype data (termed SNP heritability) [132], understanding the factors that contribute to the gap between SNP heritability and heritability estimates from family studies remains an area of active research with GxE effects being a plausible explanation for this gap. Recent work [74] suggests that genetic predictors of complex trait do not generalize even within an ancestry group, in part, due to the role of GxE. Finally, the availability of large data sets like the UK biobank which contains $\approx 300,000$ individuals, millions SNPs and hundreds of environmental factors offers the opportunity to systematically understand the role of GxE.

Linear mixed models (LMMs) have emerged as a powerful tool to estimate SNP heritability by aggregating the effects of a large number of SNPs [133]. In these models, the model parameters, *i.e.*, variance components, associated with additive genetic effects are related to the SNP heritability. More recently, LMMs have also been extended to estimate GxE [52, 12] by jointly fitting variance components associated with additive genetic effects and GxE. The most common approach to estimating variance components attempts to maximize the likelihood or the restricted maximum likelihood (REML) [79, 133] (termed genome-based REML or GREML when applied to SNP genotypes). Computing GREML estimators can be challenging even for estimating only the additive genetic variance components. Many of the efficient algorithms for computing GREML estimates leverage the specific structure of LMMs with a single additive genetic variance component [60, 146, 64] and cannot be easily extended to the setting where we have additive and GxE variance components.

3.2 Materials and Methods

We consider the problem of inferring the proportion of variance in a trait that can be explained by GxE given genotypes collected from N individuals across M SNPs and L environmental variables.

Specifically, we first consider a model that aims to estimate GxE across all of the L environmental variables and M SNPs. This setting is applicable when a large number of environmental variables have been measured (*e.g.*, features extracted from brain MRI images or questionnaire data) and it is unclear which of these contributes to GxE. In this setting, we propose a randomized method-of-moments (MoM) variance components estimation algorithm to jointly estimate the additive genetic variance component and the GxE variance components. Our MoM estimator is valid under general distributions on the effect sizes and residual error unlike REML which requires an assumption of normally distributed effect sizes and normally distributed residual error. On the other hand, our MoM estimator is less statistically efficient relative to REML. Importantly, our randomized MoM algorithm is computationally efficient. Our proposed algorithm has runtime complexity $\mathcal{O}(N(M+L)B)$ for N individuals, M SNPs, L environments and a parameter B that controls the number of random matrix-vector multiplications where $B << M$. Further, the randomized MoM estimator allows estimation is a streaming algorithm (it requires making a single pass over the genotype and the environmental variables) and does not explicitly compute the GxE matrix leading to a substantial memory efficiency.

While the previous model is useful in quantifying the total GxE, it is often of interest to identify and interpret specific environmental variables that are involved in GxE. We extend our model to this setting, where we are fitting a GxE variance component for each environmental variable.

3.2.1 Single-component GxE model

Let $\boldsymbol{X}$ denote a $N \times M$ genotype matrix, $\boldsymbol{E}$ denote a $N \times L$ matrix of environmental variables, and $\boldsymbol{y}$ denote a N-vector of phenotypes. Define $\boldsymbol{H} = \boldsymbol{X} \odot \boldsymbol{E}$ as the $N \times ML$ matrix formed by taking products of a column in $\boldsymbol{X}$ with each column in $\boldsymbol{E}$. We assume the following model:

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{H\alpha} + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{D}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N) \\
\boldsymbol{\beta} &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_g^2}{M} \boldsymbol{I}_M) \\
\boldsymbol{\alpha} &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_{ge}^2}{M} \boldsymbol{I}_{ML})
\end{aligned} \tag{3.1}$$

Here $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an arbitrary distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In this model, σ_e^2, σ_g^2, and σ_{ge}^2 are the residual variance, genetic variance and gene-by-environment variance components respectively. Here $\boldsymbol{\beta}$ denotes M-vector of SNPs effect sizes and $\boldsymbol{\alpha}$ denotes ML-vector of GxE effect sizes.

We assume without loss of generality that $\boldsymbol{y}$ is centered and the columns of $\boldsymbol{X}$ and $\boldsymbol{H}$ are standardized. To estimate the variance components of our LMM, we use a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments. Since $\mathbb{E}\left[\boldsymbol{y}\right] = 0$, we derived the MoM estimates by equating the population covariance to the empirical covariance. The population covariance is given by:

$$cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sigma_g^2 \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \sigma_{ge}^2 \frac{1}{ML}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} + \sigma_e^2 \boldsymbol{I} \tag{3.2}$$

Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to estimate the variance parameters :

$$(\tilde{\sigma_g^2}, \tilde{\sigma_{ge}^2}, \tilde{\sigma_e^2}) = argmin_{(\sigma_g^2, \sigma_{ge}^2, \sigma_e^2)} ||f(\boldsymbol{y})f(\boldsymbol{y})^T - \left(\sigma_g^2 \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \sigma_{ge}^2 \frac{1}{ML}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} + \sigma_e^2 \boldsymbol{I}\right)||_F^2 \tag{3.3}$$

The MoM estimator satisfies the normal equations:

$$\begin{bmatrix} \frac{1}{M^2}tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) & \frac{1}{M^2L}tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) & \frac{1}{M}tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) \\ \frac{1}{M^2L}tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) & \frac{1}{(ML)^2}tr(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) & \frac{1}{ML}tr(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) \\ \frac{1}{M}tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) & \frac{1}{ML}tr(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) & N \end{bmatrix} \begin{bmatrix} \tilde{\sigma_g^2} \\ \tilde{\sigma_{ge}^2} \\ \tilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{M}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \frac{1}{ML}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix} \tag{3.4}$$

In this work, we explore randomized estimators that permit efficient computation of the entries of the linear system in Equation 3.4. Here we propose an unbiased estimator of the term $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$ that can be computed in time $\mathcal{O}(N(M+L)B)$ which is linear in sizes of both genotype and environment matrices (In Section 3.5.2 of the Supplementary Notes, we describe an alternate estimator that can be computed in time $\mathcal{O}(NMLB)$ time complexity that might be practical for a small number of SNPs or environmental variables).

Approximate computation in $\mathcal{O}(N(M+L)B)$ time

The computational bottleneck in solving system 3.4 is the evaluation of $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$. We propose an unbiased randomized estimator to compute $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$ with $\mathcal{O}(N(M+L)B)$

time complexity for a parameter B. Given B, we form $2B$ random vectors: $\boldsymbol{w}_b = \boldsymbol{u}_b \otimes \boldsymbol{v}_b$ where $\boldsymbol{u}_b \in \mathbb{R}^M, \boldsymbol{v}_b \in \mathbb{R}^L$ are independent random vectors with mean zero and covariance $\boldsymbol{I}_M$ and $\boldsymbol{I}_L$ respectively. Let $\boldsymbol{r}_b = \boldsymbol{H}\boldsymbol{w}_b$, $b \in \{1, \dots, B\}$. Then our estimator Γ is defined as:

$$\Gamma \equiv \frac{1}{B}\sum_{b=1}^{B}\left(r_{2b-1}^T r_{2b}\right)^2 \tag{3.5}$$

We will show in Theorem 1 that Γ is an unbiased estimator of $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$. We also propose an alternate randomized estimator that also has $\mathcal{O}(N(M+L)B)$ complexity in Section 3.5.3 in the Supplementary Notes. While our software implements both estimators, we focus on the estimator that we described in this section, which is better suited for a streaming implementation and forms the basis of all our results.

Lemma 1 *Suppose that $\boldsymbol{X}$ and $\boldsymbol{E}$ are $N \times M$ genotype and $N \times L$ environment matrices respectively. Define $\boldsymbol{H} = \boldsymbol{X} \odot \boldsymbol{E}$ as the $N \times ML$ matrix. Assume that $\boldsymbol{w}_t = \boldsymbol{u}_t \otimes \boldsymbol{v}_t$ where $\boldsymbol{u}_t \in \mathbb{R}^M, \boldsymbol{v}_t \in \mathbb{R}^L$ are independent random vectors with mean zero and covariance $\boldsymbol{I}_M$ and $\boldsymbol{I}_L$, $t \in \{1,2\}$ respectively. Let $\boldsymbol{r}_t = \boldsymbol{H}\boldsymbol{w}_t$, $t \in \{1,2\}$. Then $(\boldsymbol{r}_1^T\boldsymbol{r}_2)^2$ is an unbiased estimator of $tr(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{H})$.*

Proof:

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{r}_1^{\mathrm{T}}\boldsymbol{r}_2)^2\right] &= \mathbb{E}\left[(\boldsymbol{r}_1^{\mathrm{T}}\boldsymbol{r}_2)(\boldsymbol{r}_2^{\mathrm{T}}\boldsymbol{r}_1)\right] \\
&= \mathbb{E}\left[\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_2\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_1\right] \\
&= \mathbb{E}\left[tr(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_2\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_1)\right] \\
&= \mathbb{E}\left[tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_2\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_1\boldsymbol{w}_1^{\mathrm{T}})\right] \qquad \text{(cyclic property of trace)} \\
&= tr(\mathbb{E}\left[\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_2\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_1\boldsymbol{w}_1^{\mathrm{T}}\right]) \qquad \text{(linearity of trace and expectation)} \\
&= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\mathbb{E}\left[\boldsymbol{w}_2\boldsymbol{w}_2^{\mathrm{T}}\right]\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\mathbb{E}\left[\boldsymbol{w}_1\boldsymbol{w}_1^{\mathrm{T}}\right]) \qquad \text{(independence of } \boldsymbol{w}_1, \boldsymbol{w}_2\text{)} \\
&= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}) \qquad (\mathbb{E}\left[\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}}\right] = \boldsymbol{I}, t \in \{1,2\})
\end{aligned}$$

□

Lemma 2 *Suppose that $\boldsymbol{X}$ and $\boldsymbol{E}$ are $N \times M$ and $N \times L$ matrices respectively. Define $\boldsymbol{H} = \boldsymbol{X} \odot \boldsymbol{E}$ be an $N \times ML$ matrix. Define $\boldsymbol{w} = \boldsymbol{u} \otimes \boldsymbol{v}$ where $\boldsymbol{u} \in \mathbb{R}^M, \boldsymbol{v} \in \mathbb{R}^L$ are two arbitrary vectors. $\boldsymbol{Hw}$ can be computed in $\mathcal{O}(N(M+L))$.*

Proof: Denoting $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} \end{pmatrix}$, $\boldsymbol{E} = \begin{pmatrix} \boldsymbol{e}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{e}_N^{\mathrm{T}} \end{pmatrix}$, where $\boldsymbol{x}_n, n \in \{1, \ldots, N\}$ are M-vectors and $\boldsymbol{e}_n, n \in \{1, \ldots, N\}$ are L-vectors, we have:

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{h}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{h}_N^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{x}_1 \otimes \boldsymbol{e}_1)^{\mathrm{T}} \\ \vdots \\ (\boldsymbol{x}_N \otimes \boldsymbol{e}_N)^{\mathrm{T}} \end{pmatrix}$$

$$\begin{aligned} \boldsymbol{h}_n^{\mathrm{T}}\boldsymbol{w} &= (\boldsymbol{x}_n \otimes \boldsymbol{e}_n)^{\mathrm{T}} (\boldsymbol{u} \otimes \boldsymbol{v}) \\ &= \left(\boldsymbol{x}_n{}^{\mathrm{T}} \otimes \boldsymbol{e}_n{}^{\mathrm{T}}\right) (\boldsymbol{u} \otimes \boldsymbol{v}) \\ &= \left(\boldsymbol{x}_n{}^{\mathrm{T}}\boldsymbol{u}\right) \otimes \left(\boldsymbol{e}_n{}^{\mathrm{T}}\boldsymbol{v}\right), \quad \text{Using the mixed-product property of } \otimes \\ &= (\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{u})(\boldsymbol{e}_n^{\mathrm{T}}\boldsymbol{v}) \equiv b_n c_n \end{aligned} \tag{3.6}$$

Thus, we have:

$$\begin{aligned} \boldsymbol{Hw} &= \begin{pmatrix} \boldsymbol{h}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{h}_N^{\mathrm{T}} \end{pmatrix} \boldsymbol{w} = \begin{pmatrix} \boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{w} \\ \vdots \\ \boldsymbol{h}_N^{\mathrm{T}}\boldsymbol{w} \end{pmatrix} \\ &= \begin{pmatrix} b_1 c_1 \\ \vdots \\ b_N c_N \end{pmatrix}, \qquad \text{From Equation 3.6} \\ &= \boldsymbol{b} \odot \boldsymbol{c} \end{aligned}$$

where $\boldsymbol{b} = \boldsymbol{Xu}$, $\boldsymbol{c} = \boldsymbol{Ev}$. Thus, we can compute $\boldsymbol{r}$ in $\mathcal{O}(NM + NL)$. □

Theorem 1 *Γ is an unbiased estimator of $tr(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{H})$ that can be computed in $\mathcal{O}(N(M+L)B)$.*

Proof: Since each term $(r_{2b-1}r_{2b})^2$ is an unbiased estimator of $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$ (Lemma 1), Γ is an unbiased estimator of $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$ (with variance $\frac{1}{B}$ times that of each term).

Since $\boldsymbol{r}_b = \boldsymbol{H}\boldsymbol{w}_b$ can be computed in $\mathcal{O}(N(M+L))$ (Lemma 2), each term of Γ can be computed in time $\mathcal{O}(N(M+L))$ so that Γ can be computed in $\mathcal{O}(N(M+L)B)$. □

The additional terms that involve $\boldsymbol{H}$ in Equation 3.4 are $tr(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})$, $tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})$, and $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}$. We use the property that $\boldsymbol{r}_t{}^{\mathrm{T}}\boldsymbol{r}_t$ is an unbiased estimator of $tr(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})$, $(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{r}_t)^2$ is an unbiased estimator of $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}$ and $\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}_t\|_2^2$ is an unbiased estimator of $tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})$ to obtain analogous estimators for each of these quantities. Given r_t, each of these estimators can be computed in time $\mathcal{O}(N)$ and $\mathcal{O}(NM)$ respectively (see Lemma 6 in Supplementary Notes 3.5.4) .

3.2.2 Multi-component GxE model

Let $\boldsymbol{X}$ denote a $N \times M$ genotype matrix, $\boldsymbol{E}$ denote a $N \times L$ matrix of environmental variables, $\boldsymbol{C}$ denote a $N \times P$ matrix of fixed-effect covariates, and $\boldsymbol{y}$ denote a N-vector of phenotypes. We assume the following linear mixed model:

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \sum_{l=1}^{L}(\boldsymbol{X} \odot \boldsymbol{E}_{:l})\boldsymbol{\alpha}_l + \sum_{l=1}^{L}(\boldsymbol{I}_N \odot \boldsymbol{E}_{:l})\boldsymbol{\delta_l} + \boldsymbol{C}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\
\boldsymbol{\beta} &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_g^2}{M}\boldsymbol{I}_M) \\
\boldsymbol{\alpha}_l &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_{gxe,l}^2}{M}\boldsymbol{I}_M) \\
\boldsymbol{\delta}_l &\sim \mathcal{D}(\boldsymbol{0}, \sigma_{nxe,l}^2\boldsymbol{I}_N) \\
\boldsymbol{\epsilon} &\sim \mathcal{D}(\boldsymbol{0}, \sigma_e^2\boldsymbol{I}_N)
\end{aligned} \tag{3.7}$$

Here $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes an arbitrary distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\boldsymbol{E}_{:l}$ denotes l-th column of $\boldsymbol{E}$, and $\odot$ denotes row-wise Kronecker product. $\boldsymbol{\beta}$ denotes the M-vector of SNP effect sizes, $\boldsymbol{\alpha}_l$ denotes the M-vector of genetic effect sizes in the context of environment l (GxE effects) while $\boldsymbol{\delta}_l$ denotes the N-vector of NxE effect sizes for environment l, and $\boldsymbol{\epsilon}$ denotes the N-vector of noise. σ_e^2, σ_g^2, $\sigma_{gxe,l}^2$, and $\sigma_{nxe,l}^2$ denote the residual variance, additive genetic, gene-by-environment, and noise-by-environment variance components respectively. These variance components can then be transformed into the additive heritability or the proportion of variance explained by additive effects (h_g^2 associated with σ_g^2) and the GxE heritability or the proportion of variance explained by interactions of genetics with a given environment ($h^2_{gxe,l}$ associated with $\sigma^2_{gxe,l}$).

We assume without loss of generality that $\boldsymbol{y}$ is centered and the columns of $\boldsymbol{X}$ and $\boldsymbol{E}$ are standardized. To estimate the variance components of our LMM, we use a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments. Since $\mathbb{E}\left[\boldsymbol{y}\right] = 0$, we derived the MoM estimates by equating the population covariance to the empirical covariance. For simplicity, we exclude the matrix of covariates $\boldsymbol{C}$ from the model in the following derivation as the covariates can be efficiently projected out of the phenotype, genotypes, and interaction terms with minimal additional cost.

For compactness, we denote $\boldsymbol{Z}_0 = \boldsymbol{X}$, $\boldsymbol{Z}_l = \boldsymbol{X} \odot \boldsymbol{E_l}$ for $l = 1, .., L$, $\boldsymbol{Z}_l = \boldsymbol{I_N} \odot \boldsymbol{E_{:l}}$ for $l = L+1, .., 2L$, and $\boldsymbol{Z}_{2L+1} = \boldsymbol{I}_N$. The population covariance is given by:

$$cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sum_{l=0}^{2L+1} \sigma_l^2 \boldsymbol{K}_l \tag{3.8}$$

where

$$\boldsymbol{K}_l = \begin{cases} \dfrac{\boldsymbol{Z}_l\boldsymbol{Z}_l^T}{M}, & l = 0, .., L \\ \boldsymbol{Z}_l\boldsymbol{Z}_l^T, & l = L+1, .., 2L+1 \end{cases}$$

and

$$\sigma_l^2 = \begin{cases} \sigma_g^2, & l = 0 \\ \sigma^2_{gxe,l}, & l = 1, .., L \\ \sigma^2_{nxe,l}, & l = L+1, .., 2L \\ \sigma_e^2, & l = 2L+1 \end{cases}$$

Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to find the variance components.

$$\tilde{\boldsymbol{\sigma^2}} = argmin_{\boldsymbol{\sigma^2}} ||\boldsymbol{y}\boldsymbol{y}^T - \sum_{l=0}^{2L+1} \sigma_l^2 \boldsymbol{K}_l||_F^2 \tag{3.9}$$

The MoM estimator satisfies the following normal equations:

$$\boldsymbol{T} = \boldsymbol{\sigma^2 q} \tag{3.10}$$

where $\boldsymbol{T}$ is matrix with entries $T_{ij} = tr(\boldsymbol{K}_i\boldsymbol{K}_j), i, j \in \{0, .., 2L+1\}$, and $\boldsymbol{q}$ and $\tilde{\boldsymbol{\sigma^2}}$ are vectors with entries $c_l = \boldsymbol{y}^T\boldsymbol{K}_l\boldsymbol{y}$, for $l \in \{0, .., 2L+1\}$.

Given estimated $\tilde{\boldsymbol{\sigma^2}}$, the heritability associated with component i for a component that represents additive genetic or GxE effects (equivalently, the proportion of variance explained by component i is defined as follows:

$$h_i^2 = \frac{\hat{\sigma}_i^2 tr(\boldsymbol{K}_i)}{\sum_k \hat{\sigma}_k^2 tr(\boldsymbol{K}_k)} \tag{3.11}$$

The aforementioned definition of heritability holds when the matrices $\boldsymbol{Z}$'s columns have zero means, and N is large. To explicitly ensure that the columns of GxE matrices also have zero means, a column consisting of all ones is included in the covariate matrix. Consequently, when the covariates are projected out of the GxE matrices, it guarantees that all columns have zero means.

Computing the coefficients of the system of linear equation 3.10 presents computational challenges. The main computational bottleneck is the evaluation of the quantities T_{ij} for $i, j \in \{0, \ldots, 2L+1\}$ which requires $\mathcal{O}(N^2M)$. Therefore, the total time complexity for exact MoM is $\mathcal{O}(N^2ML + L^3)$ imposing challenging memory or computation requirements for Biobank-scale data (N in the hundreds of thousands, M in the millions, L in the hundreds or thousands).

Instead of computing the exact value of T_{ij}, GENIE uses a randomized estimator of the trace [48]. This estimator uses the fact that for a given $N \times N$ matrix $\boldsymbol{C}$, $\boldsymbol{w}^T\boldsymbol{C}\boldsymbol{w}$ is an unbiased estimator of $tr(\boldsymbol{C})$ ($E[\boldsymbol{w}^T\boldsymbol{C}\boldsymbol{w}] = tr[\boldsymbol{C}]$) where $\boldsymbol{w}$ be a random vector with mean zero and covariance $\boldsymbol{I}_N$. Hence, we can estimate the values T_{ij}, $i, j \in \{0, \ldots, 2L+1\}$ as follows:

$$T_{ij} = tr(\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{Z}_j^T) \approx \widehat{T_{ij}} = \frac{1}{B}\sum_b \boldsymbol{w}_b^T\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{Z}_j^T\boldsymbol{w}_b \tag{3.12}$$

Here $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_B$ are B independent random vectors with zero mean and covariance $\boldsymbol{I}_N$. In GENIE, we draw these random vectors independently from a standard normal distribution. Note that computing T_{ij} by using the above estimator involves matrix-vector multiplications which are repeated B times. Therefore, the total running time is $\mathcal{O}(LNMB)$.

Moreover, we can leverage the structure of the genotype matrix which only contains entries in $\{0,1,2\}$. For a fixed genotype matrix $\boldsymbol{X}_k$, we can improve the per iteration time complexity of matrix-vector multiplication from $\mathcal{O}(NM)$ to $\mathcal{O}(\frac{NM}{max(\log_3(N),\log_3(M))})$ by using the Mailman algorithm [59]. Solving the normal equations takes $\mathcal{O}(L^3)$ time so that for a small number of components (L), the overall time complexity of our algorithm is $\mathcal{O}(\frac{LNMB}{\max(\log_3(N),\log_3(M))} + L^2NB + L^3)$.

Although the model defined in Equation 3.7 is beneficial in quantifying the total GxE effects for a given E, it is interesting to identify and interpret the interaction of E with specific regions of the genome, such as SNPs with a particular range of minor allele frequencies or SNPs that lie within genes expressed specifically in a tissue. Following our previous work [81], the genotype component $\boldsymbol{X}$ can be assigned to T (potentially overlapping) components with respect to a set of annotations (such as MAF/LD or functional annotations). Thus, we extend our model as follows:

$$\begin{aligned}
\boldsymbol{y} &= \sum_{t=1}^{T}\boldsymbol{X}_t\boldsymbol{\beta}_t + \sum_{t=1}^{T}\sum_{l=1}^{L}(\boldsymbol{X}_t \odot \boldsymbol{E}_{:l})\boldsymbol{\alpha}_{tl} + \sum_{l=1}^{L}(\boldsymbol{I}_N \odot \boldsymbol{E}_{:l})\boldsymbol{\delta}_l + \boldsymbol{C}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\
\boldsymbol{\beta}_t &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma^2_{g,t}}{M}\boldsymbol{I}_M) \\
\boldsymbol{\alpha}_{tl} &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma^2_{gxe,tl}}{M}\boldsymbol{I}_M) \\
\boldsymbol{\delta}_l &\sim \mathcal{D}(\boldsymbol{0}, \sigma^2_{nxe,l}\boldsymbol{I}_N) \\
\boldsymbol{\epsilon} &\sim \mathcal{D}(\boldsymbol{0}, \sigma^2_e\boldsymbol{I}_N)
\end{aligned} \tag{3.13}$$

Here $\boldsymbol{X}_t$ is the genotype of annotation t with M_t SNPs, $\boldsymbol{\alpha}_{tl}$ refers to the effect sizes of SNPs in annotation t in the context of environment l. Analogously, $\sigma^2_{gxe,tl}$ refers to the variance component for SNPs in annotation t in the context of environment l while $h^2_{gxe,tl}$ refers to the GxE heritability associated with annotation t in the context of environment l.

Given estimated GxE heritabilties under the above model, we define the enrichment of genetic effects in annotation t in the context of environment l (also termed GxE enrichment) as follows :

$$Enrichment(gxe,t,l) = \frac{h^2_{gxe,tl}/\sum_{t=1}^{T}(h^2_{gxe,tl})}{M_t/M}, t \in \{1,\ldots,T\}, l \in \{1,\ldots,L\} \tag{3.14}$$

GENIE uses the same randomized trace estimator approach to efficiently estimate GxE heritability.

3.3 Results

3.3.1 Calibration and power

We assess the false positive rate of tests of GxE heritability based on GENIE in simulations under different genetic architectures with no GxE heritability. For each architecture, we simulated 100 phenotype replicates across $N = 291,273$ unrelated white British individuals in the UKBB and $M = 459,792$ SNPs with MAF $> 1\%$ genotyped on the UK Biobank genotyping array. We chose statin usage in the UKBB as the environmental variable. We varied the percentage of causal SNPs while fixing the additive heritability at $h^2_g = 0.25$. We ran GENIE with $B = 10$ random vectors (see Section on Effect of the choice of the number of random vectors).

Across all simulations, the false positive rate of rejecting the null hypothesis of no GxE heritability is controlled at levels 0.05 and 0.05/200 (we consider this threshold which controls for the number of trait-E pairs that we test in UKBB): the average $P(\text{rejection at } p < t)$ is 4% and 0% for $t = 0.05$ and $t = 0.05/200$ respectively (Figure 3.1a)).

To measure the power of GENIE to detect GxE heritability, we simulated phenotypes with a non-zero GxE heritability. Across genetic architectures, we varied the GxE heritability while fixing the additive heritability at 0.25 and the percentage of causal SNPs at 10% (these are the default parameters of our simulations unless otherwise specified). We simulated 100 replicates for every genetic architecture. Let $h^2_{gxe}(i)$ be the estimate of h^2_{gxe} and SE_i be the jackknife estimate of the standard error on the i-the replicate for $i \in \{1, .., 100\}$. We computed the p-value of a test of the null hypothesis of no h^2_{gxe} on the i-th replicate from the Z-score defined as $h^2_{gxe}(i)/SE_i$ for $i \in \{1, .., 100\}$. We reported the percentage of replicates with p-value$< t$ as the power of GENIE on a given genetic architecture for a p-value threshold of t.

GENIE has adequate power to detect GxE effects with $h^2_{gxe} \geq 0.02$ in a sample of $300K$ unrelated individuals at $p < 0.05$ (Figure 3.1b)). Additionally, across all genetic architectures, GENIE yields unbiased estimates of GxE heritability (Figure 3.1c)).

Next, we assessed the accuracy of GENIE in a setting where we have multiple environmental variables. We simulated phenotypes from a sub-sampled set of UKBB genotypes choosing a subset of $N = 10,000$ individuals and $20,000$ SNPs on chromosome 1 of the UK Biobank Axiom array. We considered a setting with $L = 10$ environmental variables with $\sigma^2_g = 0.2$, five environmental variables with $\sigma^2_{gxe} = 0$, three environmental variables with $\sigma^2_{gxe} = 0.1$ and two with $\sigma^2_{gxe} = 0.01$. We generated 100 replicates of simulated phenotypes for each set of parameters. We find that GENIE obtains estimates of h^2_{gxe} that are accurate across the environmental variables (Supplementary Figure 3.12, Supplementary Table 3.1).

3.3.2 Effect of the choice of the number of random vectors

We explored the choice of the number of random vectors in two ways. First, we quantified the contribution of randomization to the SE of the GxE estimator in GENIE. We simulated 100 phenotypes where $h^2_{gxe} = 0$ and the causal ratio is 10%. We compared the SE of GxE estimates with $B = 10$ random vectors run 100 times over one of the replicates (the contribution of the randomization to the SE) to the SE of GxE estimates across 100 replicates to determine that, with $B = 10$, randomization contributes to about a third of the total SE. Second, we verified that our GxE estimates are highly correlated for the choice of random vectors $B = 10$ vs $B = 100$ (Pearson's $\rho = 0.97$; Supplementary Figure 3.11). These results lead us to conclude that $B = 10$ random vectors provide stable estimates, and we use this setting in our remaining analyses.

3.3.3 Noise heterogeneity

Previous studies have shown that accounting for noise heterogeneity (NxE component) is essential to avoid false positives and inflation in estimates of GxE effects [113, 76, 12]. To demonstrate the importance of modeling NxE, we simulated phenotypes in the presence of NxE effect such that $h^2_{gxe} = \sigma^2_{nxe} \in \{0, 0.04, 0.08, 0.10\}$ (we set σ^2_{nxe} to 0.04 when $h^2_{gxe} = 0$). We ran GENIE, in turn, with and without NxE component. Across all simulations, the model that does not account for the NxE component (G+GxE) yields statistically significant upward bias in its GxE estimates (relative bias ranges from 2.5% to 69% across genetic architectures) while the model that fits a noise heterogeneity component (G+GxE+NxE) achieves unbiased estimates of GxE (Figure 3.2a)).

Further, we compared the calibration of tests of GxE from GENIE with MEMMA [52], a recently proposed scalable method for GxE heritability estimation (we did not include GPLEMMA in this comparison as the model underlying GPLEMMA aims to infer a linear combination of multiple environmental variables that maximizes h^2_{gxe} while our experiments all focus on the setting of a single environmental variable). First, we simulated phenotypes with neither GxE effects nor NxE effects from a subset of $N = 40k$ unrelated white British individuals. In this setting, MEMMA has an inflated false positive rate while GENIE is calibrated (Figure 3.2b)). The inflated false positive rate for MEMMA in the absence of the NxE effect can be explained by a bias in their estimates of the SE of the variance components (Supplementary Figure 3.10). We then explored the setting with noise heterogeneity but no GxE. The false positive rate of MEMMA increases as it does not model noise heterogeneity while GENIE has a controlled false positive rate across all simulations (Figure 3.2b)).

3.3.4 Computational efficiency

We evaluated the runtime of GENIE, MEMMA, and GCTA(HE) (which implements an exact method-of-moments estimator) with increasing sample size ($N \in \{10K, 50K, 100K, 290K\}$) for a fixed number of SNPs ($M = 459,792$) and a single environmental variable. We ran all methods to fit a single G and GxE variance component. All methods were run on an Intel(R) Xeon(R) Gold 6140 CPU 2.30GHz, with 187GB RAM. GENIE and MEMMA were run with ten random vectors. The runtime of GCTA(HE) includes the time to compute the GRMs. We ran GENIE and GCTA(HE) on a single core while we ran MEMMA on both a single core and four cores. We set a maximum time limit of two days as a constraint for all methods. We could run GCTA(HE) on a dataset of up to $50K$ samples. GENIE is highly scalable and can estimate GxE on about $300K$ individuals and roughly $500K$ SNPs within an hour, approximately 30 times faster than MEMMA run on four cores (Supplementary Figure 3.13).

3.3.5 Estimating GxE in the UK Biobank

We applied GENIE to estimate additive heritability (h_g^2) and GxE heritability (h_{gxe}^2) for fifty quantitative phenotypes measured in UKBB across unrelated white British individuals. These fifty phenotypes fall into eight broader phenotypic categories (blood biochemistry, kidney biomarkers, anthropometry, lipid metabolism biomarkers, blood pressure, liver biomarkers, lung, and glucose metabolism biomarkers) that have been analyzed in prior works [80, 124]. Following these studies, we applied a rank-based inverse normal transformation to all phenotypes. We considered, in turn, smoking status, sex, age, and statin usage as environmental variables. We included each environmental variable as a fixed effect in the relevant analyses. First, we explored the importance of modeling noise-environment interactions (NxE) in real data (building on our simulation results). We then analyzed, in turn, common SNPs genotyped on the UK Biobank array (MAF$>$ 1%), and then common and low-frequency imputed SNPs (MAF $\geq$ 0.1%). For select combinations of phenotypes and environmental variables, we also applied GENIE to partition GxE heritability across

MAF-LD annotations and to estimate GxE heritability in genes expressed in specific tissues.

Robustness of GENIE in the UK Biobank

We first assessed the robustness of GENIE by estimating h_g^2 under three different models: G, G+GxE, and G+GxE+NxE where each model is named by the set of variance components fitted jointly. The additive heritability estimates were highly correlated across the models (Pearson's correlation $\rho \geq 0.98$ for every pair of models), leading us to conclude that GENIE provides robust estimates of additive heritability across different models (Supplementary Figure 3.14). We observe a significant difference in h_g^2 for a handful of trait-E pairs when estimated with G+GxE and G+GxE+NxE that include alcohol frequency intake, overall health, and hair color with both smoking status and sex as environmental variables, and alcohol frequency intake and overall health with age.

Our simulations in the previous section revealed the importance of modeling noise heterogeneity (Figure 3.2). To investigate the consequences of modeling NxE in real data, we fit, in turn, models without and with NxE (in addition to G and GxE components). The number of trait-E pairs with significant h_{gxe}^2 ($p < 0.05/200$) decreased from 135 under the G+GxE model to 69 under the G+GxE+NxE model: decreasing from 40 to 21 for smoking (Figure 3.3b)), 27 to 29 for sex (Supplementary Figure 3.15b)), 28 to 12 for age (Supplementary Figure 3.16b)), and 40 to 7 for statin usage (Supplementary Figure 3.17b)). For traits with significant h_{gxe}^2, the magnitudes of the estimates varied across the two models: ratio of h_{gxe}^2 estimates under the G+GxE+NxE to the G+GxE model were 137% on average (range: $43-350\%$), 110% ($70-224\%$), 131% ($99-166\%$), and 42% ($21-72\%$) for smoking (Figure 3.3a)), sex (Supplementary Figure 3.15a)), age (Supplementary Figure 3.16a)), and statin (Supplementary Figure 3.17a)) respectively. The magnitude of noise heterogeneity across trait-E pairs can be substantial: 0.05%, 164%, 10%, and 14% of the additive heritability on average for smoking, sex, age, and statin, respectively (Supplementary Figures 3.18, 3.19, 3.20, and 3.21).

To further investigate the effect of modeling NxE, we performed permutation analyses by randomly shuffling the genotypes while preserving the trait-E relationship (a setting where there is expected to be no GxE by construction while the relationship between phenotype and E is preserved). We applied GENIE under the G+GxE and G+GxE+NxE models to each trait-E pair. The false positive rate of rejecting the null hypothesis of no GxE across the trait-E pairs is substantially inflated under the G+GxE model while being controlled under the G+GxE+NxE model (Figure 3.3c), Supplementary Figures 3.15c), 3.16c), and 3.17c) for smoking, sex, age, and statin respectively). These results indicate that modeling NxE is critical to avoid spurious findings of GxE.

Gene-by-Smoking Interaction

We applied GENIE to estimate the proportion of phenotypic variance explained by gene-by-smoking interactions ($h^2_{gxSmoking}$) for 50 quantitative phenotypes. We find 21 traits showing statistically significant evidence for $h^2_{gxSmoking}$ ($p < 0.05/200$) with $h^2_{gxSmoking}$ about 6.1% of h^2_g on average (Figure 3.4). Two of the traits with the largest $h^2_{gxSmoking}$ were basal metabolic rate and body mass index (BMI) with estimates of 2.4% and 2.3% respectively (estimates remained significant when we used the binary coding of the smoking status variable obtained by merging the categories of never and previous; Supplementary Figure 3.22). Our estimates are consistent with a previous study of that analyzed BMI and lifestyle factors in the UKBB to find significant GxE for smoking behavior [93]. The $h^2_{gxSmoking}$ estimates for basal metabolic rate and BMI are about 11% and 7% of their respective h^2_g estimates.

Gene-by-Sex Interaction

We find 29 traits with statistically significant h^2_{gxSex} ($p < 0.05/200$) with h^2_{gxSex}/h^2_g observed to be 8.7% on average (Figure 3.5). Serum testosterone levels showed the largest h^2_{gxSex} of 11% with the h^2_{gxSex} nearly as large as h^2_g consistent with prior work showing differences in genetic associations [96, 105] and heritability [147] across males and females. Beyond testosterone, we observe significant h^2_{gxSex} for several anthropometric traits, such as waist-hip-ratio adjusted for BMI (WHR) ($h^2_{gxSex} = 4.3\%$ and $\frac{h^2_{gxSex}}{h^2_g} = 20\%$), and lipid measures (results consistent for binary encoding; Supplementary Figure 3.23) consistent with previous work documenting sex-specific differences in the genetic architecture of anthropometric traits [89, 127, 88, 90, 2, 147]. Consistent with prior GWAS that identified genetic variants with sex-dependent effects [15, 54], our analyses of serum urate levels show substantial point estimates of h^2_{gxSex}, although these estimates are not statistically significant.

Gene-by-Age Interaction

We find 12 traits with statistically significant h^2_{gxAge} ($p < 0.05/200$) with h^2_{gxAge}/h^2_g observed to be 4.3% on average (Figure 3.6). Lipid and blood pressure measures show some of the largest h^2_{gxAge} (about 2.5% for LDL-C and total cholesterol and 1.9% for diastolic blood pressure). Previous studies have found genetic variants in the SORT1 gene to have age-dependent effects on LDL cholesterol [101] and nominal evidence for age-dependent genetic effects on blood pressure regulation [104]. We find that BMI shows evidence for significant h^2_{gxAge} while WHR does not, expanding on prior work that identified age-dependent genetic variants for BMI but not for WHR in GWAS [127]. Interestingly, we used a standardized encoding of age so that GxAge effects capture the interaction of genetic effects on the phenotype as a function of deviation from the mean age in UKBB while previous studies typically focus on changes in genetic effects in bins of age. It is plausible that other codings of age, *e.g.*, coding age to measure interactions as a function of older vs. younger individuals, could yield differing results.

Gene-by-Statin Interaction

We find seven traits that show statistically significant evidence for $h^2_{gxStatin}$ ($p < 0.05/200$) with an average ratio of $h^2_{gxStatin}$ to h^2_g across traits of 5.2% (Figure 3.7). We find that LDL and total cholesterol show significant $h^2_{gxStatin}$ (1.7% and 1.6% respectively) while HDL cholesterol with a point estimate of $h^2_{gxStatin}$ of 0.4% does not (results consistent for binary encoding; Supplementary Figure 3.24). We observe the largest estimates of $h^2_{gxStatin}$ for HbA1c and blood glucose measurements (2% and 1.2% respectively) which are interesting in light of statin usage being shown to be associated with a small increase in risk for Type-2 Diabetes [97].

Estimating GxE heritability from imputed SNPs

We applied GENIE to estimate $h^2_{gxSmoking}$, h^2_{gxSex}, h^2_{gxAge} ,and $h^2_{gxStatin}$ attributable to $M = 7,774,235$ imputed SNPs with MAF $\geq 0.1\%$. Prior work has shown that analyzing common and low-frequency variants with a single variance component can result in biased estimates of additive heritability [109, 18]. A solution to this problem involves fitting multiple variance components obtained by partitioning SNPs based on their frequency and local LD scores (as quantified by the LD-scores [3] or the LDAK scores [109]) [108, 18, 29, 107]. We follow this approach by partitioning SNPs into eight annotations based on quartiles of the LD-scores and two MAF annotations (MAF$<5\%$ and MAF$>5\%$; Methods).

We performed simulations to show that GENIE applied with SNPs partitioned based on MAF and LD scores can accurately estimate h^2_{gxe} across varying MAF and LD-dependent genetic architectures while using a single component for all SNPs can lead to substantial biases (Supplementary Section 3.6; Supplementary Figure 3.25). We applied GENIE using MAF-LD partitions to jointly estimate h^2_g and h^2_{gxe} (Supplementary Figures 3.26, 3.27, 3.28, and 3.29). While estimates of h^2_{gxe} from imputed SNPs are largely concordant with the estimates obtained from array SNPs, we identify nine trait-E pairs for which the h^2_{gxe} estimates are significantly different ($p < 0.05/200$). In all these cases, h^2_{gxe} estimates from imputed SNPs are higher than those from array SNPs. For example, we estimate $h^2_{gxSmoking}$ for BMI $= 6.5 \pm 0.5\%$ which is larger than our estimate based on array SNPs as well as a previous estimate of $4.0 \pm 0.8\%$ based on common HapMap3 SNPs [93]. Restricting to traits with significant GxE in both array and imputed SNPs, we observed that the average ratio ($\frac{h^2_{gxe}(imputed)}{h^2_{gxe}(array)}$) is 1.79 (2.31, 1.63, 1.17, and 1.17 respectively for GxSmoking, GxSex, GxAge, and GxStatin; Supplementary Figure 3.30). Across trait-E pairs with significant h^2_{gxe}, the average h^2_{gxe} is 2.8% on the imputed data compared to 1.5% on array data while the ratio of $\frac{h^2_{gxe}}{h^2_g}$ is 14.3% on the imputed data compared to 6.8% on the array data (averaged across trait-E pairs, we estimate $h^2_{gxe} = 0.9\%$ on imputed vs 0.7% on array data).

We explored the impact of fitting multiple variance components based on MAF and LD by applying GENIE to fit a single GxE and additive variance component using Smoking status as the environmental variable. While ten traits showed significant $h^2_{gxSmoking}$ in both analyses, five traits were exclusively significant in the MAF-LD model while one was exclusively significant in the single-component model. Restricting to traits with significant GxSmoking in both models, $h^2_{gxSmoking}$ estimates in the MAF-LD model were about 3x those from the single-component model on average (Supplementary Figure 3.31). We also investigated whether MAF-LD partitioning affected estimates of $h^2_{gxSmoking}$ obtained from array SNPs. We find that $h^2_{gxSmoking}$ estimates are largely concordant whether obtained from a single component or a MAF-LD partitioned model (ratio of 0.99 on average) consistent with the array SNPs being relatively common (MAF $> 1\%$).

Our analysis suggests that partitioning by MAF and LD is helpful for estimating h^2_{gxe} from both common and low-frequency SNPs and the inclusion low-frequency SNPs can increase estimates of h^2_{gxe} for specific traits.

Partitioning GxE heritability across MAF and LD annotations

Previous studies that have shown that the additive SNP effects increase with decreasing minor allele frequency (MAF) and local levels of linkage disequilibrium (LD) [28, 81, 98, 141], likely due to the effects of negative selection. However, the MAF-LD dependence of SNP effects in the context of specific environmental factors has not been empirically explored. Our analyses in the preceding section, showing differences in the genome-wide h^2_{gxe} estimates when partitioning by MAF and LD vs. fitting a single variance component, suggest that GxE effects are expected to vary by MAF and LD in a pattern that is distinct from what would be expected when fitting a single variance component which assumes that the effect size at a SNP varies with its allele frequency f as $\frac{1}{f(1-f)}$ while not varying with local LD (for a fixed value of the allele frequency f).

To explore the MAF-LD dependence of GxE effects, we used GENIE to partition h^2_{gxe}

across MAF and LD annotations (while also simultaneously partitioning additive heritability) of $M = 7,774,235$ imputed SNPs divided into eight annotations based on quartiles of LD-scores and two MAF bins (low-frequency bins with MAF$< 5\%$ and high-frequency bins with MAF$\geq 5\%$). Within each of these eight bins, we defined the per-allele squared effect size as $\beta_k^2 = \frac{h_k^2}{2M_k f_k(1-f_k)}$ where h_k^2 is the GxE (or additive) heritability attributed to bin k, M_k is the number of SNPs in bin k and f_k is the mean MAF in bin k.

For the sake of presentation, we selected one phenotype with high genome-wide GxE heritability for each of the four environmental variables analyzed (Figure 3.8; See the Supplementary Data 1 for results on all trait-E pairs). Across bins of MAF and LD, the magnitude of additive allelic effects tends to be larger than those of the GxE effects consistent with the genome-wide results. We observe that the per-allele squared GxE effect size β_{gxe}^2 tends to increase with lower MAF within a given quartile of LD score and to increase with lower bins of LD score for a fixed MAF bin (Figure 3.8a). These trends are analogous to the relationship observed for additive per-allele effect sizes (Figure 3.8b). Across the trait-E pairs, restricting to the lowest quartile of LD scores, low-frequency SNPs tend to have higher per-allele GxE effect sizes compared to high-frequency SNPs: the ratio of β_{gxe}^2 in low vs high MAF bins is 8.2 ± 11.2, 24.6 ± 19.7, 3.4 ± 2.1, and 3.7 ± 1.2 for HbA1c-statin, BMI-smoking, LDL-age, and testosterone-sex respectively. In the highest quartile of LD scores, we found no statistically significant differences in β_{gxe}^2 across low and high MAF SNPs in any of the four trait-E pairs (we also plot the per-standardized genotype additive and GxE heritability, $\frac{h_k^2}{M_k}$, in Supplementary Figure 3.32).

Partitioning GxE heritability across tissue-specific genes

The ability of GENIE to simultaneously estimate multiple, potentially overlapping, additive and GxE variance components enables us to explore how h_{gxe}^2 is localized across the genome. Specifically, we set to answer the question of whether h_{gxe}^2 is enriched in genes specifically expressed in a given tissue as a means to identify tissues that are relevant to a trait in a specific environmental context.

We applied GENIE to estimate h_g^2 and h_{gxe}^2 across each of 53 sets of genomic annotations defined as regions around genes that are highly expressed in a specific tissue in the GTEx dataset [23]. For each of the four environmental variables, we analyzed only traits with genome-wide significant h_{gxe}^2 based on our prior analyses of the array SNPs. For every set of tissue-specific genes, we followed prior work [23] by jointly modeling the tissue-specific gene annotation as well as 28 genomic annotations that are part of the baseline LDSC annotations that include genic regions, enhancer regions, and conserved regions [21]). Specifically, our model has 29 additive variance components and 29 GxE variance components and estimates the additive and GxE heritability that can be attributed to genes specifically expressed in a tissue while controlling for the effects of the background annotations. A positive $h_{g,tissue}^2$ represents a positive contribution of genetic effects in a tissue to additive heritability [23]. Analogously, a positive $h_{gxe,tissue}^2$ represents a positive contribution of genetic effects in this tissue to trait heritability in the context of the specific environment. We test estimates of $\frac{h_{gxe,tissue}^2/h_{gxe,total}^2}{M_{tissue}/M_{total}}$ $(\frac{h_{g,tissue}^2/h_{g,total}^2}{M_{tissue}/M_{total}})$ to answer whether a tissue of interest is enriched for GxE (additive) heritability conditional on the remaining genomic annotations included in the model (Methods).

We first verified that our approach is able to detect previously reported enrichments for additive effects such as brain-specific enrichment for BMI and adipose-specific enrichment for WHR (Figure 3.9) [23]. Across 69 trait-E pairs with significant genome-wide GxE that we tested, we observed significant enrichment of $h_{gxe,tissue}^2$ (FDR < 0.10) for at least one tissue in five trait-E pairs (we plot four of these pairs in Figure 3.9 since the results from the fifth LDL cholesterol-age are highly correlated with cholesterol-age). Across these trait-E pairs, we document differential patterns of enrichments for GxE effects compared to additive effects. BMI exhibits brain-specific enrichment of $h_{gxSmoking}^2$ and h_g^2 while WHR exhibits enrichment of h_{gxSex}^2 and h_g^2 in adipose and breast tissue (in addition to the enrichment of h_g^2 in the uterus and cardiovascular tissues). The adipose tissue-specific enrichment of h_{gxSex}^2 in WHR is notable in light of known instances of genes that are associated with WHR in adipose tissue in a sex-dependent manner. ADAMTS9, a gene involved in insulin sensitivity [89], is specifically expressed in adipose tissue and has been shown to be located near GWAS hits for WHR that are specific to females [103, 127, 89]. The transcription factor, KLF14, is located near a sex-dependent GWAS variant for WHR, type-2 Diabetes, and multiple other metabolic and anthropometric traits [106]. Further, the expression level of this gene is associated with the GWAS variant in adipose but not other tissues [106]. We also find instances where tissues that are enriched for h^2_{gxe} are distinct from those that are enriched for h^2_g. We observe that the enrichment of h^2_{gxSex} for basal metabolic rate in brain and adipose tissues is distinct from the tissues that are enriched in h^2_g for the same trait (cardiovascular and digestive tissues) (Figure 3.9). Fitting this trend, we find that h^2_{gxAge} for cholesterol shows enrichment in cardiovascular tissues while h^2_g shows liver-specific enrichment. Finally, we find suggestive evidence that the liver is the most significantly enriched tissue for $h^2_{gxStatin}$ in HbA1c ($p = 0.02$) as well as for h^2_{gxSex} in testosterone ($p = 0.005$) although neither enrichment is significant at FDR of 0.10. These enrichments recapitulate known biology: the liver-specific enrichment of GxStatin effects for HbA1c reflect the tissues in which the target of statins (HMG-CoA-reductase) is expressed [110] while the liver-specific enrichment of GxSex for testosterone is consistent with previous findings implicating CYP3A7, a gene involved in testosterone metabolism that is specifically expressed in the liver and lies within a locus that contains one of the strongest GWAS signals for serum testosterone in females [105]

3.4 Discussion

We have described GENIE, a method that can jointly estimate the proportion of variation in a complex trait that can be attributed to GxE and additive genetic effects. GENIE can also partition GxE heritability across the genome with respect to annotations, such as functional and tissue-specific annotations or annotations defined based on the minor allele frequency (MAF) and local linkage disequilibrium (LD score) of each SNP to localize signals of GxE. GENIE provides well-calibrated tests for the existence of a GxE effect and has high power to detect GxE effects while being scalable to large datasets.

Our simulations and real data analysis results confirm the importance of including noise heterogeneity in GxE models. In UKBB data analyses, we observed about half of trait-E pairs with significant h^2_{gxe} under the G+GxE model are no longer significant under the G+GxE+NxE model. Consistent with this observation, we estimate a substantial contribution of noise heterogeneity to trait variation.

After accounting for noise heterogeneity, we observe significant genome-wide h^2_{gxe} across more than a quarter of the trait-E pairs analyzed. Our finding has implications for understanding trait heritability by moving beyond the definition of narrow-sense heritability that only includes additive genetic effects. Based on our analyses, it is conceivable that approaches that can jointly model the hundreds of environmental variables measured in Biobank-scale datasets will further increase estimates of h^2_{gxe}. Additionally, our recovery of additional h^2_{gxe} from low-frequency SNPs ($0.1\% \geq \text{MAF} < 1\%$) point to traits where an understanding of GxE effects can benefit from whole-exome and whole-genome studies. Further, our results point to traits where GxE has the potential to improve genome-wide polygenic scores (GPS) of complex traits (since h^2_{gxe} quantifies the maximum predictive accuracy that is achievable by a linear predictor based on GxE effects). In the context of sex as an environmental variable, sex-specific GPS has been shown to provide improved accuracy over agnostic scores [91, 25, 2, 147]. GxE has also been recently proposed as a possible explanation for why GPS may not generalize beyond the cohort on which these predictors were trained [74] so that modeling GxE in relevant traits could improve their transferability. Our finding that allelic effects for GxE increase with decreasing MAF and LD analogous to the relationship observed for additive allelic effects motivates an evolutionary understanding of these trends and can inform what we expect to learn from studies of rare genetic variation. Finally, our identification of sets of genes that are enriched for GxE can offer clues on trait-relevant tissues and pathways and has the potential to inform functional genomic studies [14, 20].

We discuss the limitations of our work as well as directions for future research. First, GENIE does not explicitly model G-E correlations [76]. While such correlations can lead to biases in estimates of GxE in the fixed-effect setting [16], it has been shown that, in the polygenic setting, the GxE variance component estimates remain unbiased when G-E correlations are independent of the polygenic GxE effects [12]. Nevertheless, there are plausible settings, where such correlations can lead to false positive or biased estimates of GxE, *e.g.*, where the phenotype directly affects the environmental variable. Developing scalable methods that are accurate in these settings is an important direction for future work. Second, estimates of GxE heritability are sensitive to the scale on which traits and environmental variables are measured and how environmental variables are encoded. In this work, we analyze quantile-normalized traits (following prior studies) and encode discrete environmental variables using a univariate parameterization (either as a 0-1 vector for each environmental variable or as a standardized version). It might be preferable to work with traits measured on their original scale and to encode each level of discrete environmental variables by a separate 0-1 covariate (leading to k environmental covariates for a k-valued environmental variable). While such choices would necessarily be guided by domain knowledge and interpretability, GENIE supports easy-to-use and rapid exploration of the consequences of these choices and can aid in assessing the robustness of these choices (we have explored a limited space of these choices here). Third, the environmental variable relevant for GxE may not be measured directly or accurately so that the environmental variable that is measured in a dataset is best viewed as a proxy for the relevant latent environmental covariate. On a related note, while GENIE can model the impact of heterogeneous noise resulting from observed environmental variables by introducing NxE components, it is important to note that the heterogeneous noise may also arise due to non-observed environmental variables. Several recent works have tried to test for GxE when the environmental variables are not observed [139, 67]. These issues along with the possibility of reverse causality, *i.e.*, where the trait affects the environmental variable, warrant caution in any causal interpretation of our results (although it might be possible to overcome some of these limitations in specific analyses such as GxSex). Fourth, the model underlying GENIE is not applicable to binary traits (either with or without ascertainment). GENIE can be extended to be applicable to binary traits (*e.g.*, disease status) along the lines proposed in the context of additive [33, 125] and GxE estimation [12].

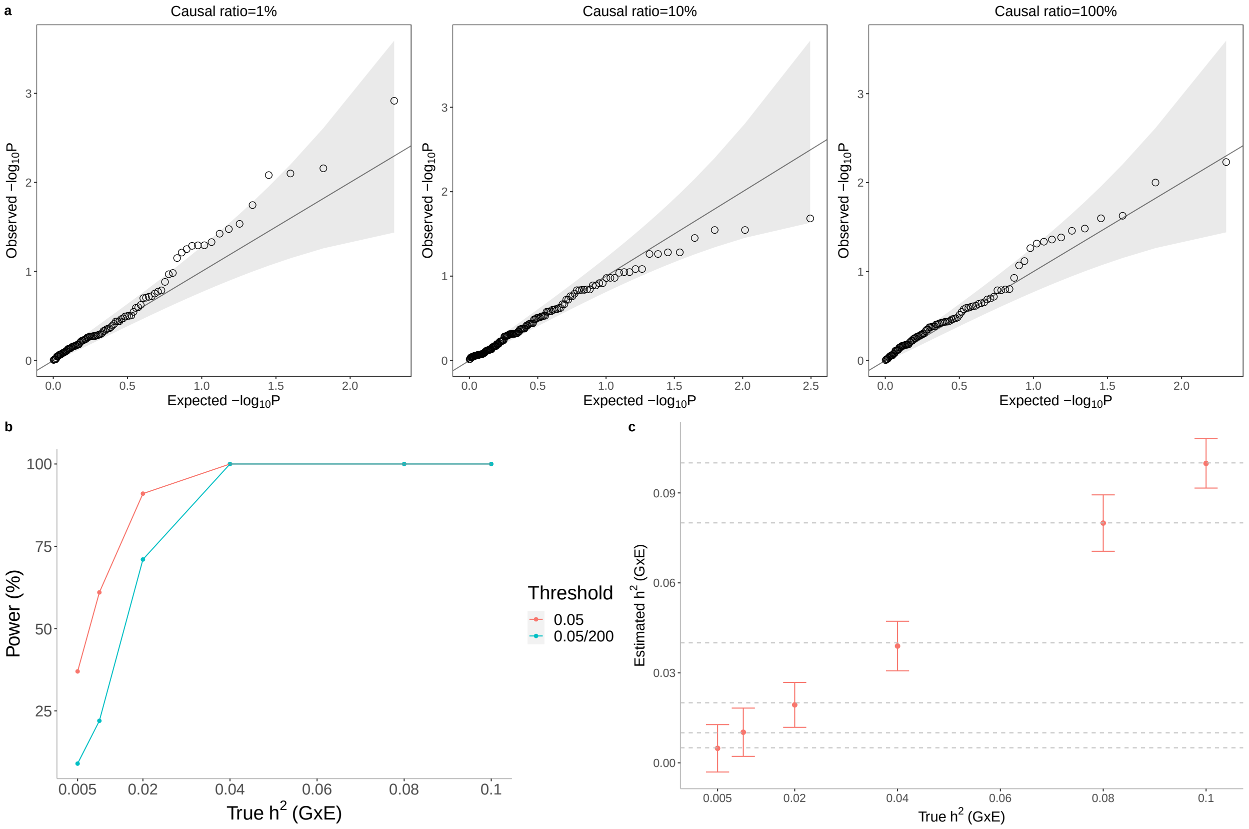


Figure 3.1: **Calibration and power of GENIE in large-scale simulations ($N = 291,273$ unrelated individuals, $M = 459,792$ SNPs).** **a**) Q-Q plot of p-values (of a test of the null hypothesis of zero GxE heritability) when GENIE is applied to phenotypes simulated in the absence of GxE effects. Each panel contains 100 replicate phenotypes simulated with additive heritability $h_g^2 = 0.25$ and varying proportions of causal variants. Across all architectures, the mean of $P(\text{rejection at } p < t)$ are 7.5% and 0% for $t = 0.05$ and $t = \frac{0.05}{200}$ respectively (7.5% is not significantly different from the nominal rate of 5%; the p-value of a test of bias of point estimates of h_{gxe}^2 is $p = 0.75$). **b**) The power of GENIE across genetic architectures as a function of GxE heritability. We report power for p-value thresholds of $t \in \{0.05, \frac{0.05}{200}\}$. **c**) The accuracy of h_{gxe}^2 estimates obtained by GENIE. Across all simulations, statin usage in UKBB was used as the environmental variable.

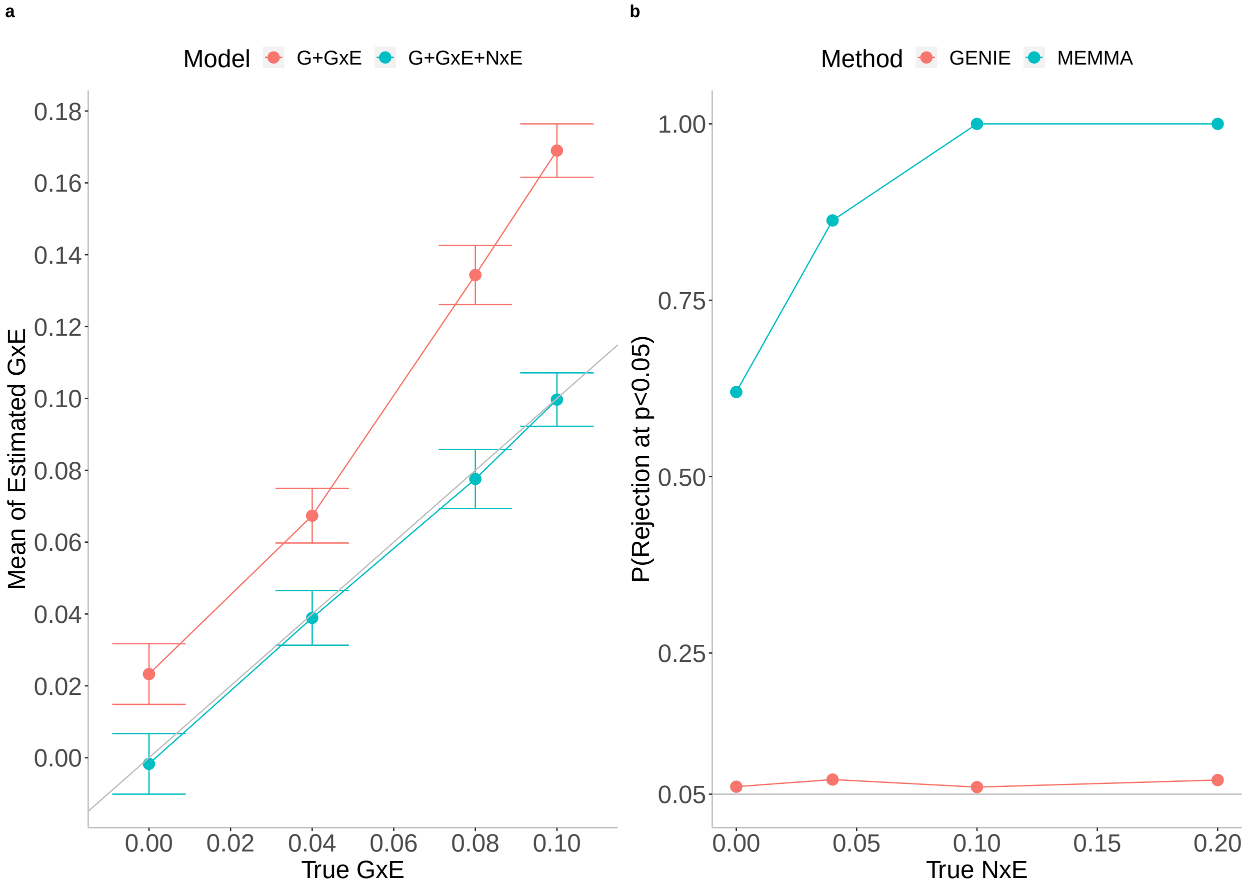


Figure 3.2: **Effect of noise heterogeneity (NxE) on the accuracy of estimates of GxE heritability in simulations.** **a**) Comparison of GxE heritability estimates from GENIE under a G+GxE model to those from a G+GxE+NxE model. Model G+GxE refers to a model with additive and gene-by-environment interaction components. Model G+GxE+NxE refers to a model with additive, gene-by-environment interaction, and noise heterogeneity (noise-by-environment interaction) components. We simulated phenotypes with NxE effects and GxE effects across $N = 291,273$ individuals genotyped at $M = 459,792$ SNPs. The x-axis and y-axis correspond to the true GxE and the mean of the estimated GxE (from 100 replicates), respectively. Points and error bars represent the mean and ± SE, respectively. **b**) Comparison of false positive rates of tests for GxE heritability across GENIE and MEMMA. We performed simulations with no GxE heritability but with varying magnitudes of the variance of the NxE effect. We compute the false positive rate as the fraction of rejections (p-value of a test of the null hypothesis of zero GxE heritability < 0.05) over 100 replicates of phenotypes simulated from $N = 40,000$ individuals genotyped at $M = 459,792$ SNPs.

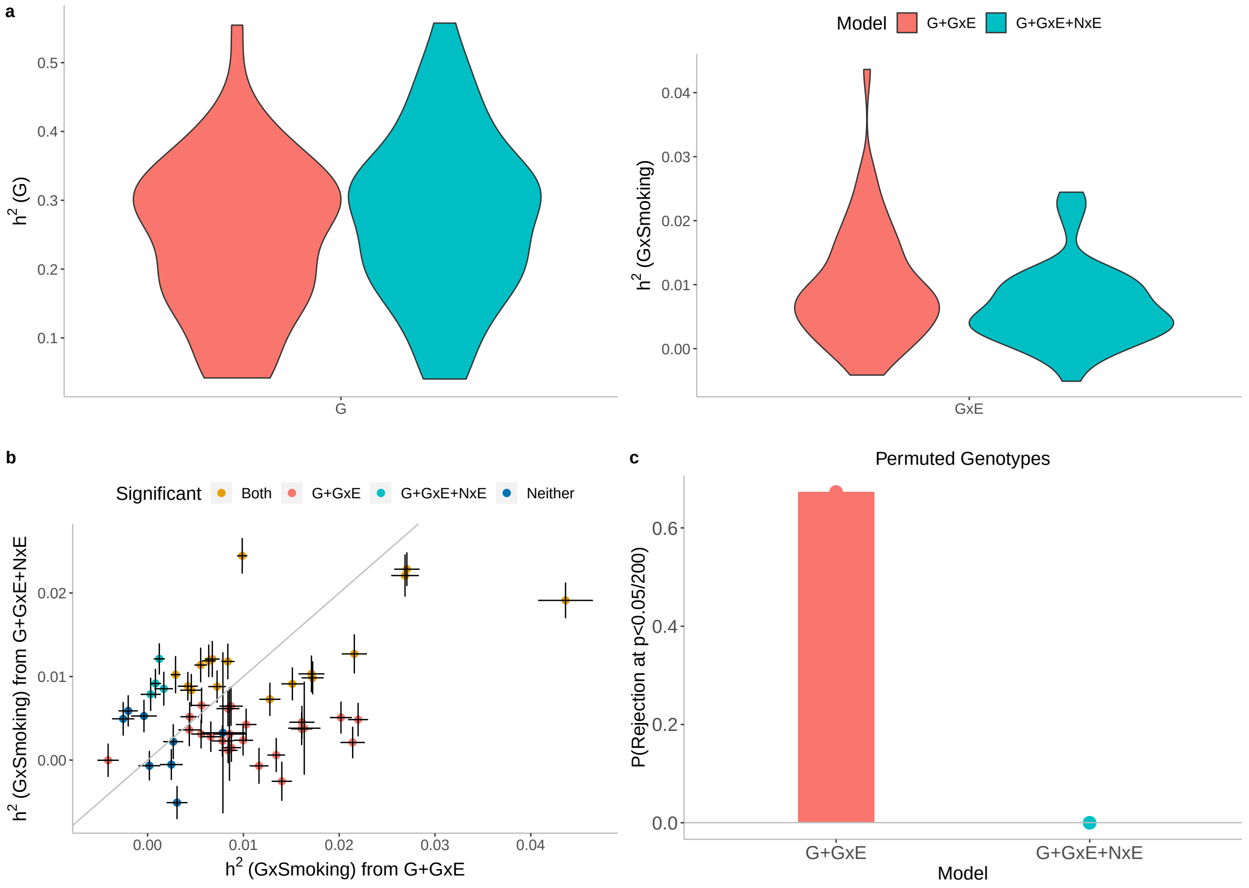


Figure 3.3: **Effect of Noise heterogeneity (NxE) on estimates of heritability associated with GxSmoking across 50 quantitative phenotypes in UKBB**. Model G+GxE refers to a model with additive and gene-by-environment interaction components where the environmental variable is smoking status. Model G+GxE+NxE refers to a model with additive, gene-by-environment interaction, and environmental heterogeneity (noise-by-environment interaction) components. **a**) We run GENIE under G+GxE and G+GxE+NxE models to assess the effect of fitting an NxE component on the additive and GxE heritability estimates. **b**) Comparison of GxE heritability estimates obtained from GENIE under a G+GxE+NxE model (x-axis) to a G+GxE model (y-axis). Black error bars mark ± standard errors centered on the estimated GxE heritability. Color of the dots indicate whether estimates of GxE heritability are significant under each model. **c**) We performed permutation analyses by randomly shuffling the genotypes while preserving the trait-E relationship and applied GENIE in each setting under G+GxE and G+GxE+NxE models. We report the fraction of rejections (p-value of a test of the null hypothesis of zero GxE heritability $< \frac{0.05}{200}$ that accounts for the number of phenotypes tested) over 50 UKBB phenotypes.

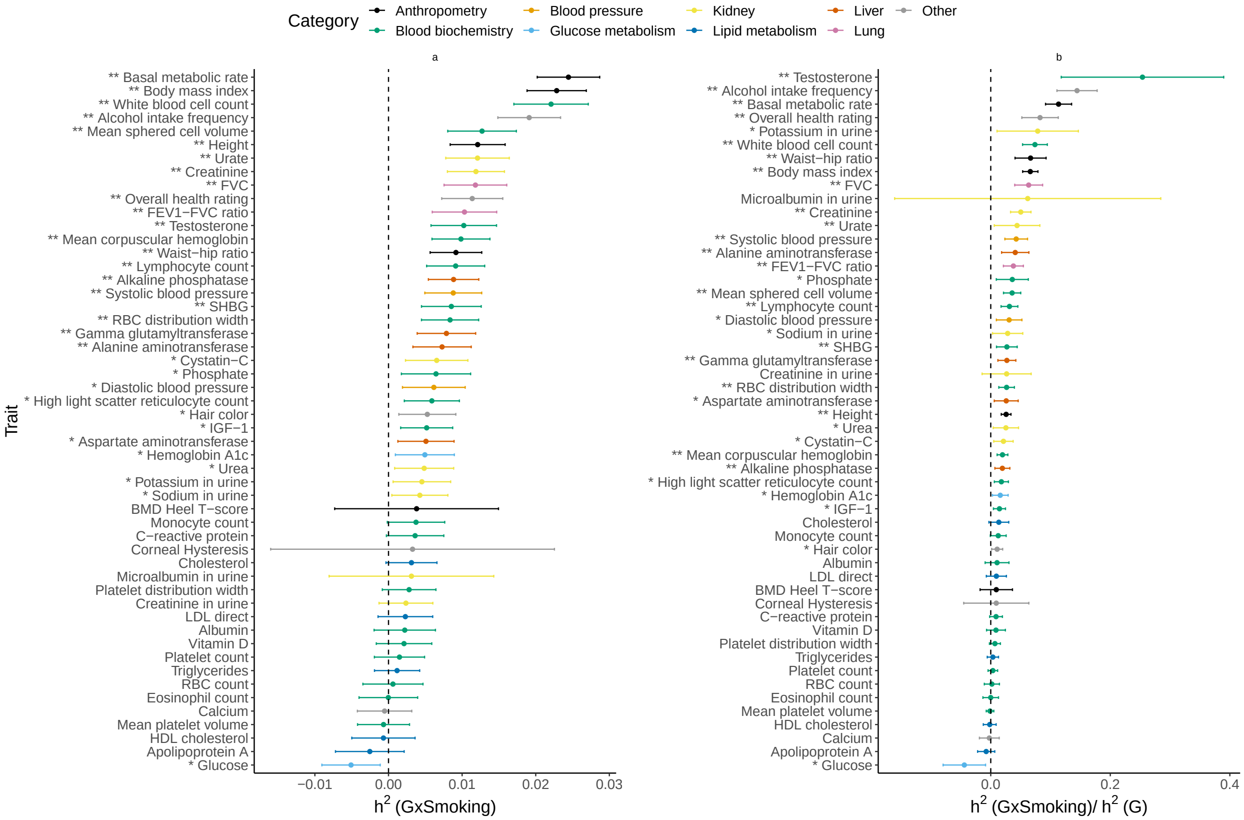


Figure 3.4: **Estimates of GxSmoking heritability across phenotypes in UK Biobank. a**) GxSmoking heritability and **b**) the ratio of GxSmoking to additive heritability. We applied GENIE to $N = 291,273$ unrelated white British individuals and $M = 459,792$ array SNPs (MAF$\geq$ 1%). Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. The environmental variable is standardized in these analyses. Error bars mark ±2 standard errors centered on the point estimates. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

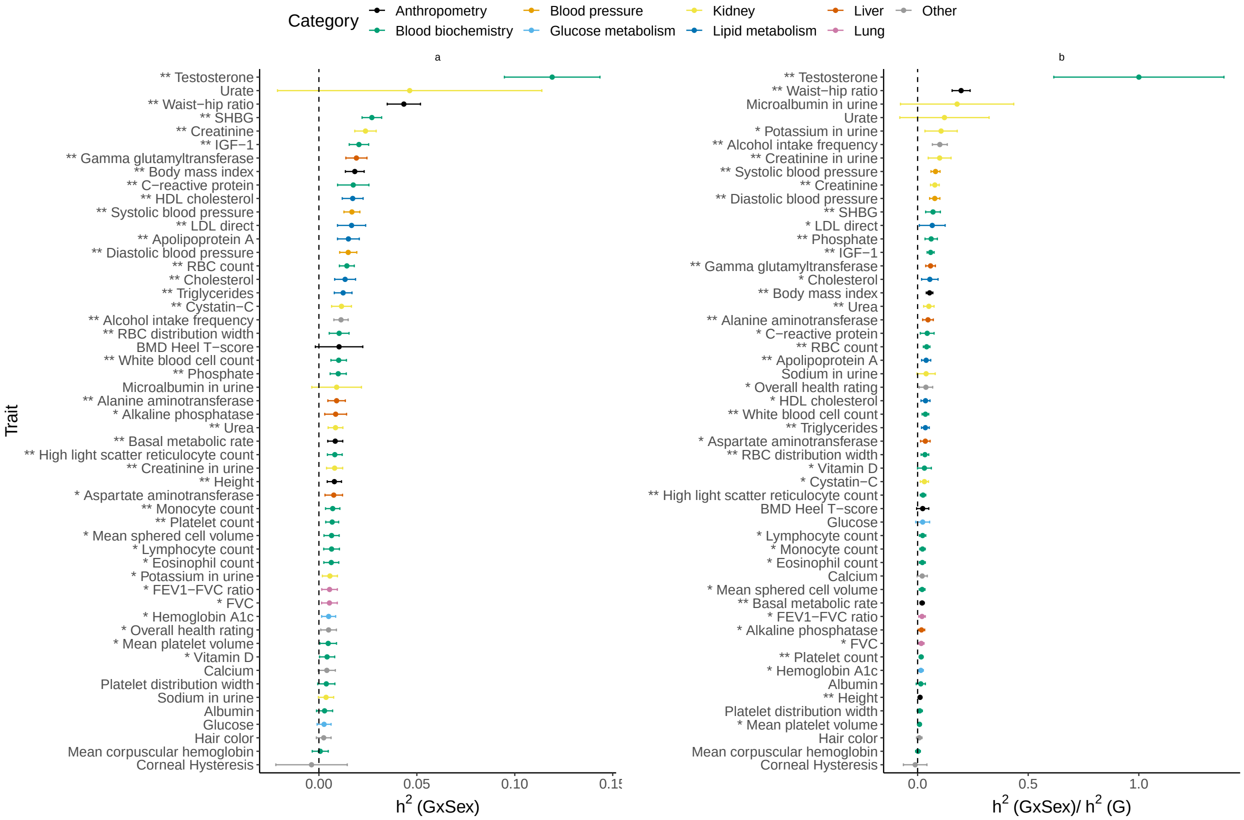


Figure 3.5: **Estimates of GxSex heritability across phenotypes in UK Biobank.** **a**) GxSex heritability and **b**) ratio of GxSex to additive heritability. We applied GENIE to $N = 291,273$ unrelated white British individuals and $M = 459,792$ array SNPs (MAF$\geq$ 1%). Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. The environmental variable is standardized in these analyses. Error bars mark ±2 standard errors centered on the point estimates. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

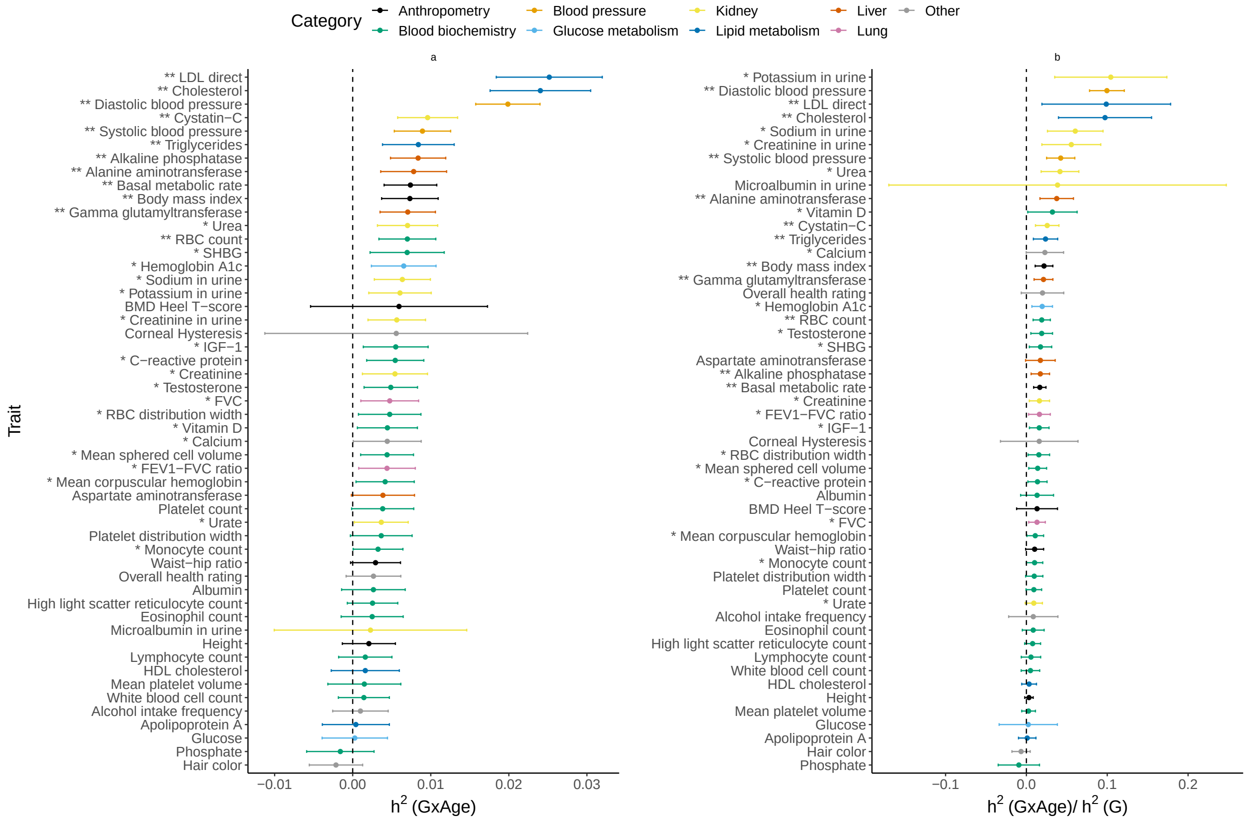


Figure 3.6: **Estimates of GxAge heritability across phenotypes in UK Biobank.** **a**) GxAge heritability and **b**) ratio of GxAge to additive heritability. We applied GENIE to $N = 291,273$ unrelated white British individuals and $M = 459,792$ array SNPs (MAF$\geq$ 1%). Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. The environmental variable is standardized in these analyses. Error bars mark ±2 standard errors centered on the point estimates. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

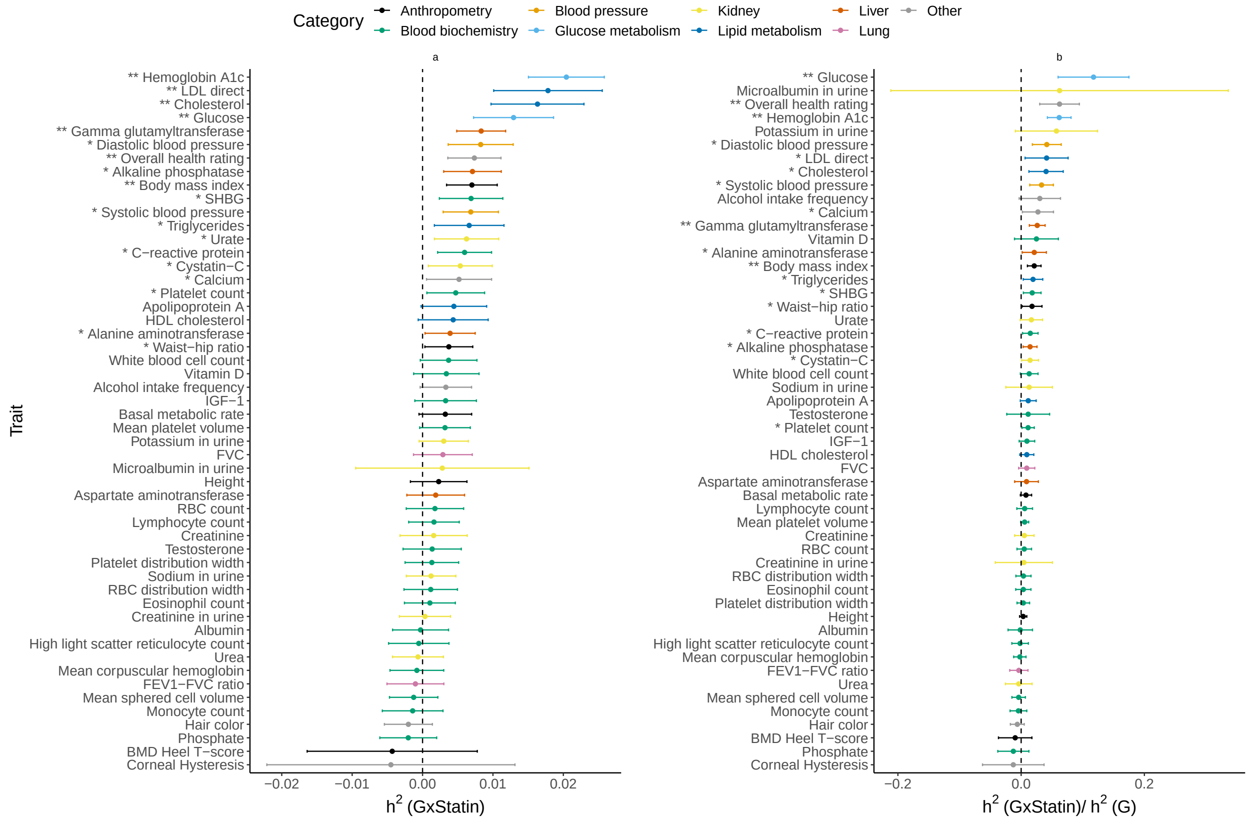


Figure 3.7: **Estimates of GxStatin heritability across phenotypes in UK Biobank.** **a**) GxStatin heritability and **b**) ratio of GxStatin to additive heritability. We applied GENIE to $N = 291,273$ unrelated white British individuals and $M = 459,792$ array SNPs (MAF$\geq$ 1%). Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. The environmental variable is standardized in these analyses. Error bars mark ±2 standard errors centered on the point estimates. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

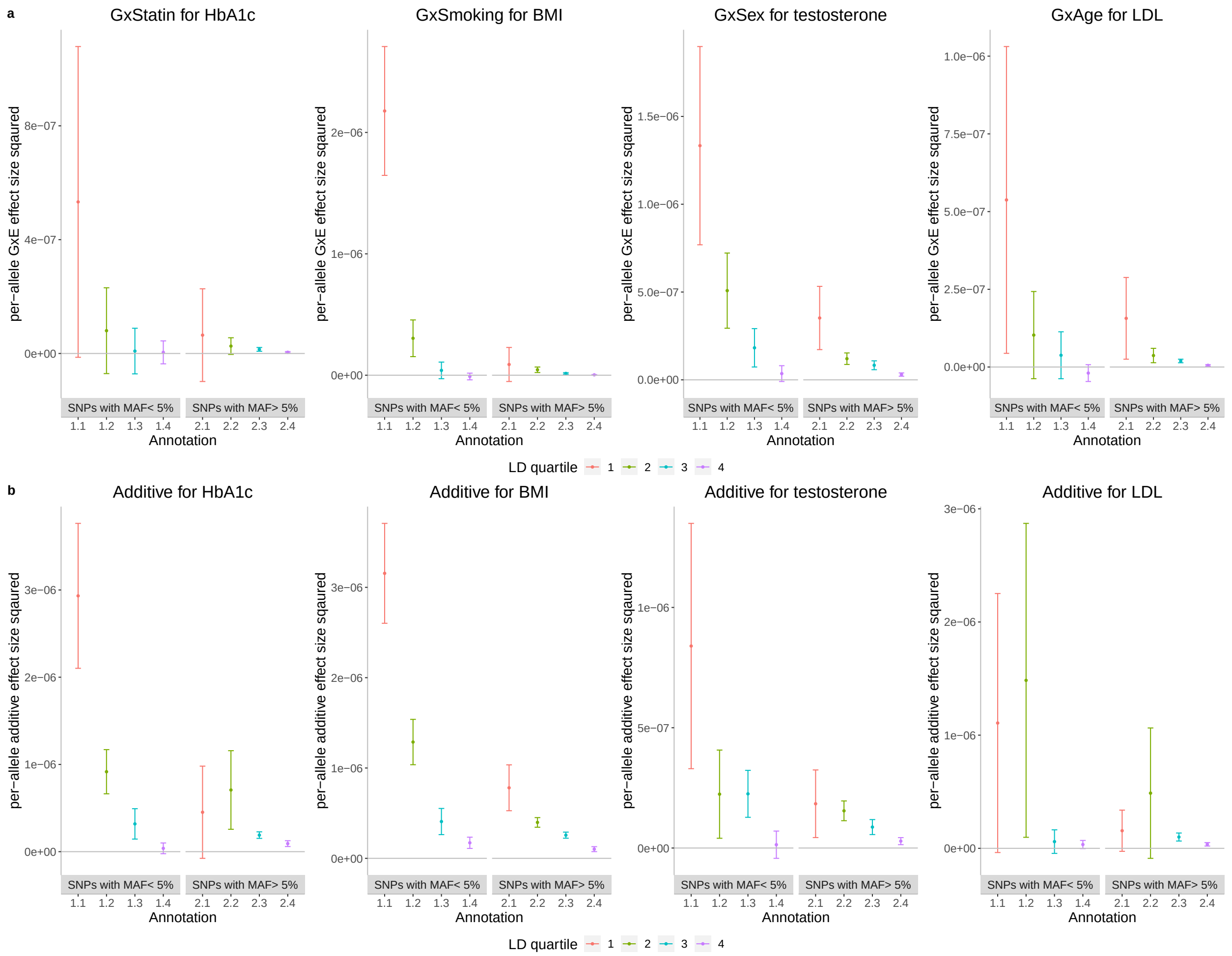


Figure 3.8: **Per-allele squared GxE and additive effect sizes as a function of MAF and LD. a**) The squared per-allele GxE effect size for four selected pairs of trait and environments (trait-E pairs). **b**) The squared per-allele additive effect size for the same trait-E pairs. The x-axis corresponds to MAF-LD annotations where annotation $i.j$ includes SNPs in MAF bin i and LD quartile j where MAF bin 1 and MAF bin 2 correspond to SNPs with MAF $\leq 5\%$ and MAF $> 5\%$ respectively while the first quartile of LD-scores correspond to SNPs with the lowest LD-scores respectively). The y-axis shows the per-allele GxE (or additive) effect size squared defined as $\frac{h_k^2}{2M_k f_k(1-f_k)}$ where h_k^2 is the GxE (or additive) heritability attributed to bin k, M_k is the number of SNPs in bin k, and f_k is the mean MAF in bin k. Error bars mark ±2 standard errors centered on the estimated effect sizes.

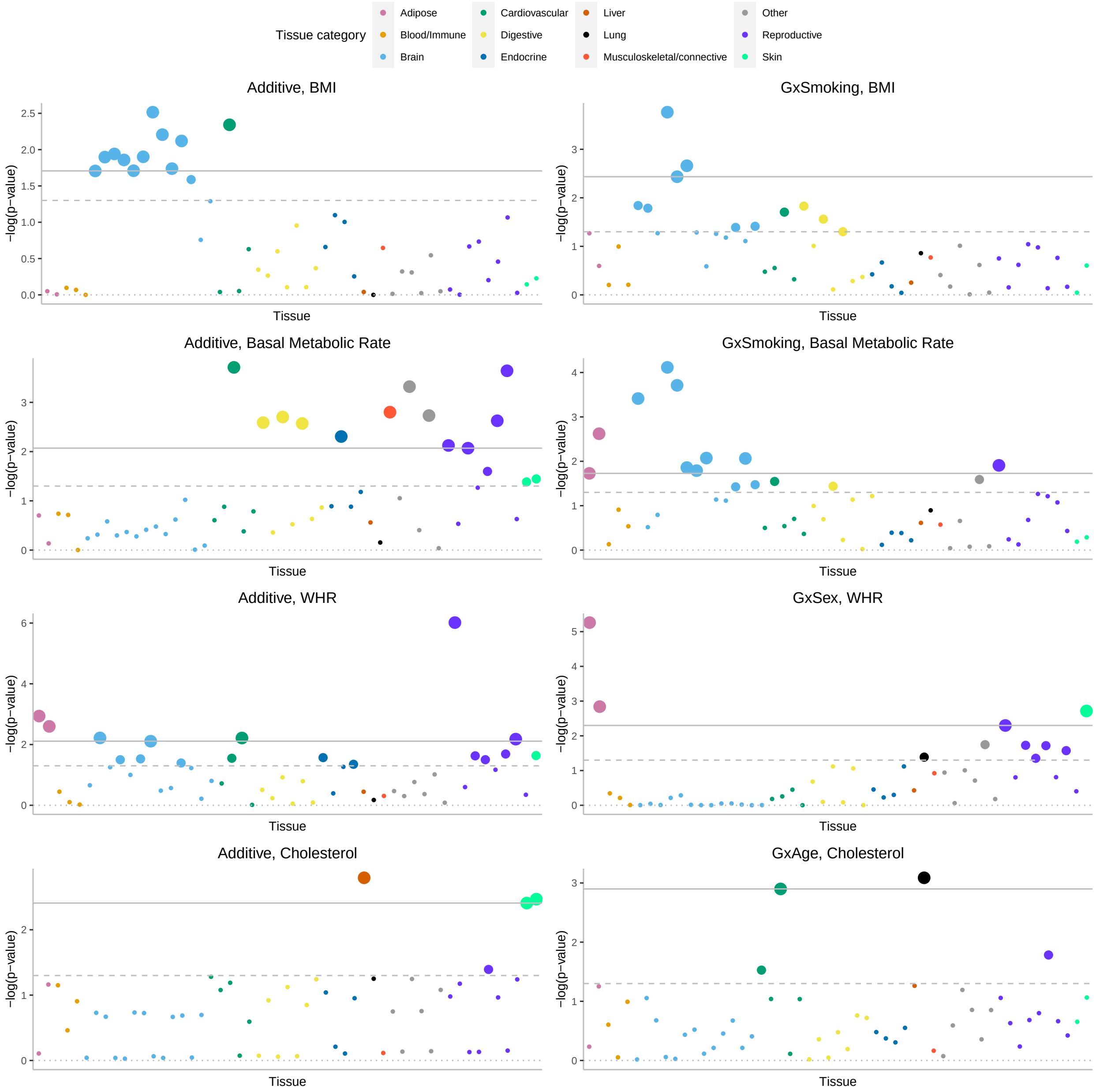


Figure 3.9: **Partitioning GxE heritability across 53 tissue-specific genes..** We plot $-log_{10}(p)$ where p is the corresponding p-value of the tissue-specific GxE enrichment defined as $\frac{h^2_{gxe,tissue}/h^2_{gxe,total}}{M_{tissue}/M_{total}}$. For every tissue-specific annotation, we use GENIE to test whether this annotation is significantly enriched for per-SNP heritability, conditional on 28 functional annotations that are part of the baseline LDSC annotations. The dashed and solid lines correspond to the nominal $p < 0.05$ and FDR< 0.1 threshold, respectively.

3.5 Supplementary Notes

3.5.1 Exact computation

$$
\begin{aligned}
tr(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) &= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}) \\
&= \sum_{i,j}^{LM} \left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}_{i,j}\right)^2 \qquad (3.15) \\
&= \sum_{i,j}^{LM} (\sum_{k} H_{ki}H_{kj})(\sum_{l} H_{li}H_{lj}) \qquad (3.16) \\
&= \sum_{i,j}^{LM} \sum_{k,l}^{N} H_{ki}H_{kj}H_{li}H_{lj}
\end{aligned}
$$

Using the expression in Equation 3.16, exact computation of $tr(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})$ requires $\mathcal{O}(N(ML)^2)$ time.

Write $H_{ki} = G_{ka}E_{kb}$ where $i = \rho(a,b)$ for some invertible mapping ρ that maps pairs $(a,b) \in \{1,\ldots,M\} \times \{1,\ldots,L\}$ to $i \in \{1,\ldots,ML\}$.

$$
\begin{aligned}
&= \sum_{i,j}^{LM} \sum_{k,l}^{N} H_{ki}H_{kj}H_{li}H_{lj} \\
&= \sum_{a,c}^{M} \sum_{b,d}^{L} \sum_{k,l}^{N} G_{ka}E_{kb}G_{kc}E_{kd}G_{la}E_{lb}G_{lc}E_{ld} \\
&= \sum_{k,l}^{N} \left(\sum_{a,c}^{M} G_{ka}G_{kc}G_{la}G_{lc}\right)\left(\sum_{b,d}^{L} E_{kb}E_{kd}E_{lb}E_{ld}\right) \\
&= \sum_{k,l}^{N} \left(\sum_{a}^{M} G_{ka}G_{la}\right)\left(\sum_{c}^{M} G_{kc}G_{lc}\right)\left(\sum_{b}^{L} E_{kb}E_{lb}\right)\left(\sum_{d}^{L} E_{kd}E_{ld}\right) \\
&= \sum_{k,l}^{N} \left(\sum_{a}^{M} G_{ka}G_{la}\right)^2\left(\sum_{b}^{L} E_{kb}E_{lb}\right)^2 \qquad (3.17)
\end{aligned}
$$

This computation requires $\mathcal{O}(N^2(M+L))$ time.

Thus, exact computation can be achieved in $\mathcal{O}(\min\left(N^2(M+L), N(ML)^2\right))$ time. The additional terms that involve $\boldsymbol{H}$ in Equation 3.4 is $tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})$. We have the following decomposition:

$$\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) \cdot (\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}),$$

where the product is element-wise (this can be seen from the steps in deriving (3.17)). We also note that similarly:

$$tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) = \sum_{i,j}^{N} (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})_{ij}^{2} \cdot (\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})_{ij}.$$

This yields an $O(N^2(M+L))$ exact algorithm for calculating all quantities in (3.4). We would like to avoid explicit computation of $\boldsymbol{H}$ forming a $N \times ML$ matrix in solving the normal equation.

3.5.2 Approximate computation in $\mathcal{O}(NMLB)$ time

Lemma 3 *[48] Let $\boldsymbol{A}$ be an $M \times M$ matrix and $\boldsymbol{w} \in \mathbb{R}^M$ a random vector with mean zero and covariance $\boldsymbol{I}_M$. Then $\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}$ is an unbiased estimator of $tr(\boldsymbol{A})$.*

Lemma 4 *Let $\boldsymbol{A}$ be an $ML \times ML$ matrix. Suppose that $\boldsymbol{u} \in \mathbb{R}^M, \boldsymbol{v} \in \mathbb{R}^L$ are independent random vectors with mean zero and covariance $\boldsymbol{I}_M$ and $\boldsymbol{I}_L$ respectively. Then $\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}$ is an unbiased estimator of $tr(\boldsymbol{A})$ where $\boldsymbol{w} = \boldsymbol{u} \otimes \boldsymbol{v}$.*

Proof:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w}\right] &= \mathbb{E}\left[\sum_{i,j}^{ML} w_i A_{ij} w_j\right] \\
&= \sum_{i,j}^{ML} \mathbb{E}\left[w_i A_{ij} w_j\right] \\
&= \sum_{i,j}^{ML} A_{ij}\mathbb{E}\left[w_i w_j\right] \\
&= \sum_{a,c}^{M}\sum_{b,d}^{L} A_{\rho(a,b)\rho(c,d)}\mathbb{E}\left[u_a v_b u_c v_d\right]
\end{aligned}$$

Here $(a,b) = \rho^{-1}(i)$ and $(c,d) = \rho^{-1}(j)$.If $i \neq j$, then either $a \neq c$ or $b \neq d$ so that $\mathbb{E}\left[u_a v_b u_c v_d\right] = 0$. Otherwise if $i = j$, then $a = c$ and $b = d$ so that $\mathbb{E}\left[u_a v_b u_c v_d\right] = \mathbb{E}\left[u_a^2 v_b^2\right] = \mathbb{E}\left[u_a^2\right]\mathbb{E}\left[v_b^2\right] = 1$.

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w}\right] &= \sum_{a=c}^{M}\sum_{b=d}^{L} A_{\rho(a,b)\rho(c,d)} \\ &= \sum_{i}^{ML} A_{ii} \\ &= tr(\boldsymbol{A})\end{aligned}$$

□

Theorem 2 *Suppose that $\boldsymbol{X}$ and $\boldsymbol{E}$ are $N \times M$ genotype and $N \times L$ environment matrices respectively. Define $\boldsymbol{H} = \boldsymbol{X} \odot \boldsymbol{E}$ as the $N \times ML$. Assume that $\boldsymbol{w} = \boldsymbol{u} \otimes \boldsymbol{v}$ where $\boldsymbol{u} \in \mathbb{R}^M, \boldsymbol{v} \in \mathbb{R}^L$ are independent random vectors with mean zero and covariance $\boldsymbol{I}_M$ and $\boldsymbol{I}_L$ respectively. Then $\boldsymbol{w}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{w}$ is an unbiased estimator of $tr(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{H})$ with $\mathcal{O}(NMLB)$ time complexity.*

Proof: In Lemma 4 we showed that $\hat{\theta} \equiv \boldsymbol{w}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}$ is an unbiased estimator of $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$. To compute $\hat{\theta}$, we consider the following intermediate computations:

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{w} \tag{3.18}$$

$$\boldsymbol{s} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{r} \tag{3.19}$$

$$\hat{\theta} = \boldsymbol{s}^{\mathrm{T}}\boldsymbol{s} \tag{3.20}$$

Based on Lemma 2 , we can compute $\boldsymbol{r}$ in $\mathcal{O}(NM + NL)$. Given $\boldsymbol{r}$, we can compute $\boldsymbol{s} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{r} = \sum_n(\boldsymbol{x}_n \otimes \boldsymbol{e}_n)r_n$ in $\mathcal{O}(NML)$ time so that $\hat{\theta}$ can be computed in $\mathcal{O}(NML)$ time. Given B random vectors, we can compute $\theta_B = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{w}_b^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_b$ in $\mathcal{O}(NMLB)$ time. □

3.5.3 Alternate estimator of $tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})$ in $\mathcal{O}(N(M+L)B)$ time

Given mutually independent random vectors $\boldsymbol{u} \in \mathbb{R}^M, \boldsymbol{v} \in \mathbb{R}^L$, $\boldsymbol{z} \in \mathbb{R}^L$ whose components are i.i.d. random variables each with mean zero and variance one. Let $\boldsymbol{b} = \boldsymbol{X}\boldsymbol{u}$, $\boldsymbol{c} = \boldsymbol{E}\boldsymbol{v}$, $\boldsymbol{F} = diag(\boldsymbol{b})\boldsymbol{X}$, and $\boldsymbol{G} = diag(\boldsymbol{c})\boldsymbol{E}$.

Lemma 5 $\hat{\theta} = \boldsymbol{F}^T\boldsymbol{G}^2$

Proof:

$$\begin{aligned}
\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G} &= \boldsymbol{X}^{\mathrm{T}} diag(\boldsymbol{bc})\boldsymbol{E} \\
&= \sum_{n} b_n c_n \boldsymbol{x}_n \boldsymbol{e}_n^{\mathrm{T}}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}^2 &= tr\left(\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\right)^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\right) \\
&= tr\left(\left(\sum_{n'} b_{n'} c_{n'} \boldsymbol{x}_{n'} \boldsymbol{e}_{n'}^{\mathrm{T}}\right)^{\mathrm{T}} \sum_{n} b_n c_n \boldsymbol{x}_n \boldsymbol{e}_n^{\mathrm{T}}\right) \\
&= tr\left(\sum_{n}\sum_{n'} b_n c_n b_{n'} c_{n'} \boldsymbol{e}_{n'} \boldsymbol{x}_{n'}^{\mathrm{T}} \boldsymbol{x}_n \boldsymbol{e}_n^{\mathrm{T}}\right) \\
&= \sum_{n}\sum_{n'} b_n c_n b_{n'} c_{n'} tr\left(\boldsymbol{e}_{n'} \boldsymbol{x}_{n'}^{\mathrm{T}} \boldsymbol{x}_n \boldsymbol{e}_n^{\mathrm{T}}\right) \\
&= \sum_{n}\sum_{n'} b_n c_n b_{n'} c_{n'} \boldsymbol{e}_n^{\mathrm{T}} \boldsymbol{e}_{n'} \boldsymbol{x}_{n'}^{\mathrm{T}} \boldsymbol{x}_n \\
&= \sum_{n}\sum_{n'} r_n r_{n'} \boldsymbol{e}_n^{\mathrm{T}} \boldsymbol{e}_{n'} \boldsymbol{x}_n^{\mathrm{T}} \boldsymbol{x}_{n'} \\
&= \sum_{n}\sum_{n'} r_n r_{n'} \boldsymbol{e}_n^{\mathrm{T}} \boldsymbol{e}_{n'} \boldsymbol{x}_n^{\mathrm{T}} \boldsymbol{x}_{n'} \\
&= \sum_{n}\sum_{n'} r_n r_{n'} \left(\boldsymbol{e}_n^{\mathrm{T}} \otimes \boldsymbol{x}_n^{\mathrm{T}}\right)\left(\boldsymbol{e}_{n'} \otimes \boldsymbol{x}_{n'}\right), \qquad \text{Mixed-product property} \\
&= \sum_{n}\sum_{n'} r_n r_{n'} (\boldsymbol{e}_n \otimes \boldsymbol{x}_n)^{\mathrm{T}}\left(\boldsymbol{e}_{n'} \otimes \boldsymbol{x}_{n'}\right) \\
&= \left(\sum_{n} r_n (\boldsymbol{e}_n \otimes \boldsymbol{x}_n)^{\mathrm{T}}\right)\left(\sum_{n'} r_{n'}\left(\boldsymbol{e}_{n'} \otimes \boldsymbol{x}_{n'}\right)\right) \\
&= \boldsymbol{s}^{\mathrm{T}}\boldsymbol{s} = \hat{\theta}
\end{aligned}$$

□

Theorem 3 $\tilde{\theta} \equiv \boldsymbol{F}^T\boldsymbol{G}\boldsymbol{z}^2$ *is an unbiased estimator of* $tr(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{H})$.

Proof:

$$\begin{aligned}
\mathbb{E}\left[\tilde{\theta}|\boldsymbol{u},\boldsymbol{v}\right] &= \mathbb{E}\left[\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}^{2}|\boldsymbol{u},\boldsymbol{v}\right] \\
&= \mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v}\right] \\
&= \mathbb{E}\left[tr\left(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\right)|\boldsymbol{u},\boldsymbol{v}\right] \\
&= \mathbb{E}\left[tr\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\right)|\boldsymbol{u},\boldsymbol{v}\right] \\
&= tr\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}|\boldsymbol{u},\boldsymbol{v}\right]\right) \\
&= tr\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}\right) \\
&= \boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}^{2} \\
&= \hat{\theta}, \qquad \text{Lemma 5}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\tilde{\theta}\right] &= \mathbb{E}\left[\mathbb{E}\left[\tilde{\theta}|\boldsymbol{u},\boldsymbol{v}\right]\right] \\
&= \mathbb{E}\left[\hat{\theta}\right] \\
&= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})
\end{aligned}$$

□

We can compute $\boldsymbol{b}$ and $\boldsymbol{F}$ in $\mathcal{O}(NM)$ time while $\boldsymbol{c}$ and $\boldsymbol{G}$ can be computed in $\mathcal{O}(NL)$ time. $\boldsymbol{z}_1 = \boldsymbol{G}\boldsymbol{z}$ can be computed in $\mathcal{O}(NL)$ time while $\boldsymbol{z}_2 = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{z}_1$ can be computed in $\mathcal{O}(NM)$ time so that $\tilde{\theta} = \|\boldsymbol{z}_2\|_2^2$ can be computed in $\mathcal{O}(M)$ time. Thus, $\tilde{\theta}$ can be computed in $\mathcal{O}(N(M+L))$ time.

3.5.4 Computation of other terms involving $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$

Lemma 6 *Suppose that $\boldsymbol{X}$ and $\boldsymbol{E}$ are $N \times M$ genotype and $N \times L$ environment matrices respectively. Define $\boldsymbol{H} = \boldsymbol{X} \odot \boldsymbol{E}$ as the $N \times ML$. Assume that $\boldsymbol{w}_t = \boldsymbol{u}_t \otimes \boldsymbol{v}_t$ where $\boldsymbol{u}_t \in \mathbb{R}^M, \boldsymbol{v}_t \in \mathbb{R}^L$ are independent random vectors with mean zero and covariance $\boldsymbol{I}_M$ and $\boldsymbol{I}_L$ respectively. Let $\boldsymbol{r}_t = \boldsymbol{H}\boldsymbol{w}_t$. Then $(\boldsymbol{y}^T\boldsymbol{r}_t)^2$ and $\|\boldsymbol{X}^T\boldsymbol{r}_t\|_2^2$ are unbiased estimators of $\boldsymbol{y}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{y}$ and $tr(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{H}^T)$ respectively.*

Proof:

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{r}_t)^2\right] &= \mathbb{E}\left[(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{r}_t)(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{r}_t)\right] \\
&= \mathbb{E}\left[(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t)(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t)\right] \\
&= \mathbb{E}\left[(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y})(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t)\right] \\
&= \mathbb{E}\left[tr(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t)\right] \\
&= \mathbb{E}\left[tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}})\right] \qquad \text{(cyclic property of trace)} \\
&= tr(\mathbb{E}\left[\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}}\right]) \qquad \text{(linearity of trace and expectation)} \\
&= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\mathbb{E}\left[\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}}\right]) \\
&= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}) \qquad (\mathbb{E}\left[\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}}\right] = \boldsymbol{I}) \\
&= tr(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}) \qquad \text{(cyclic property of trace)} \\
&= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y}
\end{aligned}$$

Further, we have :

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}_t\|_2^2\right] &= \mathbb{E}\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}_t)^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}_t)\right] \\
&= \mathbb{E}\left[\boldsymbol{r}_t^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}_t\right] \\
&= \mathbb{E}\left[(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X})(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t)\right] \\
&= \mathbb{E}\left[tr(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t)\right] \\
&= \mathbb{E}\left[tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}})\right] \qquad \text{(cyclic property of trace)} \\
&= tr(\mathbb{E}\left[\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}}\right]) \qquad \text{(linearity of trace and expectation)} \\
&= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\mathbb{E}\left[\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}}\right]) \\
&= tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}) \qquad (\mathbb{E}\left[\boldsymbol{w}_t\boldsymbol{w}_t^{\mathrm{T}}\right] = \boldsymbol{I}) \\
&= tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) \qquad \text{(cyclic property of trace)}
\end{aligned}$$

□

3.5.5 Computing the standard errors of the estimates

$$(\tilde{\boldsymbol{\sigma^2}}) = argmin_{\boldsymbol{\sigma^2}}||\boldsymbol{y}\boldsymbol{y}^T - \sum_{k=1}^{K} \sigma_k^2 \boldsymbol{K}_k||_F^2 \tag{3.21}$$

where $\boldsymbol{K}_K = \boldsymbol{I}_N$ and $\boldsymbol{K}_i = \frac{\boldsymbol{X}_i \boldsymbol{X}_i^T}{M_i}$ for $i = 1, .., K-1$. The MoM estimator satisfies the following normal equations:

$$\boldsymbol{T} = \boldsymbol{\sigma^2}\boldsymbol{q} \tag{3.22}$$

$\boldsymbol{T}$ is a $K \times K$ matrix with entries $T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l), k, l \in \{1, \ldots, K\}$, and $\boldsymbol{q}$ is a K-vector with entries $q_k = \boldsymbol{y}^T \boldsymbol{K}_k \boldsymbol{y}$. We have:

$$\tilde{\boldsymbol{\sigma^2}} \;\; = \;\; \boldsymbol{T}^{-1}\boldsymbol{q} \tag{3.23}$$

The covariance matrix of $\tilde{\sigma^2}$ is:

$$\text{Cov}\left[\tilde{\boldsymbol{\sigma^2}}\right] \;\; = \;\; \boldsymbol{T}^{-1}\text{Cov}\left[\boldsymbol{q}\right]\boldsymbol{T}^{-1}$$

we have :

$$\text{Cov}\left[\boldsymbol{q}\right] = E[\boldsymbol{q}\boldsymbol{q}^T] - E[\boldsymbol{q}]E[\boldsymbol{q}]^T$$

where :

$$\text{Cov}\left[\boldsymbol{q}\right]_{ij} \;\; = \;\; E[\boldsymbol{y}^T\boldsymbol{K}_i\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{K}_j\boldsymbol{y}] - E[\boldsymbol{y}^T\boldsymbol{K}_i\boldsymbol{y}]E[\boldsymbol{y}^T\boldsymbol{K}_j\boldsymbol{y}]$$

Lemma 7 *For a random vector $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$ and symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have :*

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

Proof:

$$\text{Cov}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}, \boldsymbol{z}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{z}\right] \;\; = \;\; \mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{z}\right] - \mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}\right]\mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{z}\right] \tag{3.24}$$

$$
\begin{align}
\mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{z}\right] &= \mathbb{E}\left[\left(\sum_{i,j} z_i A_{ij} z_j\right)\left(\sum_{k,l} z_k B_{kl} z_l\right)\right] \notag \\
&= \mathbb{E}\left[\sum_{i,j,k,l} A_{ij} B_{kl} z_i z_j z_k z_l\right] \notag \\
&= \sum_{i,j,k,l} A_{ij} B_{kl} \mathbb{E}\left[z_i z_j z_k z_l\right] \tag{3.25} \\
&= \sum_{i,j,k,l} A_{ij} B_{kl} [\mathbb{E}\left[z_i z_j\right]\mathbb{E}\left[z_k z_l\right] + \mathbb{E}\left[z_i z_k\right]\mathbb{E}\left[z_j z_l\right] + \mathbb{E}\left[z_i z_l\right]\mathbb{E}\left[z_k z_j\right]] \tag{3.26} \\
&= \sum_{i,j,k,l} A_{ij} B_{kl} [C_{ij} C_{kl} + C_{ik} C_{jl} + C_{il} C_{jk}] \tag{3.27} \\
&= \sum_{i,j,k,l} A_{ij} B_{kl} C_{ij} C_{kl} + \sum_{i,j,k,l} A_{ij} B_{kl} C_{ik} C_{jl} + \sum_{i,j,k,l} A_{ij} B_{kl} C_{il} C_{jk} \notag \\
&= \sum_{i,j,k,l} A_{ij} C_{ij} B_{kl} C_{kl} + \sum_{i,j,k,l} A_{ij} C_{ik} B_{kl} C_{jl} + \sum_{i,j,k,l} A_{ij} C_{il} B_{kl} C_{jk} \notag \\
&= \sum_{i,j} A_{ij} C_{ij} \sum_{k,l} B_{kl} C_{kl} + \sum_{i,j,k,l} A_{ij} C_{ik} B_{kl} C_{jl} + \sum_{i,j,k,l} A_{ij} C_{ik} B_{lk} C_{jl} \tag{3.28} \\
&= tr(\boldsymbol{AC}) tr(\boldsymbol{BC}) + 2 \sum_{i,j,k,l} A_{ij} C_{ik} B_{kl} C_{jl} \tag{3.29} \\
&= tr(\boldsymbol{AC}) tr(\boldsymbol{BC}) + 2 \sum_{j,k} \left(\sum_{i} A_{ji} C_{ik}\right)\left(\sum_{l} B_{kl} C_{lj}\right) \tag{3.30} \\
&= tr(\boldsymbol{AC}) tr(\boldsymbol{BC}) + 2 \sum_{j,k} (\boldsymbol{AC})_{jk} (\boldsymbol{BC})_{kj} \tag{3.31} \\
&= tr(\boldsymbol{AC}) tr(\boldsymbol{BC}) + 2 tr(\boldsymbol{ACBC}) \tag{3.32}
\end{align}
$$

Equation 3.25 follows by linearity of expectation while Equation 3.26 follows from an application of Isserlis' theorem and Equation 3.27 follows from the fact that $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$. Equation 3.28 follows by grouping factors in the first term and by interchanging indices l and k in the second term. Equation 3.29 follows from the fact that $\boldsymbol{B}$ is symmetric so that $B_{kl} = B_{lk}$ so that the last two terms are identical. Equation 3.30 uses the fact that matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ are symmetric. Equation 3.31 is the definition of matrix multiplication, and Equation 3.32 follows from the definition of the trace.

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}\right] &= \mathbb{E}\left[\sum_{i,j} z_i A_{ij} z_j\right] \\
&= \mathbb{E}\left[\sum_{i,j} A_{ij} z_i z_j\right] \\
&= \sum_{i,j} A_{ij}\mathbb{E}\left[z_i z_j\right] \\
&= \sum_{i,j} A_{ij} C_{ij} \\
&= tr(\boldsymbol{A}\boldsymbol{C}) \qquad (3.33)
\end{aligned}$$

Equation 3.24 follows by combining Equations 3.32 and 3.33. □

By using Lemma 7 and the following fact $E(xy) = cov(x, y) + E(x)E(y)$ we have :

$$\begin{aligned}
\mathrm{Cov}\left[\boldsymbol{q}\right]_{ij} &= 2tr(\boldsymbol{\Sigma}\boldsymbol{K}_i\boldsymbol{\Sigma}\boldsymbol{K}_j) + tr(\boldsymbol{K}_i\boldsymbol{\Sigma})tr(\boldsymbol{K}_j\boldsymbol{\Sigma}) - (tr(\boldsymbol{K}_i\boldsymbol{\Sigma})tr(\boldsymbol{K}_j\boldsymbol{\Sigma})) \\
&= 2tr(\boldsymbol{\Sigma}\boldsymbol{K}_i\boldsymbol{\Sigma}\boldsymbol{K}_j)
\end{aligned}$$

Replacing $\boldsymbol{\Sigma}$ by its estimate, $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ and $\boldsymbol{\Sigma}$ by its estimate $\tilde{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \tilde{\sigma}_k^2 \boldsymbol{K}_k$, the plug-in estimate of $\mathrm{Cov}\left[\boldsymbol{q}\right]_{kl}$:

$$\begin{aligned}
\widehat{\mathrm{Cov}\left[\boldsymbol{q}\right]}_{kl} &= 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_k\tilde{\boldsymbol{\Sigma}}\boldsymbol{K}_l\boldsymbol{y} \\
&= 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_k\left(\sum_{t=1}^{K}\tilde{\sigma}_t^2\boldsymbol{K}_t\right)\boldsymbol{K}_l\boldsymbol{y} \\
&= 2\left(\sum_{t=1}^{K}\tilde{\sigma}_t^2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_k\boldsymbol{K}_t\boldsymbol{K}_l\boldsymbol{y}\right) \\
&= 2\sum_{t=1}^{K}\tilde{\sigma}_t^2\left(\boldsymbol{w}_k^{\mathrm{T}}\frac{\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}}{M_t}\boldsymbol{w}_l\right)
\end{aligned}$$

where $\boldsymbol{w}_t = \boldsymbol{K}_t\boldsymbol{y}, t \in \{1, \ldots, K\}$. Therefore, $\widehat{\mathrm{Cov}\left[\boldsymbol{q}\right]}_{kl}$ can be computed in time $\mathcal{O}(\frac{NMK^2}{\max(\log_3(N),\log_3(M))})$.

3.6 Simulations of MAF and LD-dependent genomic architectures

To simulate MAF and LD-dependent architectures, we simulated phenotypes from genotypes using the following model that extends prior models of additive genetic architecture [18, 45] to include GxE effects:

$$
\begin{aligned}
\sigma^2_{g,m} &= S c_m w_m^b [f_m(1-f_m)]^a \\
(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, .., \boldsymbol{\beta}_m)^T &\sim \mathcal{N}(\mathbf{0}, diag(\sigma^2_{g,1}, \sigma^2_{g,2}, ..., \sigma^2_{g,m})) \\
\sigma^2_{gxe,m} &= S^{'} c^{'}_m w_m^b [f_m(1-f_m)]^a \\
(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, .., \boldsymbol{\alpha}_m)^T &\sim \mathcal{N}(\mathbf{0}, diag(\sigma^2_{gxe,1}, \sigma^2_{gxe,2}, ..., \sigma^2_{gxe,m})) \qquad (3.34) \\
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\alpha} &\sim \mathcal{N}(\boldsymbol{X\beta} + (\boldsymbol{X} \odot \boldsymbol{E})\boldsymbol{\alpha}, (1 - h^2_g - h^2_{gxe} - h^2_{nxe})\boldsymbol{I}_N + h^2_{nxe}(\boldsymbol{I}_N \odot \boldsymbol{E}))
\end{aligned}
$$

where $h^2_g, h^2_{gxe}, h^2_{nxe} \in [0,1]$, $a \in \{0, 0.75\}$, $b \in \{0,1\}$. Here S and $S^{'}$ are normalizing constants chosen so that $\sum_{m=1}^{M} \sigma^2_{g,m} = h^2_g$, $\sum_{m=1}^{M} \sigma^2_{gxe,m} = h^2_{gxe}$. Additive and GxE effect sizes are denoted by $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ respectively. f_m and w_m are the minor allele frequency and LD score of m^{th} SNP respectively. In this model, $c_m, c^{'}_m \in \{0,1\}$ are indicator variables for the causal status of SNP m ($c_m = 1$ and $c^{'}_m = 1$ for all SNPs). The LD score of a SNP is defined to be the sum of the squared correlation of the SNP with all other SNPs that lie within a specific distance. Setting $a = 0, b = 0$ for additive effects results in the GCTA model where the per-standardized genotypic effect sizes at a SNP do not vary with MAF or LD score [134, 3] while setting $a = 0.75, b = 1$ results in the LDAK model [109]. Our simulations assume that the GxE effects follow the same coupling as the additive effects. We used phenotypes from $N = 40k$ individuals genotyped at $M = 459,792$ SNPs on the UKBB array.

3.7 Supplementary Figures

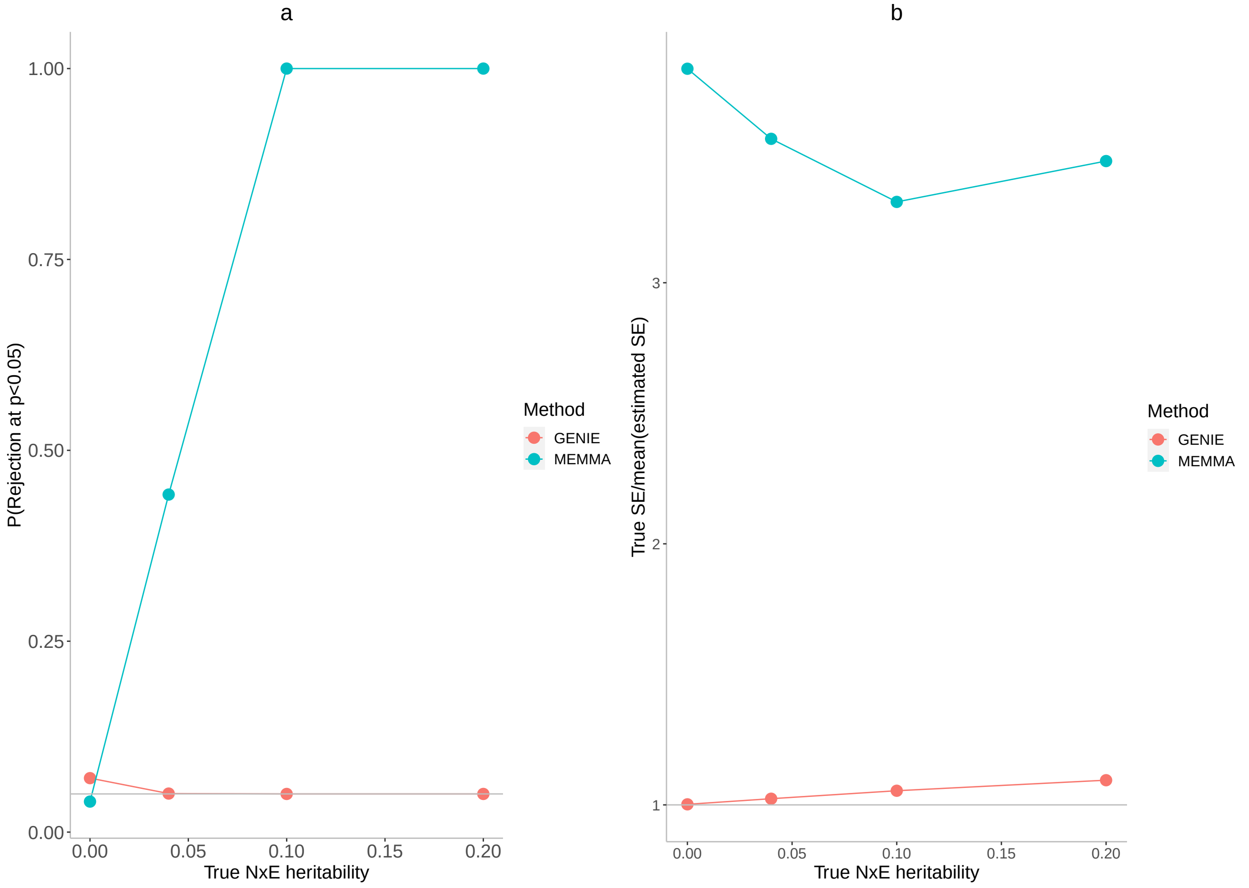


Figure 3.10: **Effect of an estimated standard error on controlling False positive rate** . **a**) We assessed the calibration of GENIE and MEMMA using their true SE instead of the estimation of SE in simulations. MEMMA has biased estimates of SE, leading to a high false positive rate even in the absence of a NxE effect. **b**) We plot the ratio of true SE over the mean of estimated SE across 100 replicates as a function of the variance of the NxE effect.

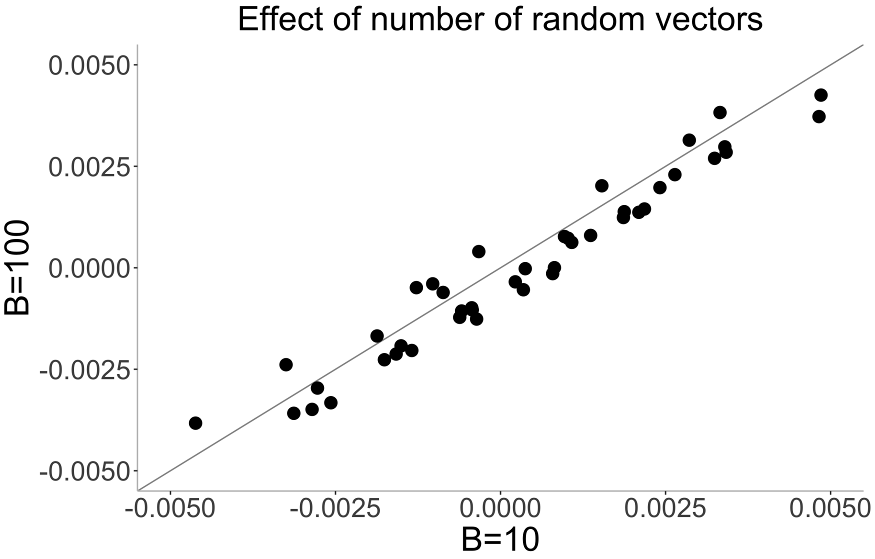


Figure 3.11: **Comparison of h^2_{gxe} estimates with $B = 10$ and $B = 100$ on large scale data**. We simulated phenotypes from $M = 459,792$ array SNPs and $N = 291,273$ individuals where $h^2_g = 0.25$, $h^2_{gxe} = 0$ and the causal ratio is 10% .

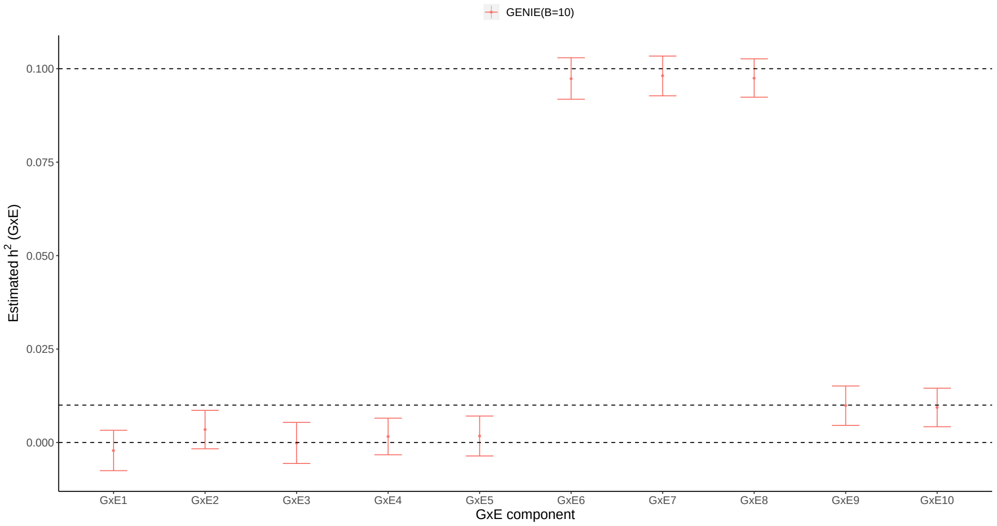


Figure 3.12: **Accuracy of GENIE when applied to multiple environmental variables**. We evaluated h^2_{gxe} estimates of GENIE for different values of σ^2_{gxe}. We simulated phenotypes with 10 environmental variables where $\sigma^2_g = 0.2$, $\sigma^2_{ge1} = \sigma^2_{ge2} = \sigma^2_{ge3} = \sigma^2_{ge4} = \sigma^2_{ge5} = 0$, $\sigma^2_{ge6} = \sigma^2_{ge7} = \sigma^2_{ge8} = 0.10$ and $\sigma^2_{ge9} = \sigma^2_{ge10} = 0.01$. Points and error bars represent the mean and ±2 SE. Mean and SE are computed from 100 replicates. Here B is the number of random vectors used by GENIE with $B = 10$ the value that we use as default (we reported values of means, SEs, and p-values of a test of the null hypothesis of no bias in the estimates of variance components in Supplementary Table 1)

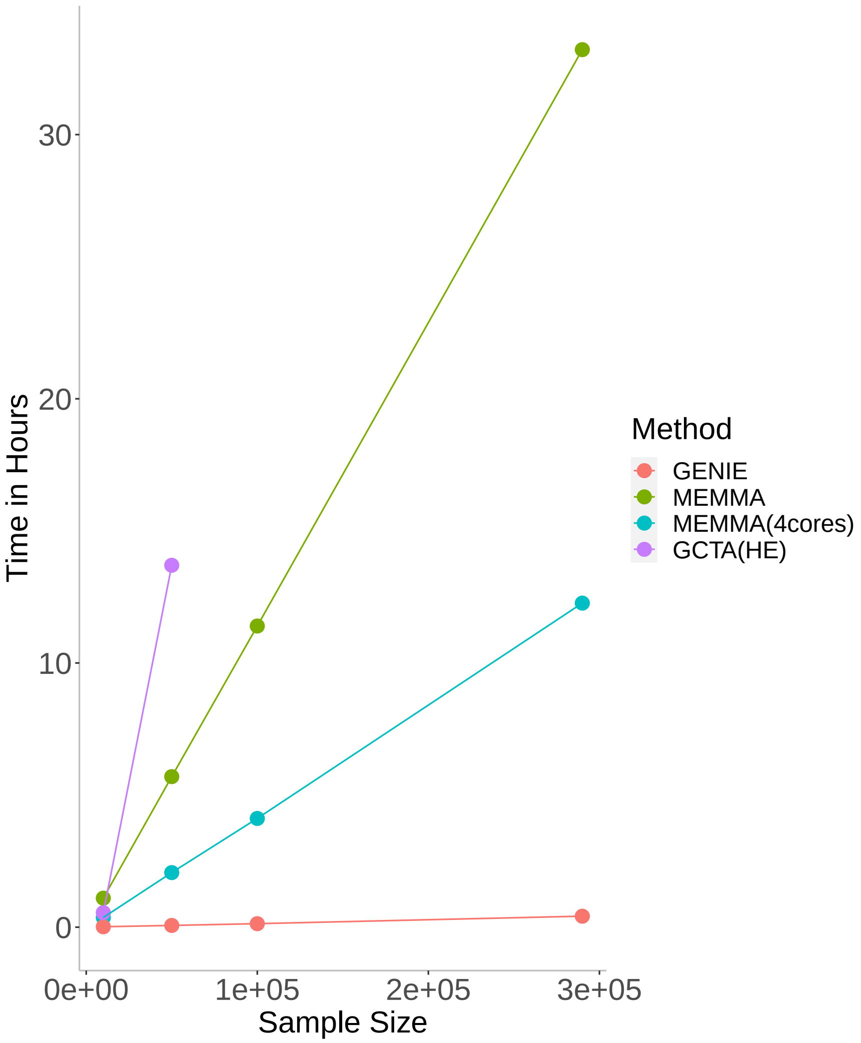


Figure 3.13: **Runtime scaling of GENIE**. We evaluated the runtime of GENIE, MEMMA, and GCTA(HE) with increasing sample size N (for a fixed number of SNPs $M = 459,792$ and single environmental variable). We fit single G and GxE variance components. All methods were run on an Intel(R) Xeon(R) Gold 6140 CPU 2.30GHz with 187 GB RAM. Ten random vectors are used by GENIE and MEMMA. The runtime of GCTA(HE) includes the computation of the GRM. GENIE and GCTA(HE) are executed on a single core while MEMMA is run on both a single core and four cores.

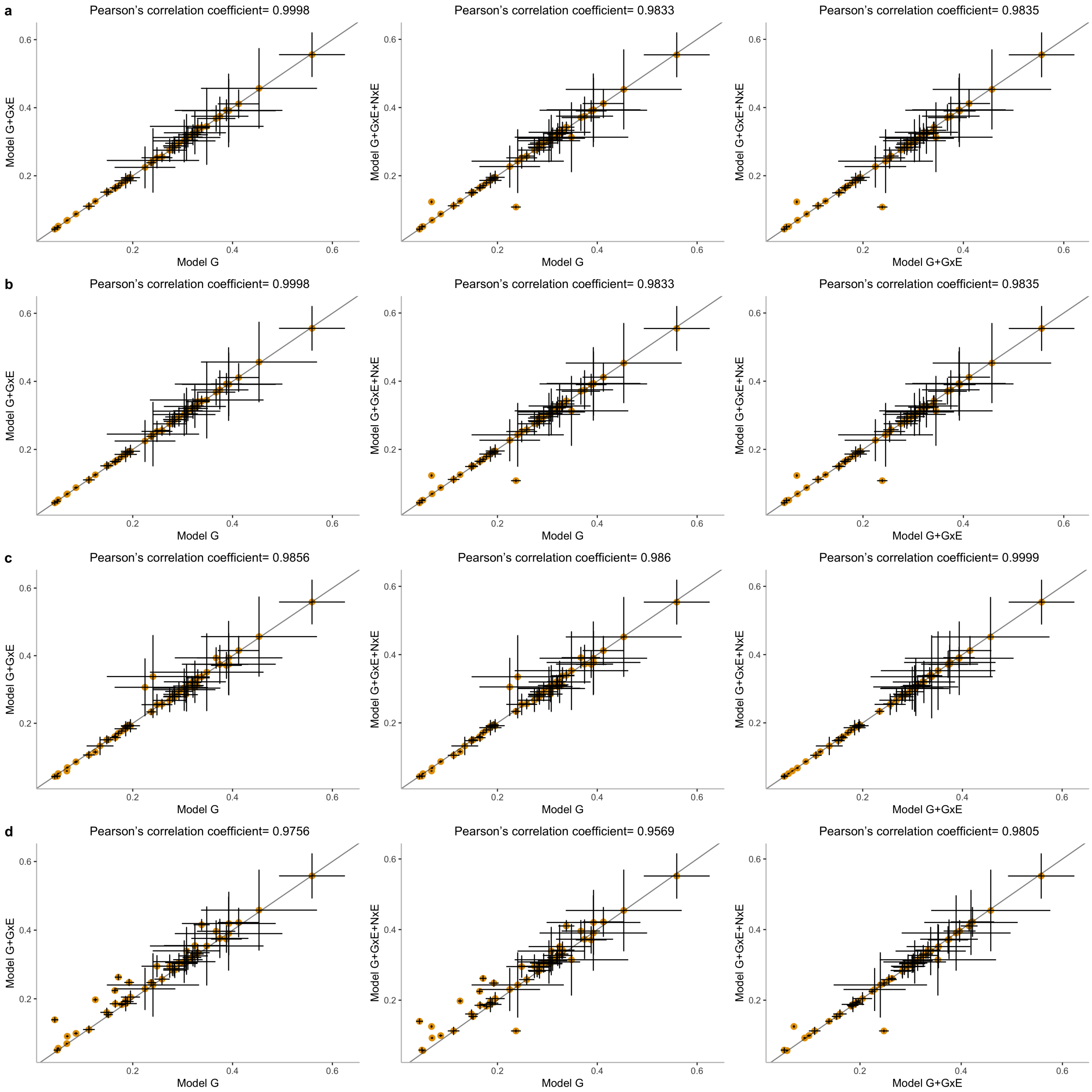


Figure 3.14: **Estimated additive variance component from three different models.** In this figure, G, GxE and NxE refer to additive, gene-by-environment, and noise-by-environment components, respectively. Every model is named by a set of variance components fitted jointly under that model. Estimates of additive components under the three models where the environmental variable is **a**) smoking status, **b**) sex, **c**) statin usage, and **d**) age. The estimates of additive heritability obtained by GENIE are consistent under these three models across environmental variables.

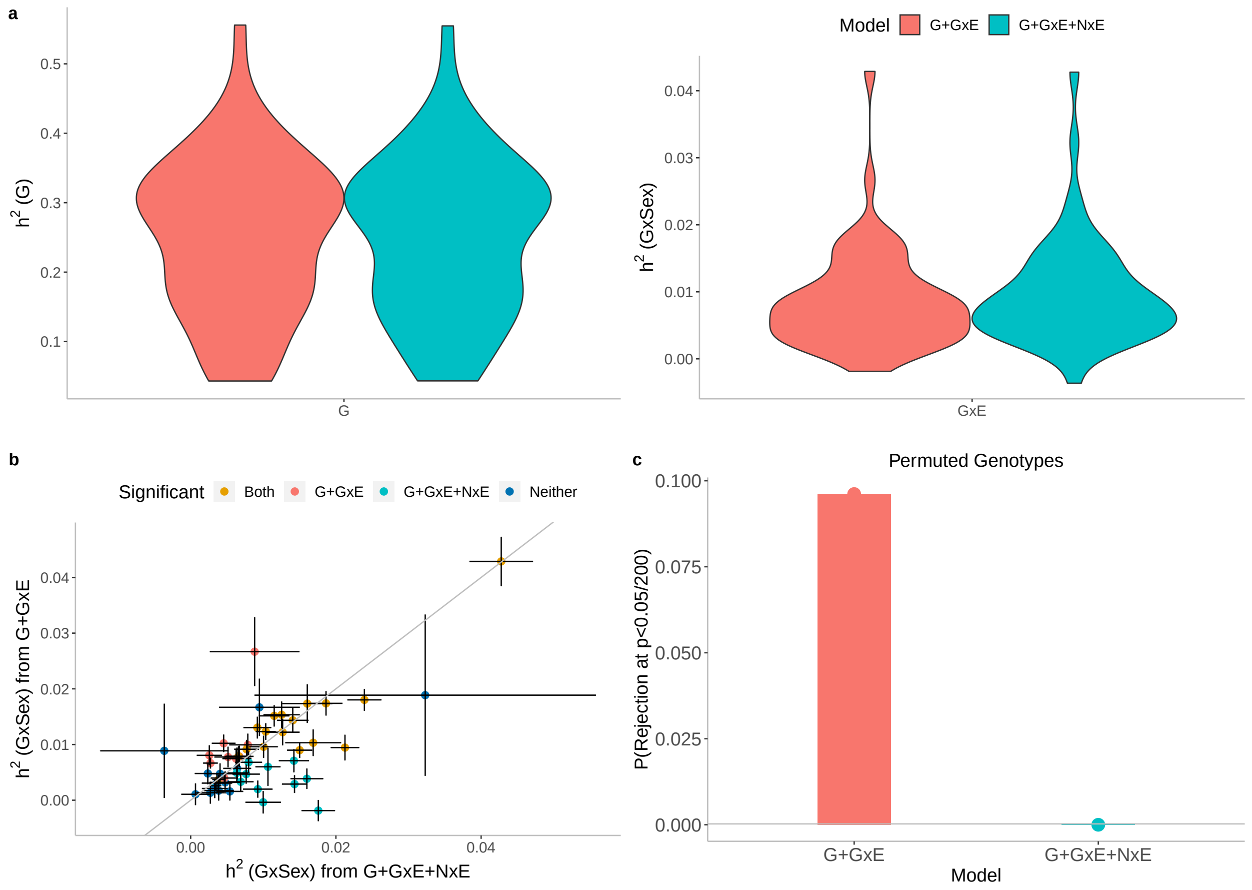


Figure 3.15: **Effect of Noise heterogeneity (NxE) on estimates of heritability associated with GxSex across 50 quantitative phenotypes in UKBB**. Model G+GxE refers to a model with additive and gene-by-environment interaction components. Model G+GxE+NxE refers to a model with additive, gene-by-environment interaction and environmental heterogeneity (noise-by-environment interaction) components. **a**) We run GENIE under G+GxE and G+GxE+NxE models to assess the effect of fitting a NxE component on the GxE and additive heritability estimates. **b**) Comparison of GxE heritability estimates obtained from GENIE under a G+GxE+NxE model (x-axis) to a G+GxE model (y-axis). Black error bars mark ± standard errors centered on the estimated GxE heritability. **c**) We performed permutation analyses by randomly shuffling, in turn, the environmental values, phenotypes or genotypes and applied GENIE in each setting under G+GxE and G+GxE+NxE models. We report the fraction of rejections (p-value of a test of the null hypothesis of zero GxE heritability < 0.001 that accounts for the number of phenotypes tested) over 50 UKBB phenotypes.

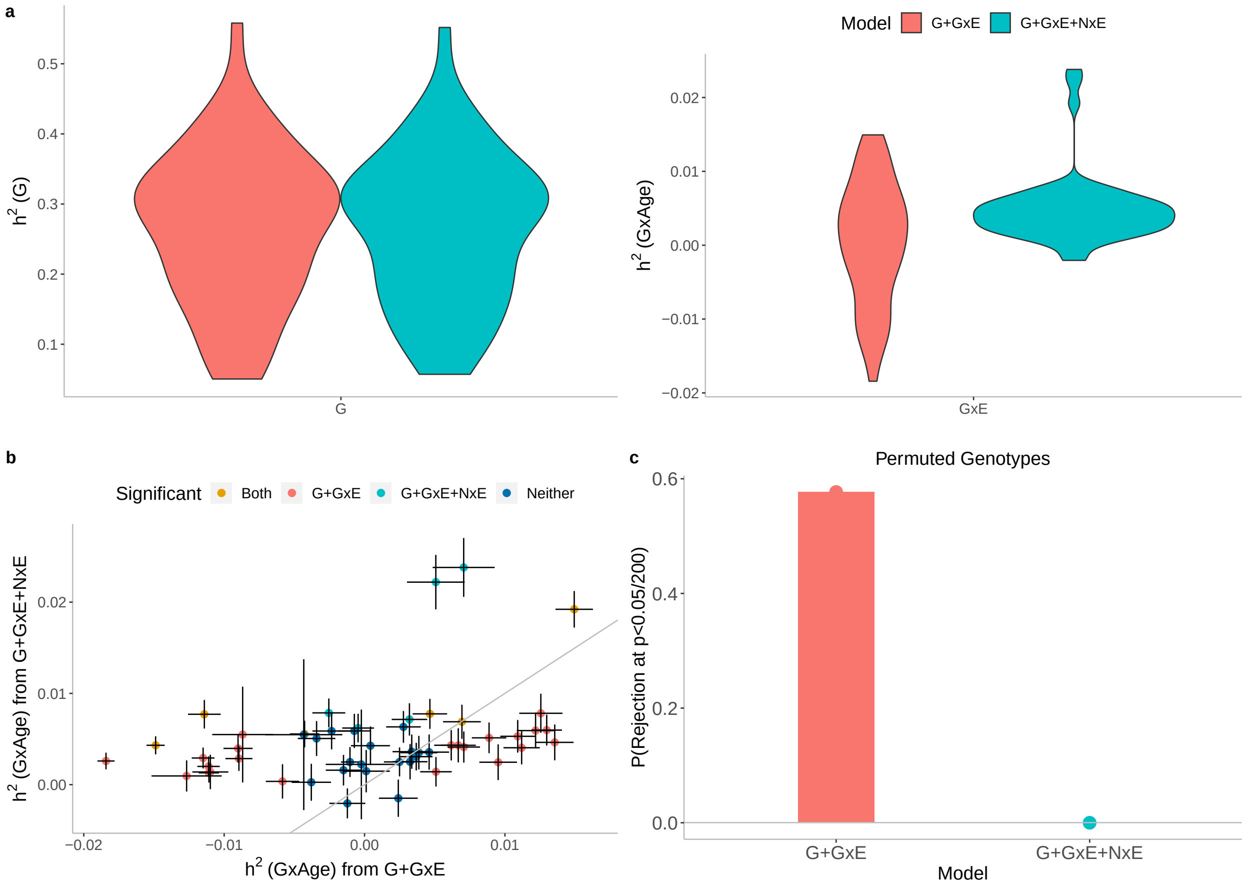


Figure 3.16: **Effect of Noise heterogeneity (NxE) on estimates of heritability associated with GxAge across 50 quantitative phenotypes in UKBB**. Model G+GxE refers to a model with additive and gene-by-environment interaction components. Model G+GxE+NxE refers to a model with additive, gene-by-environment interaction and environmental heterogeneity (noise-by-environment interaction) components. **a**) We run GENIE under G+GxE and G+GxE+NxE models to assess the effect of fitting an NxE component on the GxE and additive heritability estimates. **b**) Comparison of GxE heritability estimates obtained from GENIE under a G+GxE+NxE model (x-axis) to a G+GxE model (y-axis). Black error bars mark ± standard errors centered on the estimated GxE heritability. Color of the dots indicate whether estimates of GxE heritability are significant under each model. **c**) We performed permutation analyses by randomly shuffling the genotypes while preserving the trait-E relationship and applied GENIE in each setting under G+GxE and G+GxE+NxE models. We report the fraction of rejections (p-value of a test of the null hypothesis of zero GxE heritability $< \frac{0.05}{200}$ that accounts for the number of phenotypes tested) over 50 UKBB phenotypes.

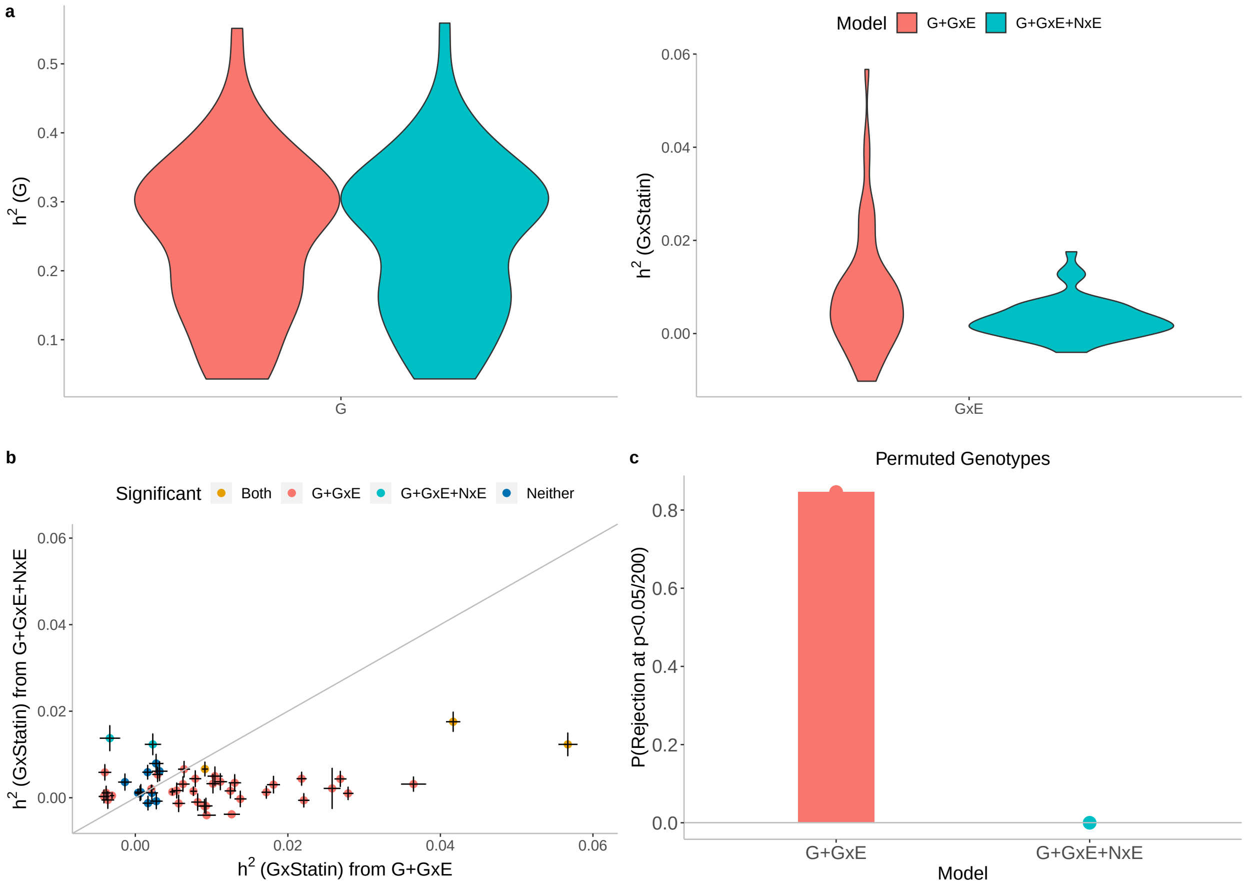


Figure 3.17: **Effect of Noise heterogeneity (NxE) on estimates of heritability associated with GxStatin across 50 quantitative phenotypes in UKBB**. Model G+GxE refers to a model with additive and gene-by-environment interaction components. Model G+GxE+NxE refers to a model with additive, gene-by-environment interaction and environmental heterogeneity (noise-by-environment interaction) components. **a**) We run GENIE under G+GxE and G+GxE+NxE models to assess the effect of fitting a NxE component on the GxE and additive heritability estimates. **b**) Comparison of GxE heritability estimates obtained from GENIE under a G+GxE+NxE model (x-axis) to a G+GxE model (y-axis). Black error bars mark ± standard errors centered on the estimated GxE heritability. **c**) We performed permutation analyses by randomly shuffling, in turn, the environmental values, phenotypes or genotypes and applied GENIE in each setting under G+GxE and G+GxE+NxE models. We report the fraction of rejections (p-value of a test of the null hypothesis of zero GxE heritability < 0.001 that accounts for the number of phenotypes tested) over 50 UKBB phenotypes.

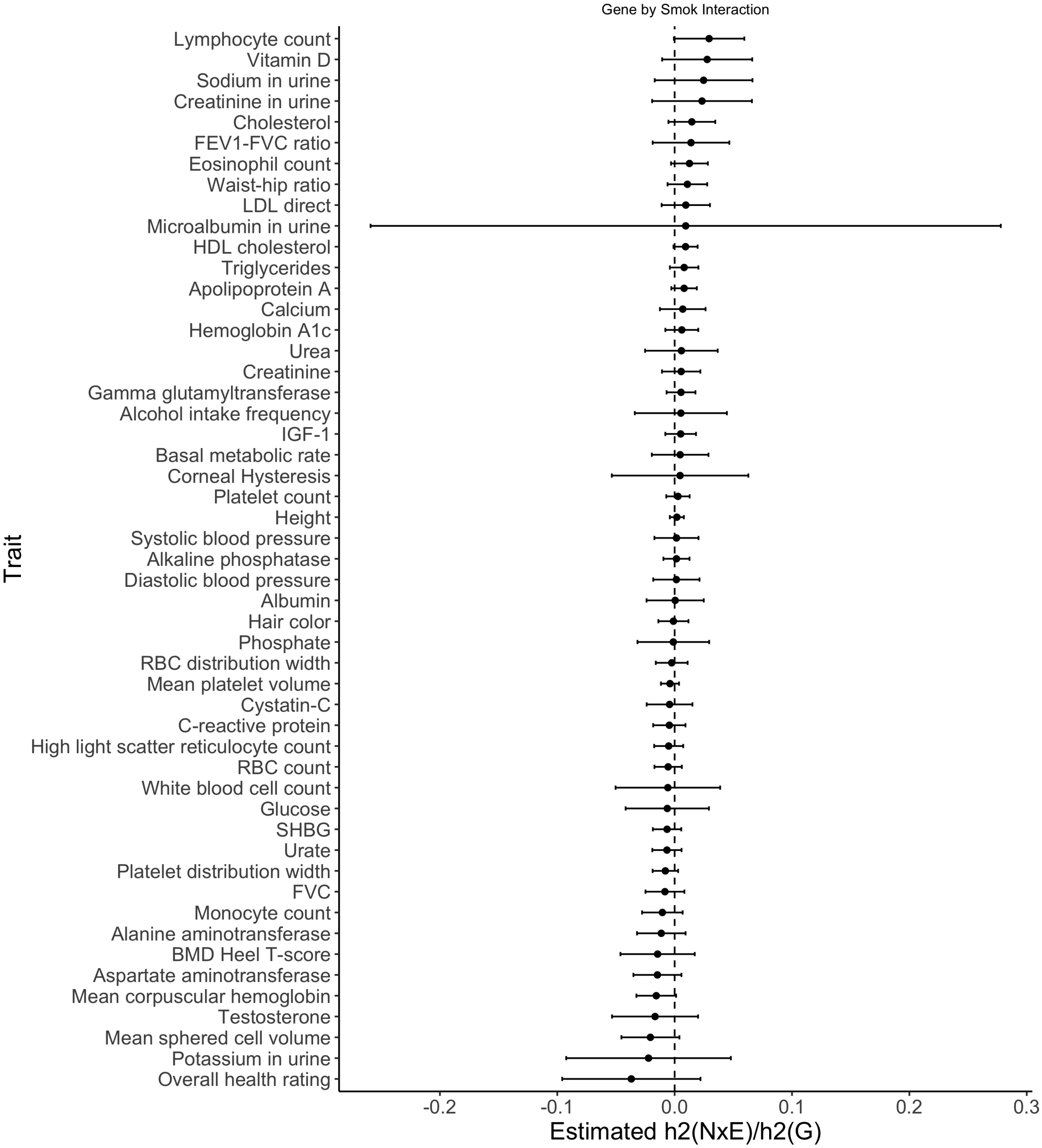


Figure 3.18: **Estimated ratio of variance attributed to noise heterogeneity over additive heritability for Smoking**. Black error bars mark ±2 standard errors centered on the estimated ratio.

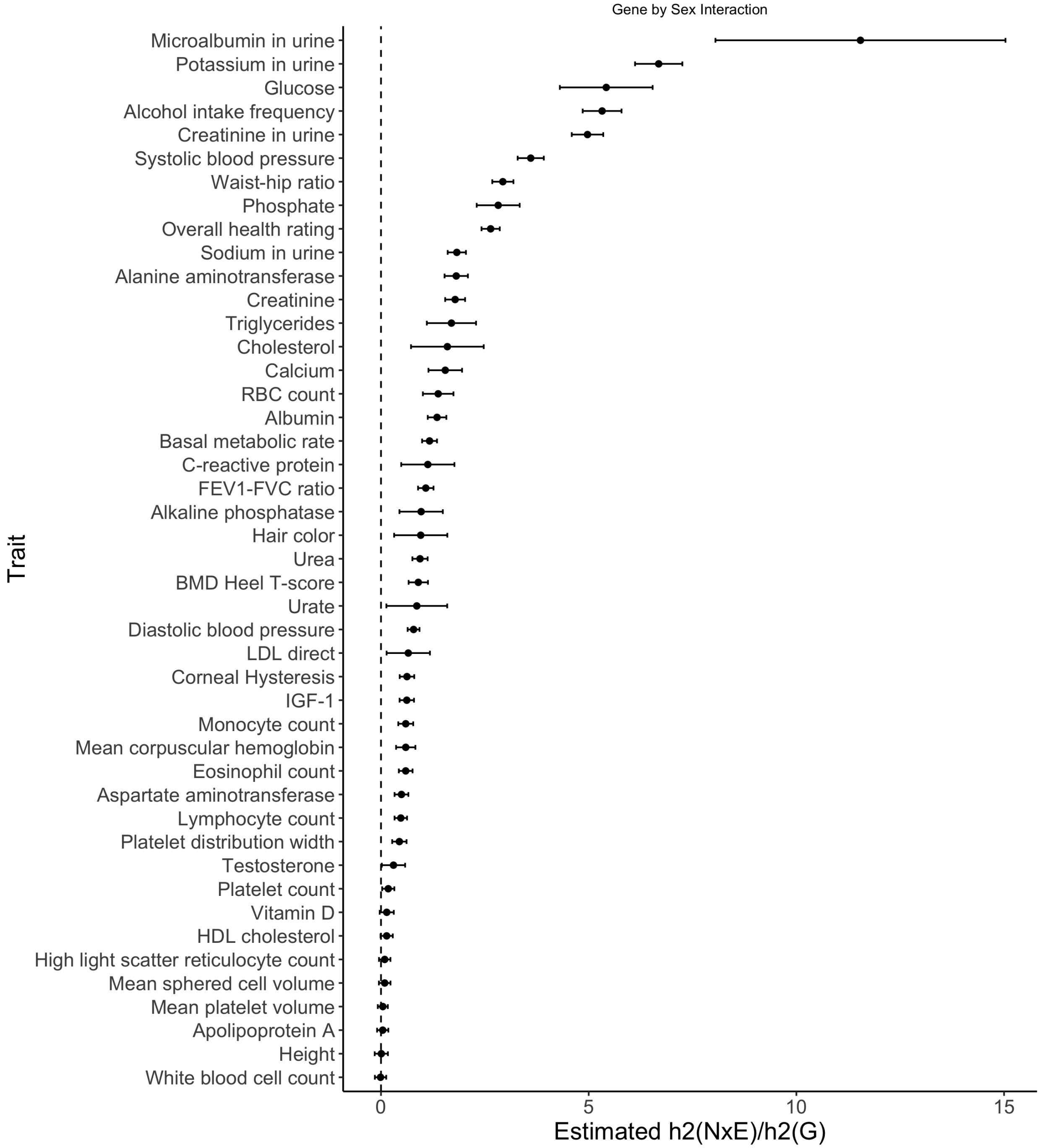


Figure 3.19: **Estimated ratio of variance attributed to noise heterogeneity over additive heritability for Sex**. Black error bars mark ±2 standard errors centered on the estimated ratio.

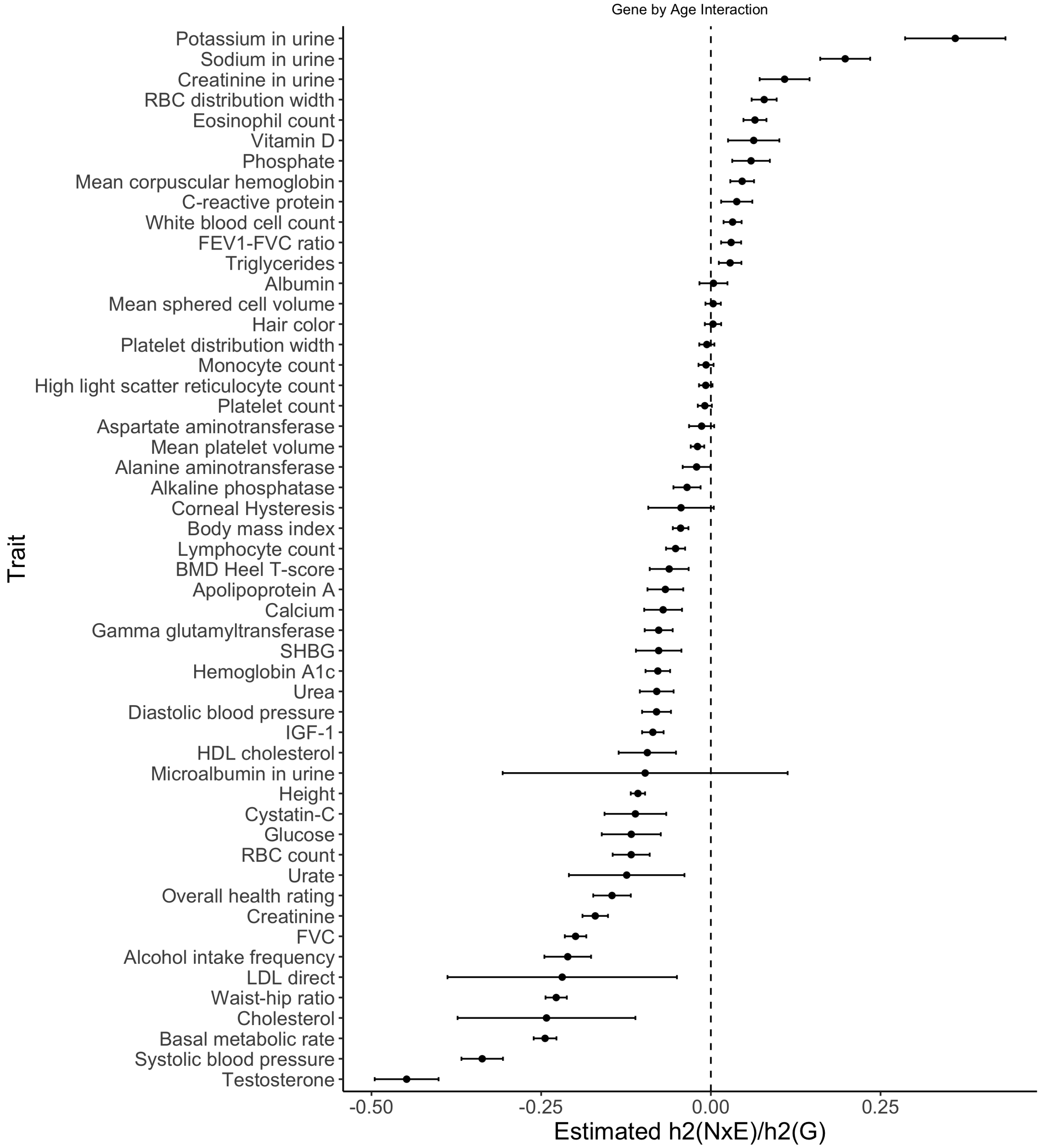


Figure 3.20: **Estimated ratio of variance attributed to noise heterogeneity over additive heritability for Age**. Black error bars mark ±2 standard errors centered on the estimated ratio.

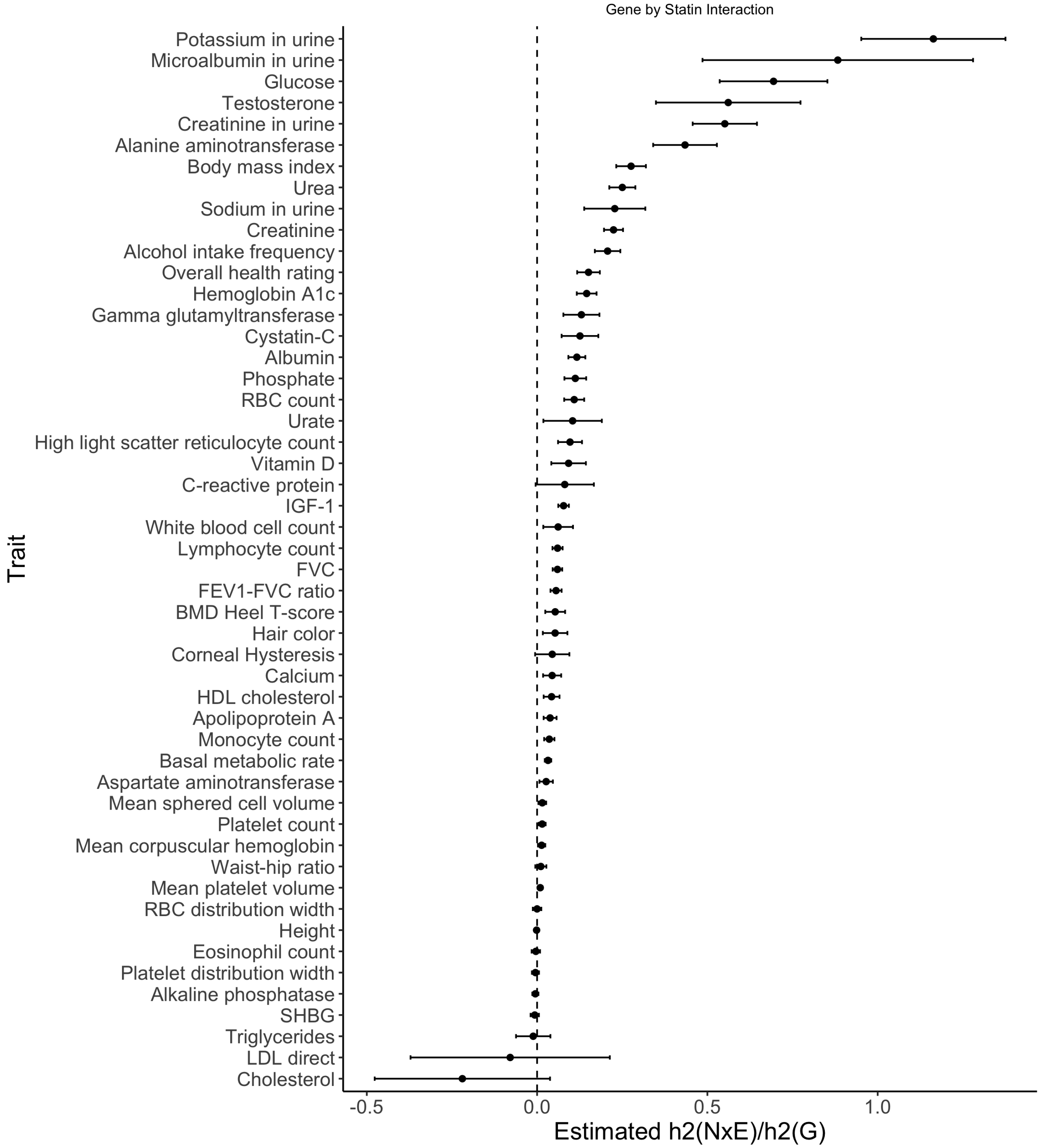


Figure 3.21: **Estimated ratio of variance attributed to noise heterogeneity over additive heritability for Statin usage**. Black error bars mark ±2 standard errors centered on the estimated ratio.

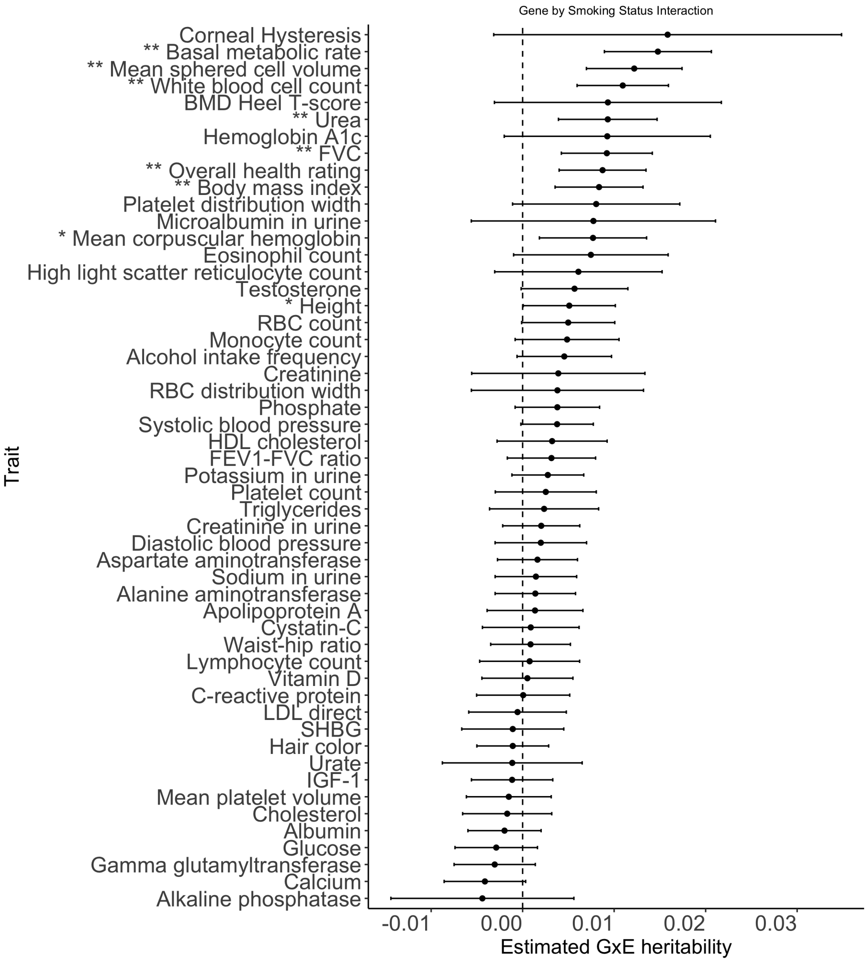


Figure 3.22: **GxSmoking across phenotypes in UK Biobank with the environmental variable coded as binary.** Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. Black error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$ respectively.

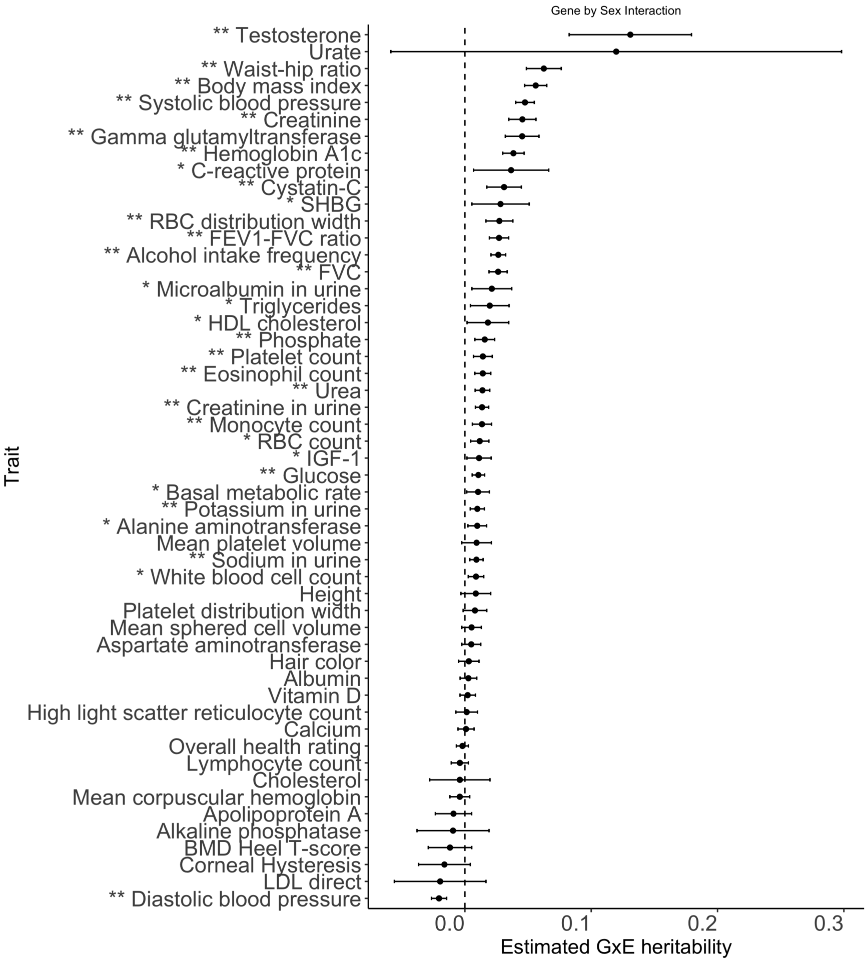


Figure 3.23: **GxSex across phenotypes in UK Biobank with the environmental variable coded as binary.** Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. Black error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/50$ respectively.

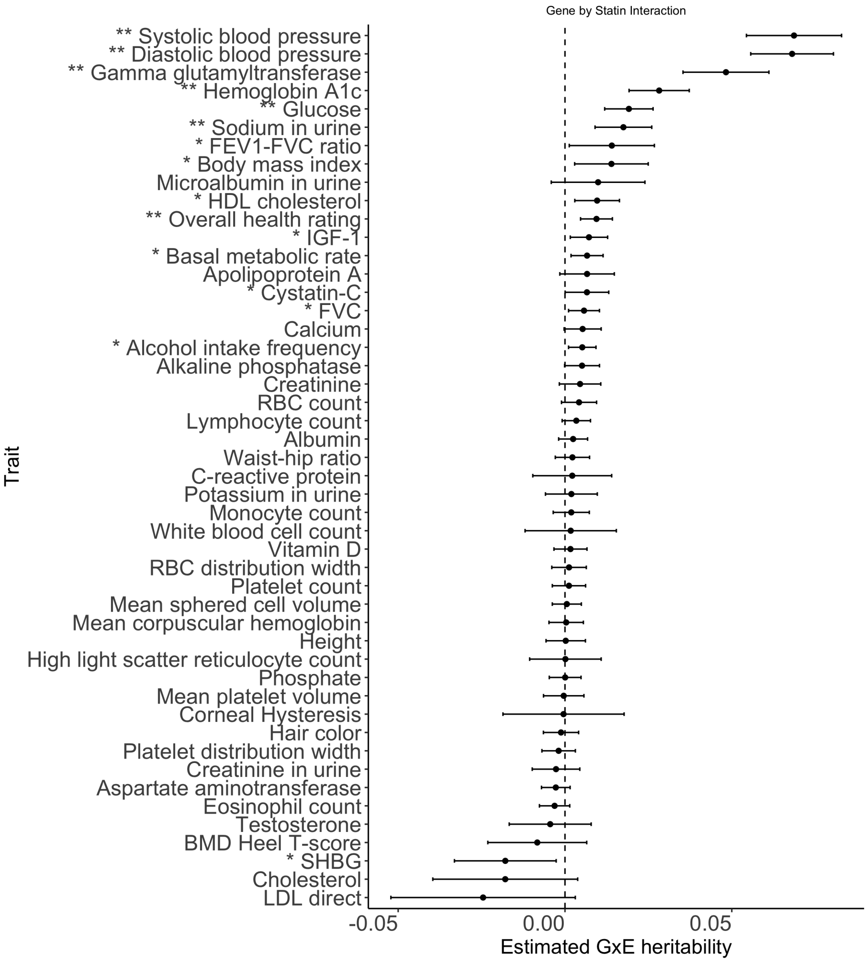


Figure 3.24: **GxStatin across phenotypes in UK Biobank with the environmental variable coded as binary.** Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. Black error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/50$, respectively.

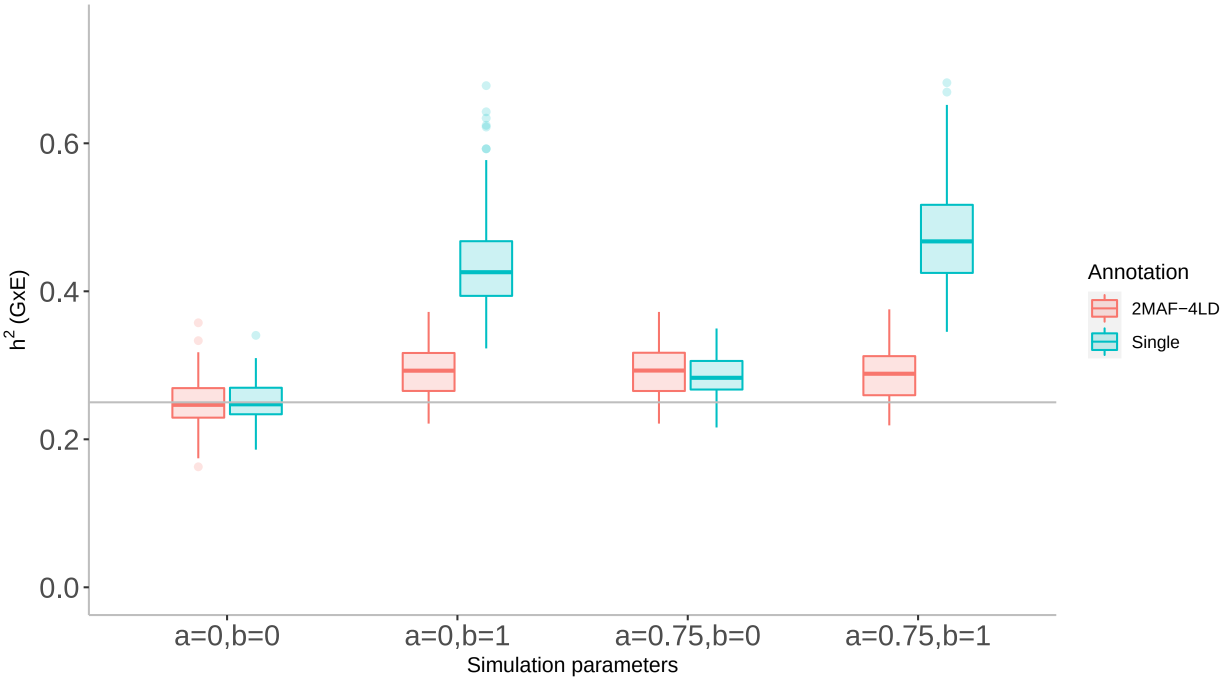


Figure 3.25: **Effect of MAF-LD partitioning on estimated GxE heritability in simulation.** We assessed the effect of MAF-LD partitioning on estimates of h^2_{gxe} in simulations. We ran GENIE in two settings: 1) fitting a model with a single additive and a single GxE variance component, 2) fitting a model with eight additive and eight GxE components defined based on four LD annotations (quartiles of LD scores) and two MAF annotations. we simulated phenotypes with GxE effects and G effects from a subset of $N = 40k$ individuals genotyped at array SNPs $M = 459,792$ by varying the coupling of MAF with effect size (a) and the effect of local LD on effect size (b) (see Supplementary note S2 for details). Here we have $h^2_g = h^2_{gxe} = 0.25$,$h^2_{nxe} = 0.05$, and all SNPs are causal for both additive and GxE effects. Each box plot represents estimates from 100 simulations.

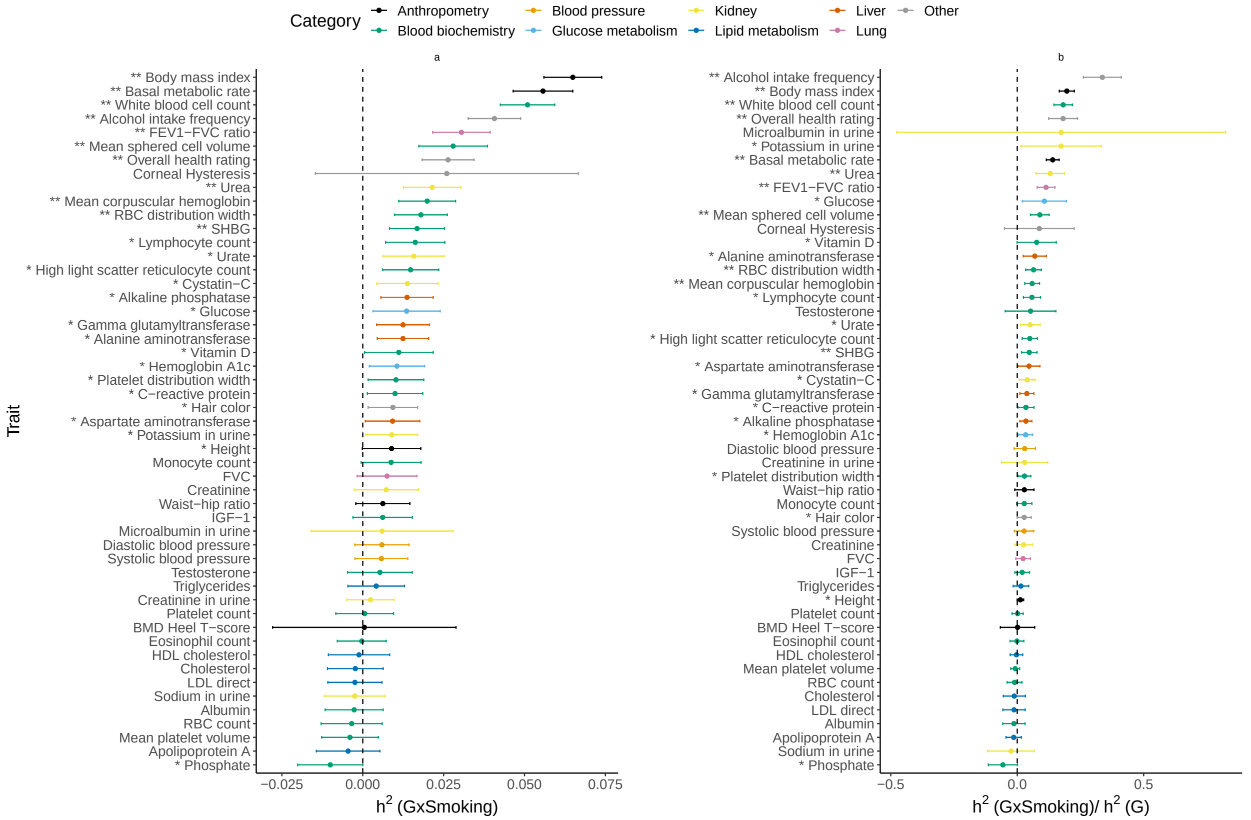


Figure 3.26: **GxSmoking across phenotypes from imputed SNPs in UK Biobank by MAF-LD partitioning.** Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. Black error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

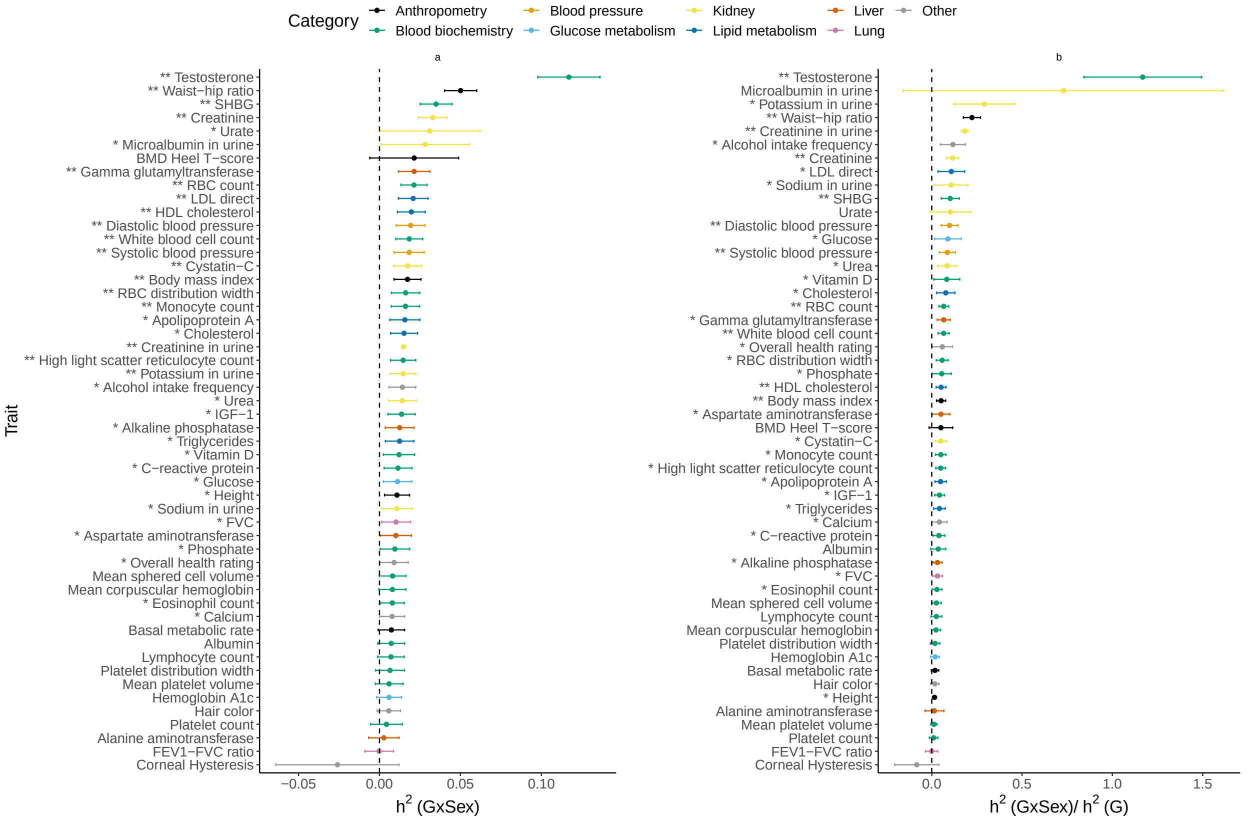


Figure 3.27: **GxSex across phenotypes from imputed SNPs in UK Biobank by MAF-LD partitioning.**Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. Error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

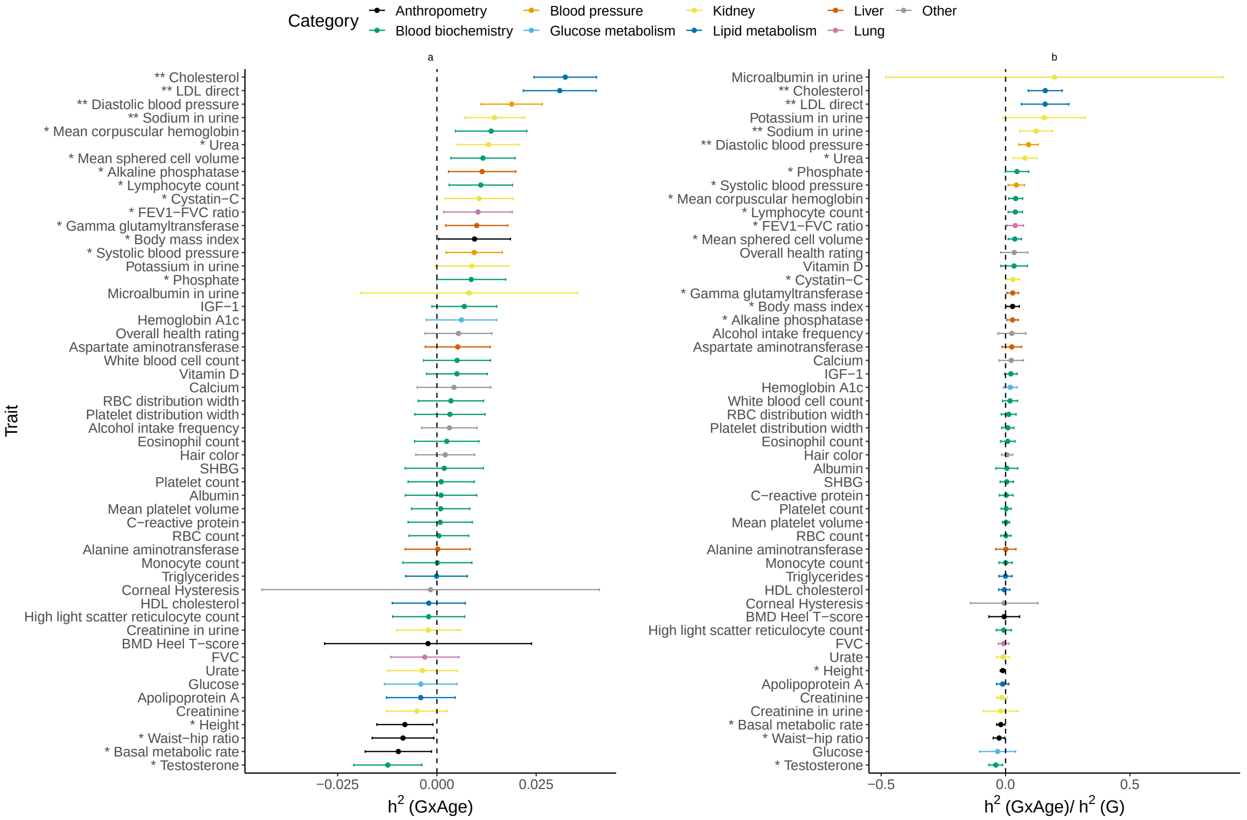


Figure 3.28: **GxAge across phenotypes from imputed SNPs in UK Biobank by MAF-LD partitioning.** Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. Error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

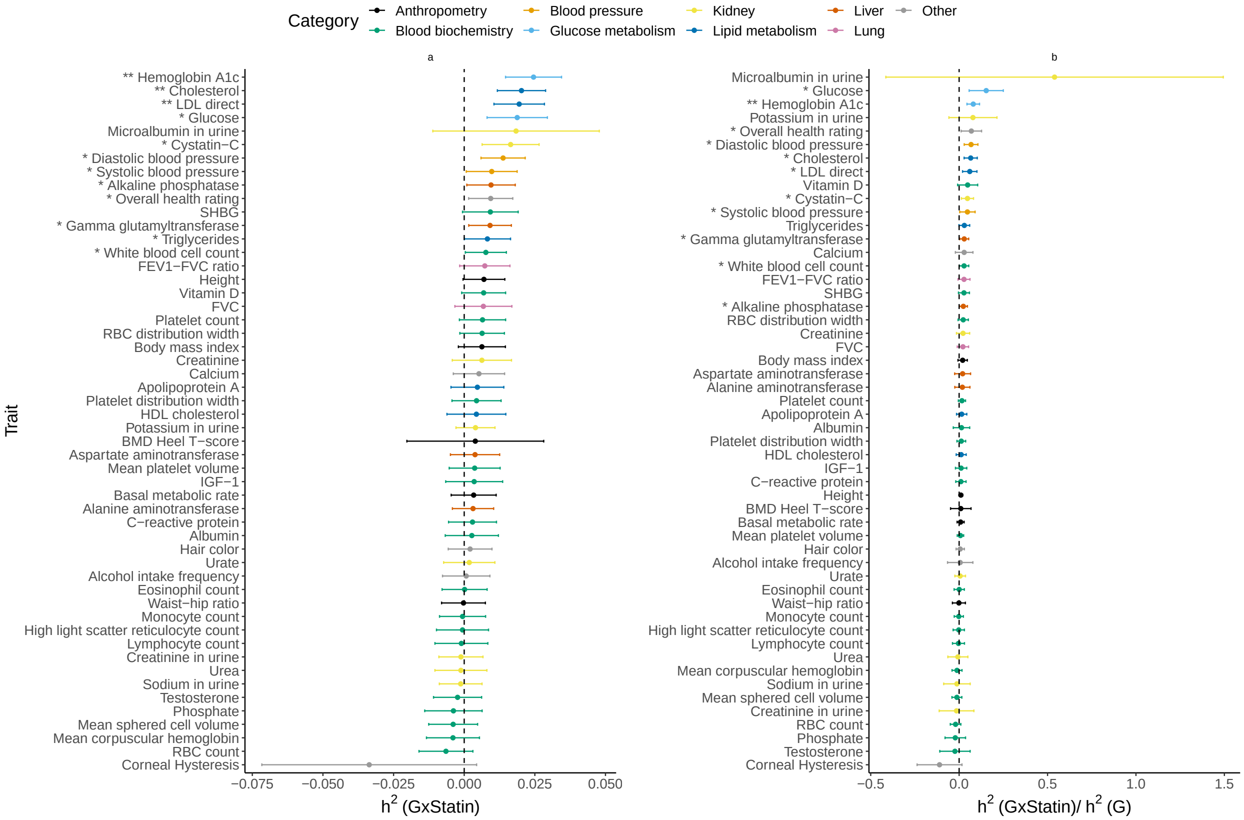


Figure 3.29: **GxStatin across phenotypes from imputed SNPs in UK Biobank by MAF-LD partitioning.** Our model includes the environmental variable as a fixed effect and accounts for environmental heterogeneity. Error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$, respectively.

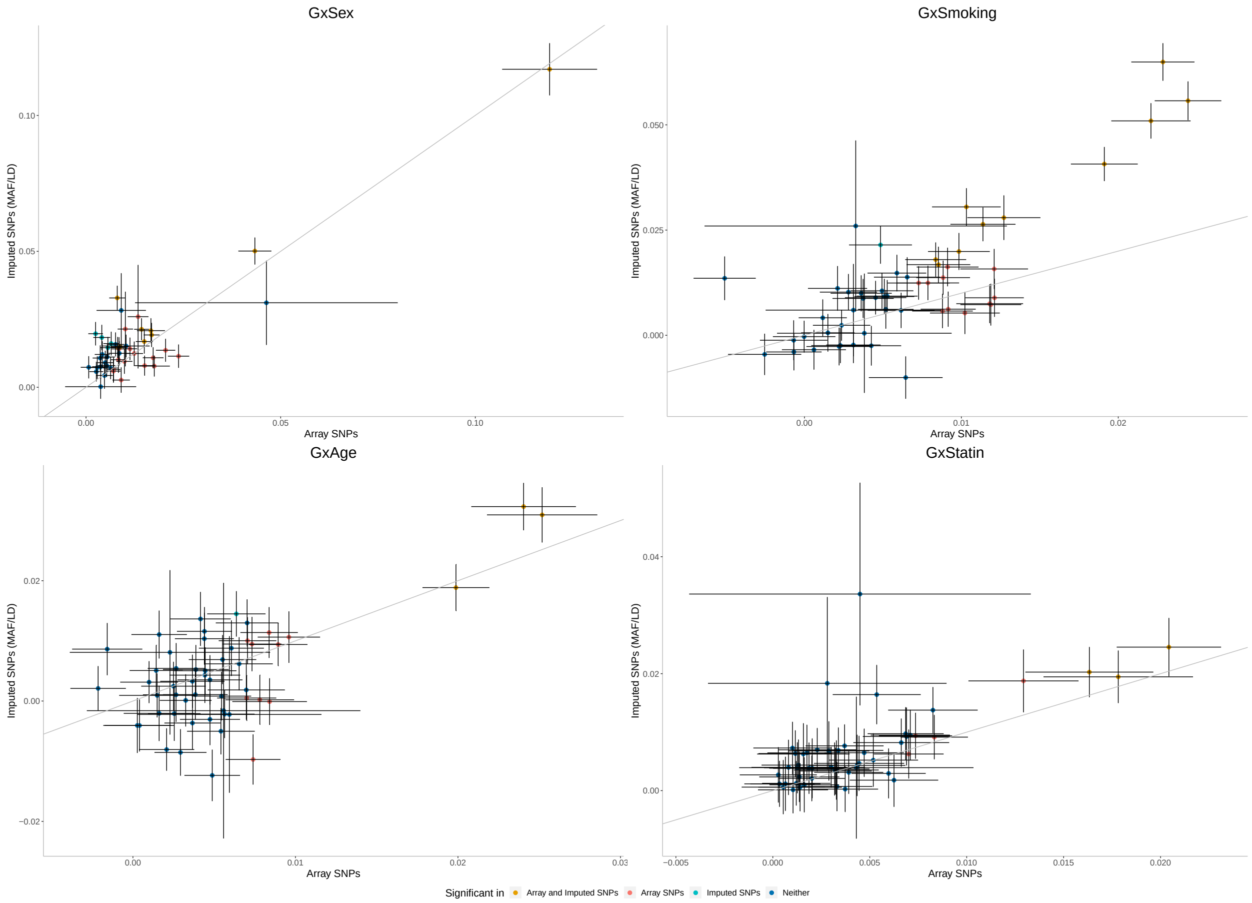


Figure 3.30: **Comparing GxSex, GxSmoking, GxAge, and GxStatin estimates from imputed SNPs (MAF$\geq$ 0.1%) and array SNPs (MAF$\geq$ 1%).** In this analysis, we applied GENIE to imputed SNPs with MAF/LD stratification and array SNPs with a single component. Black error bars mark ±2 standard errors. The asterisk and double asterisk correspond to the nominal $p < 0.05$ and $p < 0.05/200$ respectively. Color of the dots indicate whether estimates of GxE heritability are significant under each model.

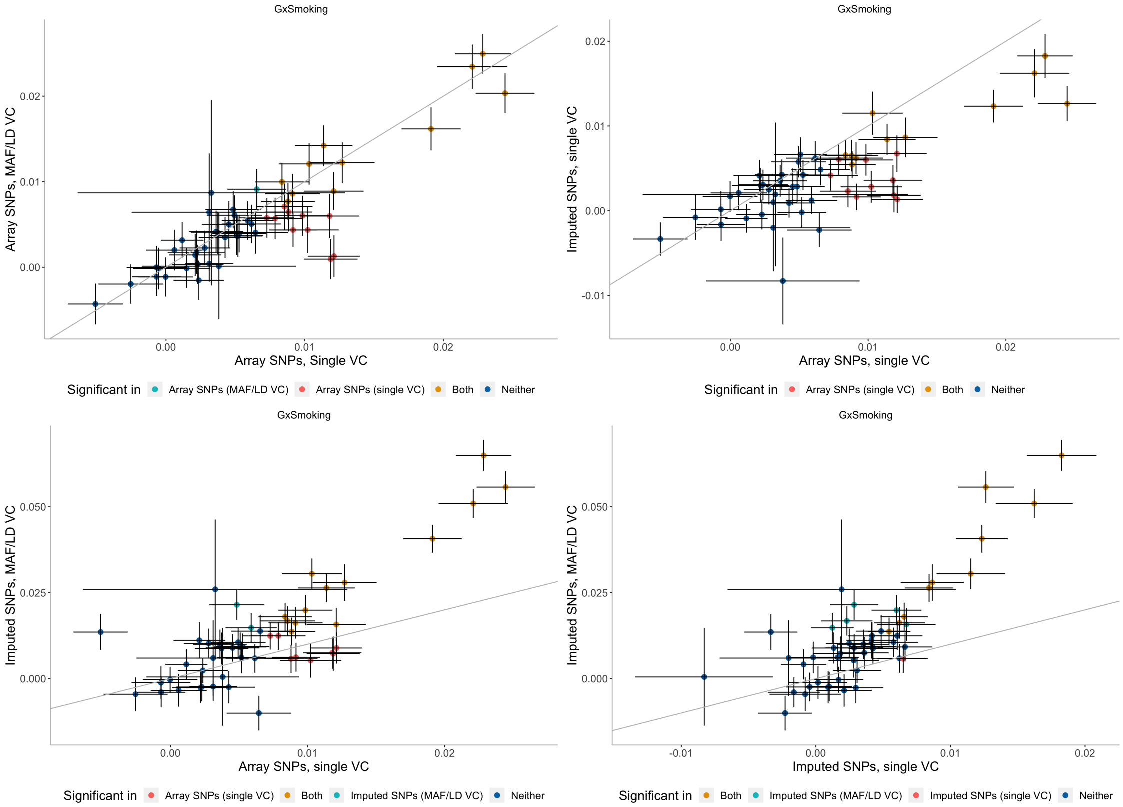


Figure 3.31: **Effect of MAF-LD partitioning on estimated GxE heritability.** We assessed the effect of MAF-LD partitioning on estimates of $h^2_{gxSmoking}$ from array SNPs and imputed SNPs. We ran GENIE in two settings: 1) fitting a model with a single additive and a single GxE variance component, 2) fitting a model with eight additive and eight GxE components defined based on four LD annotations (quartiles of LD scores) and two MAF annotations. Black error bars mark ±2 standard errors centered on the estimates of $h^2_{gxSmoking}$. Color of the dots indicate whether estimates of $h^2_{gxSmoking}$ are significant under each model.

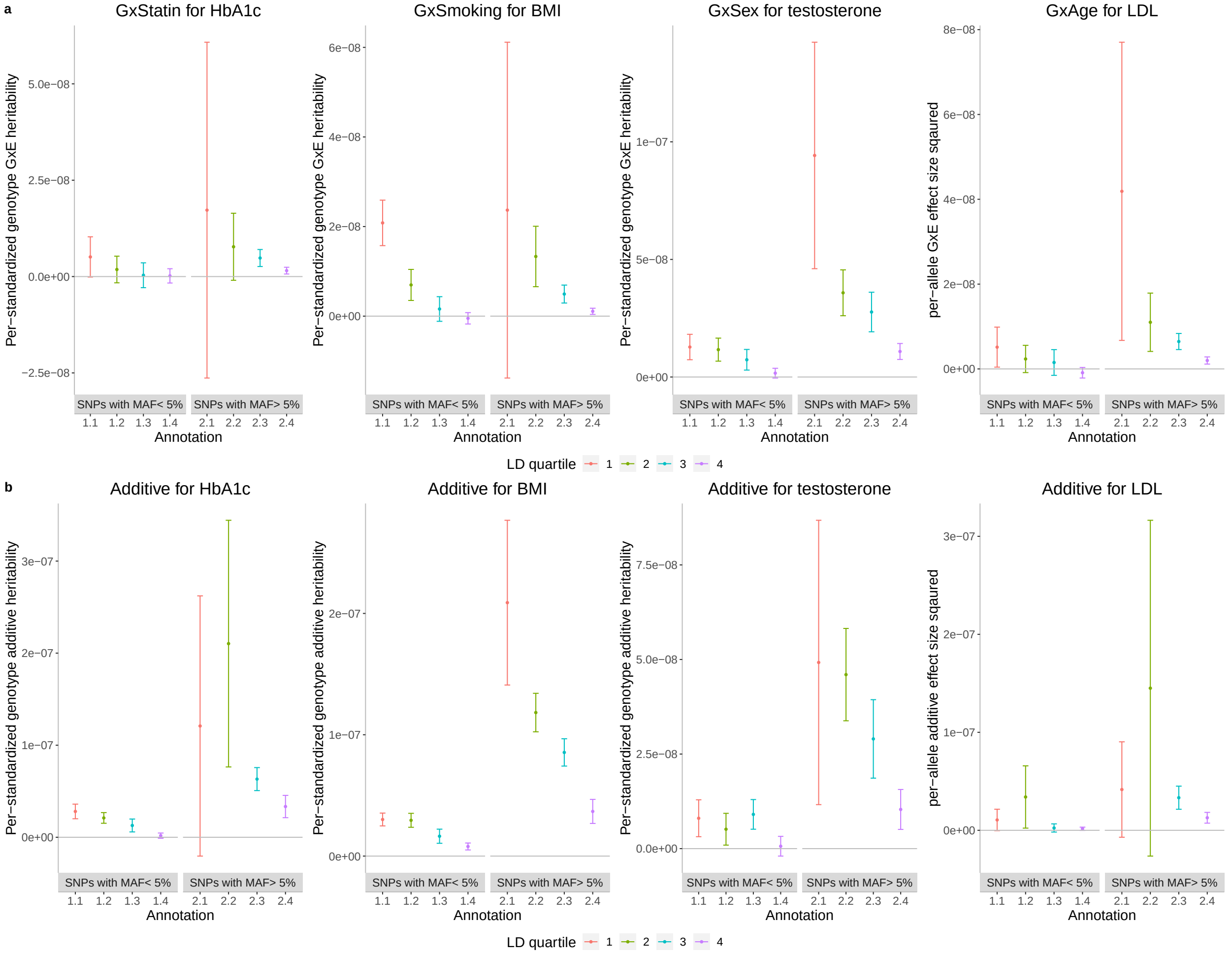


Figure 3.32: **Per-standardized genotype GxE and additive heritability as a function of MAF and LD. a**) The per-standardized genotype GxE heritability for four selected pairs of traits and environments (trait-E pairs). **b**) The per-allele additive heritability for the same trait-E pairs. The x-axis corresponds to MAF-LD annotations where annotation $i.j$ includes SNPs in MAF bin i and LD quartile j where MAF bin 1 and MAF bin 2 correspond to SNPs with MAF $\leq 5\%$ and MAF $> 5\%$ respectively while the first quartile of LD-scores corresponds to SNPs with the lowest LD-scores respectively). The y-axis shows the per-standardized genotype GxE (or additive) heritability defined as $\frac{h_k^2}{2M_k}$ where h_k^2 is the GxE (or additive) heritability attributed to bin k, M_k is the number of SNPs in bin k. Error bars mark ±2 standard errors centered on the estimated effect sizes.

3.8 Supplementary Tables

Method	Variance component	Mean	SE	Bias	Test of bias p-value
GENIE	G	0.2005	0.0382	5e-04	0.9006
GENIE	GxE1	-0.0021	0.027	-0.0021	0.4302
GENIE	GxE2	0.0035	0.0257	0.0035	0.1778
GENIE	GxE3	-1e-04	0.0275	-1e-04	0.9692
GENIE	GxE4	0.0016	0.0245	0.0016	0.5094
GENIE	GxE5	0.0017	0.0267	0.0017	0.5159
GENIE	GxE6	0.0974	0.0278	-0.0026	0.3449
GENIE	GxE7	0.0981	0.0266	-0.0019	0.4673
GENIE	GxE8	0.0975	0.0257	-0.0025	0.3329
GENIE	GxE9	0.0098	0.0263	-2e-04	0.9537
GENIE	GxE10	0.0094	0.0257	-6e-04	0.8082

Table 3.1: **Accuracy of GENIE in the setting of multiple environmental variables**: We reported the bias, and SE of GENIE under different settings with $L = 10$ environmental variables. Bias, mean and SE are computed from 100 replicates. We report p-value of a test of the null hypothesis of no bias in the estimates of variance components.

CHAPTER 4

Dominance effects

4.1 Background

Variation in complex traits can be partitioned into variation due to additive, dominance and epistatic effects [24]. Despite decades of theoretical and experimental efforts, the quantification of non-additive genetic variation in outbred populations such as humans remains challenging [7, 42, 66, 47, 84]. One approach to estimate non-additive sources of heritability in humans have been focused on comparing phenotypic similarity between close relatives [116]. These estimates, however, can be biased by confounding due to shared environmental factors. Further, the limited sample sizes of family and twin studies lead to large standard errors in estimates of non-additive effects. An alternative approach relies on the analysis of unrelated individuals. The relatively small estimates of dominance heritability from prior studies [42, 148] suggest that achieving sufficient power to detect dominance heritability will require the analysis of large numbers of unrelated individuals and methods that can be run on these large sample sizes.

To this end, we extend our previously proposed variance components method [82] to jointly estimate the heritability due to additive and dominance deviation effects attributed to SNPs genotyped across hundreds of thousands of individuals. Additive variance refers to the variance in genotypic value (the conditional mean of phenotype given genotype) explained by regression of the genotypic value on an additive representation of the genotype while dominance variance denotes the residual variance that is not explained by a model with only additive effects. Using this definition, the additive variance component captures the variance attributed to breeding values and includes both additive and dominant genetic effects [**?**, 119]. The additive (dominance) heritability refers to the ratio of the additive (dominance) variance to the phenotypic variance. Further, our method can jointly fit multiple additive and dominance variance components thereby allowing it to provide unbiased estimates of heritability for genetic architectures in which SNP effect sizes vary as a function of minor allele frequency (MAF) and linkage disequilibrium (LD).

Our method obtains unbiased estimates of additive and dominance heritability under a range of MAF and LD-dependent architectures while controlling the false positive rate of rejecting the null hypothesis of no dominance heritability under genetic architectures that assume no dominance. Applying our method to a total of 50 continuous traits measured in $291,273$ unrelated white British individuals in the UK Biobank, we find that additive heritability is 21.86% on average while dominance heritability is 0.13% on average (about 0.48% of the heritability attributed to additive effects) across common array SNPs ($M = 459,792$ SNPs, MAF $> 1\%$). Analyzing common imputed SNPs ($M = 4,824,392$, MAF $> 1\%$), we find that additive heritability is 22.83% on average while dominance heritability is 0.06% on average (about 0.47% of the heritability attributed to additive effects). We find no evidence for traits that have non-zero dominance heritability after correcting for multiple testing ($p < \frac{0.05}{50}$). Based on the power estimates of our method, we estimate that dominance heritability is unlikely to exceed 1% for the traits analyzed.

4.2 Materials and Methods

4.2.1 Additive and dominance variance components model

We are interested in estimating how much extra genetic variance can be explained by dominance variation on top of a model with only additive effects. We aim to fit a variance components model that relates phenotypes $\boldsymbol{y}$ measured across N individuals to their additive values and dominant deviations over M SNPs (while allowing for multiple additive and dominance components).

$$\begin{aligned}
\boldsymbol{y} &= \sum_{i=1}^{K} \boldsymbol{X}_i \boldsymbol{\beta}_i + \sum_{j=1}^{L} \boldsymbol{D_j} \boldsymbol{\alpha}_j + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{D}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N) \\
\boldsymbol{\beta}_i &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_{A,i}^2}{M_i} \boldsymbol{I}_{M_i}), i \in \{1, \dots, K\} \\
\boldsymbol{\alpha}_j &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_{D,j}^2}{M_j} \boldsymbol{I}_{M_j}), j \in \{1, \dots, L\}
\end{aligned}$$

Here $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an arbitrary distribution over a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. SNPs are partitioned into K additive categories and L dominance categories where $\boldsymbol{X}_i$ and $\boldsymbol{D}_j$ are the $N \times M_i$ and $N \times M_j$ matrices consisting of standardized additive and dominance deviation encodings of SNPs belonging to additive category i and dominance category j respectively, σ_e^2 is the residual variance, and $\sigma_{A,i}^2$ and $\sigma_{D,i}^2$ are the variance component of i-th additive and dominance categories respectively.

Our representation captures dominance deviation, which is different from the dominance effect of markers[119]. We encode additive and dominance deviation effects using a representation that leads to uncorrelated variance components [148, 119]. For alleles A and B at a SNP with the frequency of allele B denoted by f_B, the additive and dominance deviation encodings of the genotypes are defined as follows :

$$v_A(AA) = 0, v_A(AB) = 1, v_A(BB) = 2$$
$$v_D(AA) = 0, v_D(AB) = 2f_B, v_D(BB) = (4f_B - 2)$$

The proportion of phenotypic variance explained by additive variation (*additive heritability*) at all SNPs is defined as:

$$h_A^2 = \frac{\sum_{i=1}^{K} \sigma_{A,i}^2}{\sum_{i=1}^{K} \sigma_{A,i}^2 + \sum_{j=1}^{L} \sigma_{D,j}^2 + \sigma_e^2} \tag{4.1}$$

The proportion of phenotypic variance explained by dominance deviation (*dominance heritability*) at all SNPs is defined as:

$$h_D^2 = \frac{\sum_{j=1}^{L} \sigma_{D,j}^2}{\sum_{i=1}^{K} \sigma_{A,i}^2 + \sum_{j=1}^{L} \sigma_{D,j}^2 + \sigma_e^2} \tag{4.2}$$

The proposed model extends previous models by introducing the component corresponding to dominance deviation effects in addition to the additive effects [82]. Further, the proposed model allows for the joint estimation of multiple additive and dominance components, *e.g.*, corresponding to SNPs with varying minor allele frequency (MAF) and linkage disequilibrium (LD) annotations that have been previously shown to lead to relatively unbiased estimates of SNP heritability [18, 82].

The key inference problem in this model is the estimation of the variance components: $(\boldsymbol{\sigma^2_A}, \boldsymbol{\sigma^2_D}, \sigma^2_e)$ where $\boldsymbol{\sigma^2_A} = (\sigma^2_{A,1}, .., \sigma^2_{A,K})$ and $\boldsymbol{\sigma^2_D} = (\sigma^2_{D,1}, .., \sigma^2_{D,L})$. We use a scalable method-of-moments estimator, *i.e.*, finding values of the variance components such that the population moments match the sample moments [37, 30, 117, 33, 145]. Our method uses a randomized algorithm that avoids explicitly computing genetic relatedness matrices. Instead, it operates on a smaller matrix formed by multiplying the input genotype matrix with a small number of random vectors allowing it to scale to large samples. We estimate standard errors (SE) using an efficient block Jackknife over SNPs with 100 blocks. To estimate the variance components, we use a Method-of-Moments (MoM) estimator that estimates parameter values so that the population moments are close to the sample moments [41]. Since $\mathbb{E}\left[\boldsymbol{y}\right] = 0$, we derived the MoM estimates by equating the population covariance to the empirical covariance. The population covariance is given by:

$$cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sum_i \sigma^2_{A,i}\boldsymbol{K}_{A,i} + \sum_j \sigma^2_{D,j}\boldsymbol{K}_{D,j} + \sigma^2_e \boldsymbol{I}_N \tag{4.3}$$

Here $\boldsymbol{K}_{A,k} = \frac{\boldsymbol{X}_k\boldsymbol{X}_k^T}{M_k}$ $(\boldsymbol{K}_{D,k} = \frac{\boldsymbol{D}_k\boldsymbol{D}_k^T}{M_k})$ is the additive (dominance) genetic relatedness matrix (GRM) computed from all SNPs of k-th category. Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to find the variance components.

$$(\tilde{\boldsymbol{\sigma^2_A}}, \tilde{\boldsymbol{\sigma^2_D}}, \tilde{\sigma^2_e}) = argmin_{(\boldsymbol{\sigma^2_A}, \boldsymbol{\sigma^2_D}, \sigma^2_e)} ||\boldsymbol{y}\boldsymbol{y}^T - (\sum_i \sigma^2_{A,i}\boldsymbol{K}_{A,i} + \sum_j \sigma^2_{D,j}\boldsymbol{K}_{D,j} + \sigma^2_e \boldsymbol{I}_N)||^2_F \tag{4.4}$$

For simplicity, we denote $\boldsymbol{K}_i = \boldsymbol{K}_{A,i}$ for $i = 1, .., K$,$\boldsymbol{K}_{K+j} = \boldsymbol{K}_{D,j}$ for $j = 1, .., L$ and

$J = K + L$. The MoM estimator satisfies the following normal equations:

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{b} \\ \boldsymbol{b}^T & N \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\sigma^2}} \\ \tilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{y}^T \boldsymbol{y} \end{bmatrix} \tag{4.5}$$

Here $\tilde{\boldsymbol{\sigma^2}} = \begin{bmatrix} \tilde{\boldsymbol{\sigma_A^2}} \\ \tilde{\boldsymbol{\sigma_D^2}} \end{bmatrix}$, $\boldsymbol{T}$ is a $J \times J$ matrix with entries $T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l), k, l \in \{1, \ldots, J\}$, $\boldsymbol{b}$ is a J-vector with entries $b_k = tr(\boldsymbol{K}_k) = N$ (because $\boldsymbol{X}_k$s and $\boldsymbol{D}_k$s are standardized), and $\boldsymbol{c}$ is a J-vector with entries $c_k = \boldsymbol{y}^T \boldsymbol{K}_k \boldsymbol{y}$. Each GRM $\boldsymbol{K}_k$ can be computed in time $\mathcal{O}(N^2 M_k)$ and $\mathcal{O}(N^2)$ memory. Given J GRMs, the quantities $T_{k,l}$, c_k, $k, l \in \{1, \ldots, J\}$, can be computed in $\mathcal{O}(J^2 N^2)$. Given the quantities $T_{k,l}$, c_k, the normal Equation (4.5) can be solved in $\mathcal{O}(J^3)$. Therefore, the total time complexity for estimating the variance components is $\mathcal{O}(N^2 M + J^2 N^2 + J^3)$

The key bottleneck in solving the normal Equation (4.5) is the computation of $T_{k,l}$, $k, l \in \{1, \ldots, J\}$ which takes $\mathcal{O}(N^2 M)$. Instead of computing the exact value of $T_{k,l}$, we use an unbiased estimator of the trace [48] based on the following identity: for a given $N \times N$ matrix $\boldsymbol{C}$, $\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z}$ is an unbiased estimator of $tr(\boldsymbol{C})$ ($E[\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z}] = tr[\boldsymbol{C}]$) where $\boldsymbol{z}$ be a random vector with mean zero and covariance $\boldsymbol{I}_N$. Hence, we can estimate the values $T_{k,l}$, $k, l \in \{1, \ldots, J\}$ as follows:

$$T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l) \approx \widehat{T_{k,l}} = \frac{1}{B} \frac{1}{M_k M_l} \sum_b \boldsymbol{z}_b^T \boldsymbol{E}_k \boldsymbol{E}_k^T \boldsymbol{E}_l \boldsymbol{E}_l^T \boldsymbol{z}_b \tag{4.6}$$

where $\boldsymbol{E_i}$ matrix can be standardized additive $\boldsymbol{X_i}$ or dominance $\boldsymbol{D_i}$ matrix. Here $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$ are B independent random vectors with zero mean and covariance $\boldsymbol{I}_N$. We draw these random vectors independently from a standard normal distribution. Computing $T_{k,l}$ using the unbiased estimator involves four multiplications of sub-matrices of the genotype matrix with a vector, repeated B times. Therefore, the total running time for estimating the matrix $\boldsymbol{T}$ is $\mathcal{O}(NMB + J^2 NB)$.

4.2.2 Simulations

We simulated phenotypes for UK Biobank genotypes consisting of $M = 459,792$ array SNPs and $N = 291,273$ unrelated white British individuals (see Section on UK Biobank data). We simulated phenotypes from genotypes using the following model:

$$\begin{aligned}
\sigma^2_{A,m} &= Sc_m w_m^b [f_m(1-f_m)]^a \\
(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, .., \boldsymbol{\beta}_m)^T &\sim \mathcal{N}(\mathbf{0}, diag(\sigma^2_{A,1}, \sigma^2_{A,2}, ..., \sigma^2_{A,m})) \\
\sigma^2_{D,m} &= S^{'} c^{'}_m \\
(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, .., \boldsymbol{\alpha}_m)^T &\sim \mathcal{N}(\mathbf{0}, diag(\sigma^2_{D,1}, \sigma^2_{D,2}, ..., \sigma^2_{D,m})) \\
y|\boldsymbol{\beta}, \boldsymbol{\alpha} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\alpha}, (1 - h^2_A - h^2_D)\boldsymbol{I}_N) \qquad (4.7)
\end{aligned}$$

where $h^2_A, h^2_D \in [0,1]$, $a \in \{0, 0.75\}$,$b \in \{0,1\}$. Here S and $S^{'}$ are normalizing constants chosen so that $\sum_{m=1}^{M} \sigma^2_{A,m} = h^2_A$, $\sum_{m=1}^{M} \sigma^2_{D,m} = h^2_D$. Additive and dominance deviation effect sizes are denoted by $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ respectively. f_m and w_m are the minor allele frequency and LDAK score of m^{th} SNP respectively. In this model, $c_m, c^{'}_m \in \{0,1\}$ are indicator variables for the causal status of SNP m .The LD score of a SNP is defined to be the sum of the squared correlation of the SNP with all other SNPs that lie within a specific distance, and the LDAK score of a SNP is computed based on local levels of LD such that the LDAK score tends to be higher for SNPs in regions of low LD [109] . The above models relating genotype to phenotype are commonly used in methods for estimating SNP heritability: the GCTA Model (when $a = b = 0$ in Equation 4.7), which is used by the software GCTA [134], and the LDAK Model (where $a = 0.75, b = 1$ in Equation (4.7)) used by software LDAK [109]. Moreover, under each model, we varied the proportion and minor allele frequency (MAF) of causal variants (CVs). Proportion of causal variants were set to be either 100% or 1%, and MAF of causal variants drawn uniformly from $[0, 0.5]$ or $[0.01, 0.05]$ or $[0.05, 0.5]$ to consider genetic architectures that are either infinitesimal or sparse as well genetic architectures that include a mixture of common and rare SNPs as well as one that includes only common SNPs. We generated 100 sets of simulated phenotypes for each setting of parameters.

In experiments to assess the false positive rate, the additive heritability was set to 0.5

while the dominance heritability was set to 0. We computed p-values of a test of the null hypothesis of no h_D^2 by computing the Z-score of the estimated h_D^2 to its standard error and computing a p-value of the two-tailed test. Let $h_D^2(i)$ be the estimate of h_D^2 while $\hat{SE}_i$ is the jackknife estimate of standard error on the i-the replicate for $i \in \{1, .., 100\}$. We computed the p-value of a test of the null hypothesis of no h_D^2 on the i-th replicate from the Z-score defined as $h_D^2(i)/\hat{SE}_i$ for $i \in \{1, .., 100\}$ To test the bias of the estimator, for every simulation setting, first we compute $mean(\hat{h_D^2})$ and $SE(\hat{h_D^2})$ from all replicates, then we reported p-values of a test of no bias from the Z-score defined as $\frac{mean(\hat{h_D^2})-h_D^2}{SE/10}$.

4.2.3 Power

To assess the power of our method to detect dominance heritability, we considered simulations under different genetic architectures with a non-zero dominance heritability. Across 16 different genetic architectures, we vary the additive and dominance heritabilities and proportion of causal dominance variants (Methods). We simulated 100 replicates for every genetic architecture. Let $h_D^2(i)$ be the estimate of h_D^2 while $\hat{SE}_i$ is the jackknife estimate of standard error on the i-the replicate for $i \in \{1, .., 100\}$. We computed the p-value of a test of the null hypothesis of no h_D^2 on the i-th replicate from the Z-score defined as as $h_D^2(i)/\hat{SE}_i$ for $i \in \{1, .., 100\}$. Finally, we reported the percentage of replicates with p-value$< t$ as the power of our method on a given simulated genetic architecture for a p-value threshold of t.

4.2.4 UK Biobank Data

We applied our method to UK Biobank data. We restricted our study to self-reported British white ancestry individuals which are $> 3^{rd}$ degree relatives that is defined as pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$ [4]. Furthermore, we removed individuals who are outliers for genotype heterozygosity and/or missingness. We removed SNPs with greater than 1% missingness and minor allele frequency $< 1\%$, and that fail the test of Hardy-Weinberg equilibrium at significance threshold 10^{-7}. Finally we obtained a set of $N = 291,273$ individuals and $M = 459,792$ SNPs to use in the real data analyses. We included age, sex, and the top 20 genetic principal components (PCs) as covariates in our analysis for all traits. We used PCs precomputed by the UK Biobank from a superset of $488,295$ individuals. Additional covariates were used for waist-to-hip ratio (adjusted for BMI) and diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy, and oral contraceptives). Further, we also analyzed $M = 4,824,392$ imputed SNPs with MAF $> 1\%$ minor allele frequency (excluding SNPs with missingness $> 1\%$ and SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7}) across $N = 291,273$ unrelated white British individuals.

4.3 Results

4.3.1 Accuracy of estimates of dominance heritability in simulations

Previous studies estimate a relatively small contribution of dominance heritability for complex traits [148] so that we would like to test the false positive rate of a test of the hypothesis of no dominance heritability. To assess the false positive rate of our method, we performed simulations in the absence of dominance deviation effects ($M = 459,792$ SNPs, $N = 291,273$ individuals). Since additive SNP effects tend to vary as a function of MAF and LD patterns at the SNP [18, 29] and SNP heritability estimates tend to be sensitive to these assumptions, we simulated phenotypes according to 16 MAF and LD-dependent architectures by varying the additive heritability, the proportion of variants that have non-zero effects (causal variants or CVs), the distribution of causal variants across minor allele frequencies (CVs distributed across all minor allele frequency bins or CVs restricted to either common or low-frequency bins), and the form of coupling between the SNP effect size and MAF as well as LD. The key parameter in applying our method is the number of random vectors B which we set to 10.The key parameter in applying RHE-mc is the number of random vectors B.We have performed a set of experiments to explore the choice of B. We simulated 100 phenotypes based on $M = 459,792$ array SNPs and $N = 291,273$ individuals where $h^2_A = 0.25$ and $h^2_D = 0.02$, $p_{causal}(A) = 1$ and $p_{causal}(D) = 0.02$. We observe that the Pearson's correlation coefficients (r) between estimates with $B = 10$ and estimates with $B = 100$ are 0.94 (additive heritability) and 0.91 (dominance heritability). Therefore, $B = 10$ is sufficient for the applications considered. To obtain unbiased estimates, we also do not constrain the estimates of the variance components (allowing for negative estimates).

Recent studies have shown that methods that fit a single additive variance component yield biased estimates of SNP heritability due to the LD and MAF dependent architecture of complex traits [18, 29, 44] while models that allow for SNP effects to vary with MAF and LD obtain relatively unbiased estimates [18, 29, 82]. Thus, we ran our method using 24 bins for additive effects (based on 6 MAF and 4 LD bins) and a single bin for dominance deviation effects (although our method allows for fitting multiple dominance bins). Across the range of genetic architectures, we obtained accurate estimates of h_A^2 when we jointly fit additive and dominance heritability: biases range from -2×10^{-3} to 2×10^{-3} where $h_A^2 = 0.5$ (Figure 4.1). We also obtain unbiased estimates of h_D^2 with biases ranging from -5×10^{-5} to 6×10^{-4} where $h_D^2 = 0.0$ (Figure 4.1). Importantly, the false positive rate of rejecting the null hypothesis of no dominance heritability across 16 diverse genetic architecture is controlled at level 0.05 (see Table 4.1). We performed additional simulations that demonstrate accurate heritability estimates for a smaller sample size of $N = 10,000$ individuals.

Next, we considered simulations under genetic architectures with a non-zero dominance heritability. We evaluated the accuracy of additive and dominance heritability estimates across 16 different genetic architecture where we vary the additive and dominance heritabilities and proportion of causal dominance variants. We ran our method using 24 bins for additive effects (based on 6 MAF and 4 LD bins) and a single bin for dominance deviation effects. We obtained accurate estimates of h_A^2 and h_D^2 when we jointly fit additive and dominance heritability : biases range from -1.6×10^{-3} to 2.7×10^{-4} where $h_D^2 \in \{0.05, 0.02, 0.01, 0.001\}$ for dominance heritability while the biases range from -2.3×10^{-3} to 1.4×10^{-4} where $h_A^2 \in \{0.5, 0.25\}$ for additive heritability (Figure 4.2).

In addition, we observe high power ($> 95\%$ for a p-value threshold of 0.05) to detect dominance heritability as low as 1% in a sample size of $\approx 300,000$ (Table 4.2). A more realistic assessment of power would consider the multiple testing burden incurred when testing a collection of phenotypes with the goal of discovering traits with significant dominance heritability. Assuming we test fifty phenotypes (matching our analyses of the UK Biobank), we estimate 100% power to detect $h_D^2 = 2\%$ and $> 50\%$ power to detect $h_D^2 = 1\%$ in a sample of $\approx 300,000$ individuals ($p < \frac{0.05}{50}$).

We performed simulations to compare the accuracies of RHE-mc to REML and HE regression implemented in the GCTA software. For computational reasons, we simulate phenotypes from a subsampled set of $10,000$ genotypes across $M = 459,792$ array SNPs from the UK Biobank data. We simulated 100 phenotypes where $h_A^2 = 0.25$ and $h_D^2 = 0.05$, $p_{causal}(A) = 1$ and $p_{causal}(D) = 0.05$. All three methods obtain unbiased estimates of additive and dominance heritability. The standard error of RHE-mc is 3% and 12% larger than REML(GCTA) for additive and dominance heritability respectively. The standard error of RHE-mc is same as HE(GCTA) for additive heritability and 3% less than HE(GCTA) for dominance heritability.

Further, we evaluated the accuracy of the jackknife estimate of standard error in simulations ($N = 291,273$ unrelated individuals, $M = 459,792$ array SNPs) across diverse genetic architectures. We observe that the jackknife SE yields estimates with relative bias -1.7% on average over 13 genetic architectures.

Finally, we performed experiments to measure the extent to which we are able to capture additive and dominance variation of causal SNPs when only a subset of causal SNPs are observed due to imperfect tagging. In the first set of experiments, we simulated phenotypes based on array SNPs ($N = 291,273$ unrelated individuals, $M = 459,792$ array SNPs) where $h_A^2 = 0.25$ and $h_D^2 = 0.02$, the proportion of causal variants in the additive component is varied between 1% and 100% while the proportion of causal variants in the dominance variance component is set to 1%. We ran RHE-mc on genotypes with varying proportions of observed causal SNPs, $p_{observed} \in \{0\%, 50\%, 75\%, 100\%\}$. While estimates of additive heritability remain relatively unbiased, estimates of dominance heritability are biased downwards with the magnitude of the bias being proportional to the percentage of observed causal SNPs. These experiments suggest that dominance heritability is more sensitive to imperfect tagging than additive heritability (although this sensitivity might also be partly explained by the smaller magnitudes of the dominance heritability in our simulations). To further explore this issue, we repeated this experiment using $M = 4,824,392$ imputed genotypes with MAF $> 1\%$ with the same genetic architecture used in the analysis of array SNPs. We observe that both additive and dominance heritability estimates are relatively unbiased even when the percentage of observed causal SNPs is as low as 0%. These observations likely reflect the better tagging of SNPs that encode additive and dominance genotypes in the imputed data.

4.3.2 Additive and dominance effects in the UK Biobank

We applied our method to estimate additive and dominance heritability for 50 quantitative traits in the UK Biobank [4] by partitioning the additive component into 8 bins (based on two MAF bins (MAF≤ 0.05, MAF> 0.05) and quartiles of the LD-scores) and a single dominance bin. We restricted our analysis to $N = 291,273$ unrelated white British individual and $M = 459,792$ SNPs (MAF$> 1\%$) that were present in the UK Biobank Axiom array. Further, we chose a subset of 50 traits out of a total of 57 traits that have evidence for non-zero additive heritability (Z-score > 3).

Across the 50 traits, we observe that the average additive heritability (h_A^2) is 21.86% (standard deviation of 9.21% across traits) (Figure 4.3). On the other hand, we estimate average dominance heritability (h_D^2) to be 0.13% (SD = 0.39%). On average, we observe that dominance heritability is about 0.48% of additive heritability. We find no evidence for traits that have statistically significant non-zero dominance heritability after correcting for multiple testing ($p < \frac{0.05}{50}$).

Applying our method with a single additive component (no MAF/LD partitioning), we obtain an average $h_A^2 = 27.72\%$ (SD=12.14%) and average $h_D^2 = 0.17\%$ (SD=0.42%) across 50 traits with no evidence for statistically significant non-zero h_D^2(Table S4).

To assess the effect of population stratification on our results, we repeated our analyses retaining the first 10 PCs and 40 PCs. While our original results with first 20 PCs suggested that average $h_D^2 = 0.13\%$ (SD=0.39%), we observe average $h_D^2 = 0.13\%$ (SD= 0.38%) with the first 10 PCs while average $h_D^2 = 0.11\%$ (SD= 0.42%) with the first 40 PCs. Across these analyses, none of the traits show evidence for non-zero h_D^2 estimates that are statistically significant.

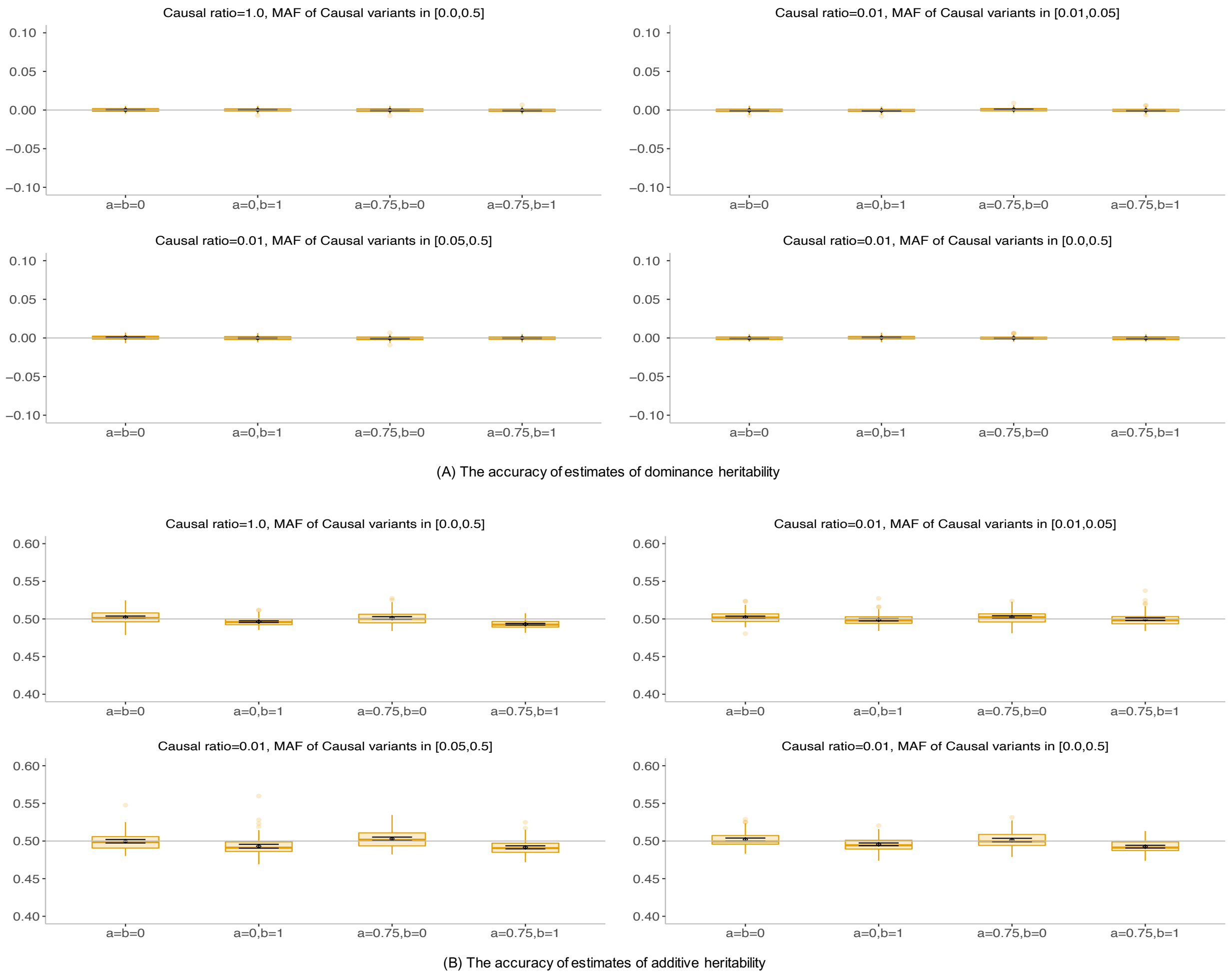
To explore the impact of imperfect tagging of causal variants on our results, we analyzed $M = 4,824,392$ imputed genotypes with MAF $> 1\%$. We observed average $h_A^2 = 22.83\%$ (SD=9.49%) across the 50 traits (Figure 4.4; Pearson's correlation between the point estimates of h_A^2 across array and imputed genotypes is 0.998) with no statistically significant differences between the h_A^2 estimates ($p < \frac{0.05}{50}$). On the imputed genotypes, we estimated average h_D^2 to be 0.06% (SD=0.19%) with the dominance heritability being about 0.47% of additive heritability. We also did not observe any statistically significant differences between the h_D^2 estimates across array and imputed genotypes suggesting that imperfect tagging of common causal SNPs (MAF $> 1\%$) is unlikely to explain our results. Although we did not find evidence for statistically significant non-zero h_D^2 after correcting for multiple testing, we found suggestive evidence for non-zero dominance heritability for blood biochemistry traits: aspartate, basal metabolic rate, blood reticulocyte count , glucose, and calcium ($p < 0.05$).

4.4 Discussion

The contribution of non-additive genetic effects to complex trait variation has been intensely debated [7, 42, 11, 148, 47]. Here, we have extended our previously developed variance components method [82] to jointly estimate multiple additive and dominance variance components on biobank-scale genotype-trait data. We find that our method accurately estimates additive and dominance heritability across a range of MAF and LD-dependent genetic architectures. While tests for the existence of a dominance component have well-controlled false positive rates, our method has high power to detect dominance components with $h_D^2 >= 1\%$ in a sample of $\approx 300K$ unrelated individuals. In application to 50 quantitative traits in the UK Biobank with genotypes measured across $459,792$ array SNPs (MAF $> 1\%$) as well as genotypes measured across $4,824,392$ imputed SNPs (MAF$> 1\%$), we observe substantial additive heritability (21.86% on average for array SNPs, 22.83% on average for imputed SNPs). On the other hand, estimates of dominance heritability tend to be low (0.13% for array and 0.06% for imputed SNPs) so that we do not find any trait with statistically significant evidence of dominance heritability.

While a previous study [148] estimated a 3% dominance heritability (point estimate averaged across 79 traits), we estimate a dominance heritability of 0.13% (point estimate averaged across 50 traits). The differences in the point estimates could be due to the differences in the set of phenotypes and individuals analyzed as well as methodology used. However, our results are concordant with Zhu et al. in that we find no statistically significant estimate of dominance heritability across the traits analyzed. Further, Zhu et al. analyzed a 7,000 individuals which leads to larger SEs than our results based on 300K individuals. The authors of Zhu et al. note that the power to estimate a dominance heritability of 0.05 with a sample size of 7,000 is only about 12%. On the other hand, our power calculations indicate that it is unlikely that h_D^2 is larger than 1% at the traits analyzed. Taken together, our results suggest that systematic identification of dominance heritability will require analysis of even larger sample sizes than the $\approx 300K$ individuals that we analyzed here. While the growth of Biobank-scale datasets will facilitate such estimates, such analyses will also require the development of novel methods that can analyze data at scale.

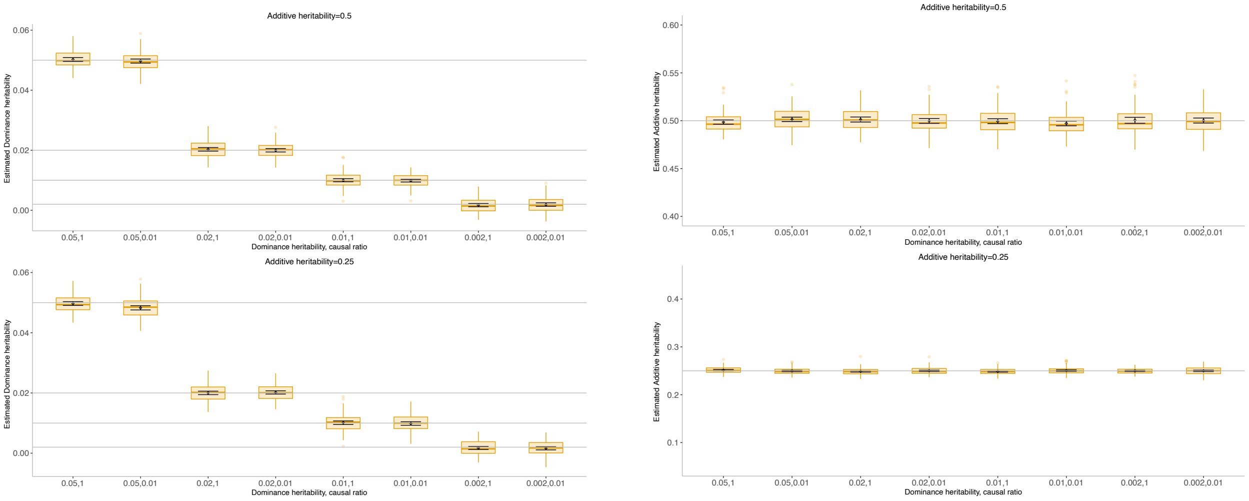
We discuss several limitations of our study as well as directions for future work. The analysis of dominance variance that we have undertaken relies on a specific encoding of dominance and additive effects that leads to uncorrelated components [148]. Due to the choice of this representation, the additive variance component that we estimate includes a contribution from dominant genetic effects while the dominance variance component quantifies the extra genetic variance that can be explained by dominance deviation on top of the additive only model. Alternative encodings might be associated with different statistical and biological interpretation [47]. Second, while our analysis has focused primarily on common SNPs (MAF > 1%), previous work has shown that dominance deviation effects tend to decay faster due to imperfect tagging relative to additive effects leading to a larger bias in estimates of these effects [148]. The concordance of our results across array and imputed genotypes suggests that our estimates of dominance heritability attributed to common SNPs are likely to be robust although we would still underestimate the contribution from low-frequency SNPs. The scalability of our method allows for the exploration of alternative encodings and low-frequency variants at scale. Finally, while our current work focuses on quantitative traits, methods that have previously proposed to estimate heritability in case-control studies [33, 125] can be extended to estimate dominance heritability for binary traits.



(A) The accuracy of estimates of dominance heritability

(B) The accuracy of estimates of additive heritability

Figure 4.1: **The accuracy of estimates of dominance and additive heritabilities in simulations with no dominance heritability ($N = 291,273$ unrelated individuals, $M = 459,792$ array SNPs)**. In **A** and **B**: We plot estimates from our method in the absence of dominance deviation effects under 16 different genetic architectures. We varied the MAF range of causal variants (MAF of CV), the coupling of MAF with effect size (a), and the effect of local LD on effect size ($b = 0$ indicates no LDAK weights and $b = 1$ indicates LDAK weights. We ran 100 replicates where the true additive and dominance heritabilities of the phenotype are 0.5 and 0.0 respectively. We ran our method using a single dominance bin and 24 additive bins formed by the combination of 6 bins based on MAF as well as 4 bins based on quartiles of the LDAK score of a SNP. Black points and error bars represent the mean and ±2 SE. Each box plot represents estimates from 100 simulations. Box plot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles respectively.



(A) The accuracy of estimates of dominance heritability

(B) The accuracy of estimates of additive heritability

Figure 4.2: **The accuracy of estimates of dominance and additive heritabilities in simulations with non-zero dominance heritability (**$N = 291,273$ **unrelated individuals,** $M = 459,792$ **array SNPs)**. In **A**, **B**: We plot estimates from our method under 16 different genetic architectures. We varied the additive heritability h_A^2, dominance heritability h_D^2 , and the proportion of dominance causal variants (causal ratio). Black points and error bars represent the mean and ±2 SE. Each boxplot represents estimates from 100 simulations.Box plot whiskers extend to the minimum and maximum estimates located within 1.5× interquartile range (IQR) from the first and third quartiles respectively.

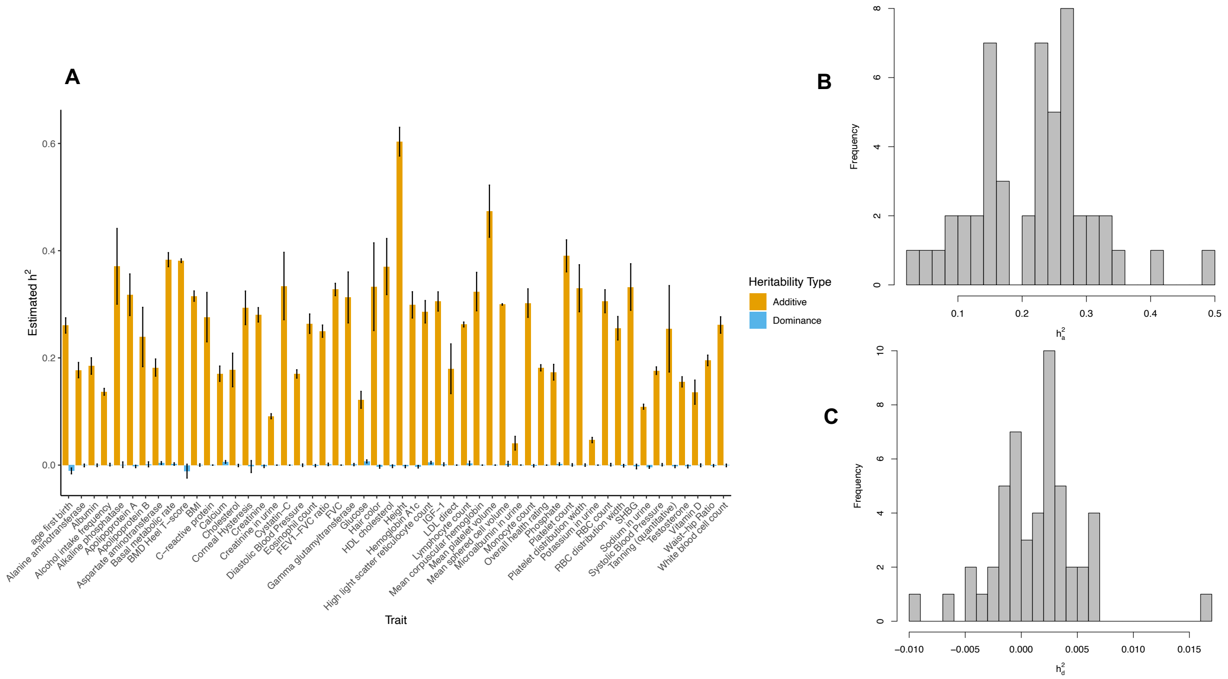


Figure 4.3: **Estimates of additive and dominance heritability for 50 quantitative phenotypes in the UK Biobank ($N = 291,273$ unrelated white British individuals, $M = 459,792$ common array SNPs (MAF $> 1\%$)).** We ran our method partitioning the additive component into 8 bins defined based on two MAF bins (MAF≤ 0.05, MAF> 0.05) and quartiles of the LD-scores and a single dominance bin. We summarize the estimates of additive and dominance heritability across the 50 phenotypes. In **A** : Black error bars mark ±2 standard errors centered on the estimated heritability. In **B** and **C** we plot the histogram of h^2_A and h^2_D respectively. Point estimates and SE's are reported in Table S2.

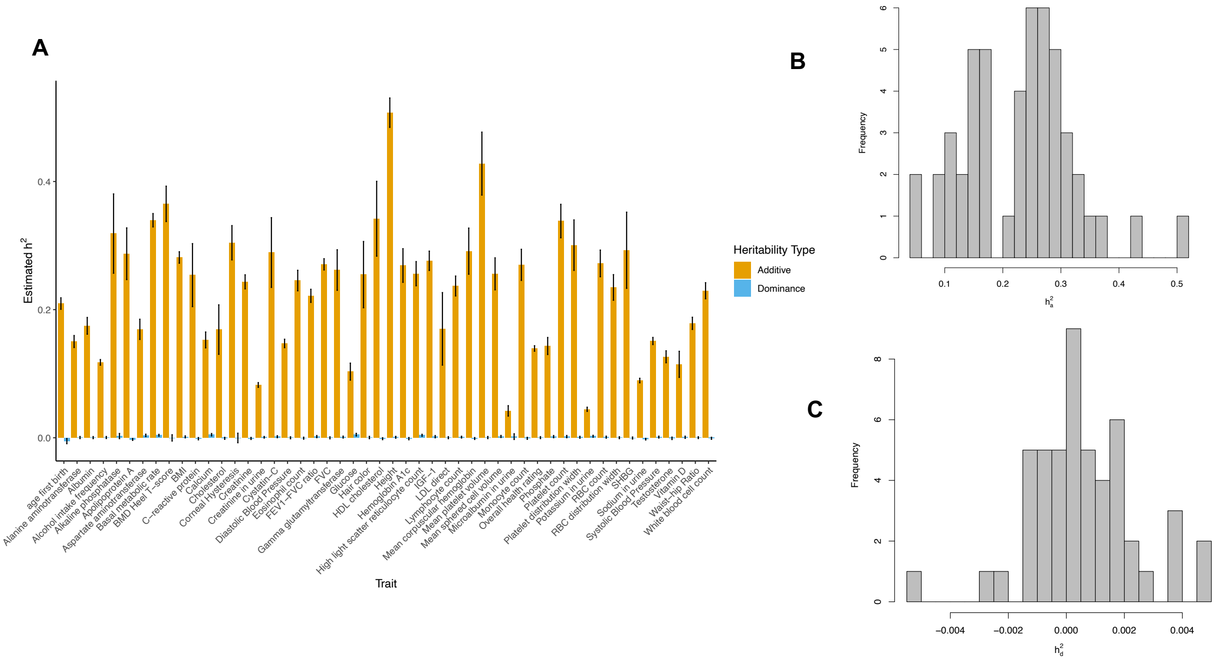


Figure 4.4: **Estimates of additive and dominance heritability for 50 quantitative phenotypes in the UK Biobank** ($N = 291,273$ **unrelated white British individuals,** $M = 4,824,392$ **common imputed SNPs (MAF** $> 1\%$**)).** We ran our method partitioning the additive component into 8 bins defined based on two MAF bins (MAF≤ 0.05, MAF> 0.05) and quartiles of the LD-scores and a single dominance bin. We summarize the estimates of additive and dominance heritability across the 50 phenotypes. In **A** : Black error bars mark ±2 standard errors centered on the estimated heritability. In **B** and **C** we plot the histogram of h^2_A and h^2_D respectively. Point estimates and SE's are reported in Table S3.

Genetic architecture			P(rejection at $p < t$)		Test of bias
% of causal SNPs	MAF of causal SNPs	MAF/LD coupling	$t = 0.05$	$t = 10^{-3}$	p-value
0.01	[0.01,0.05]	a=b=0	6%	0%	0.192
0.01	[0.01,0.05]	a=0,b=1	5%	0%	0.006
0.01	[0.01,0.05]	a=0.75,b=0	6%	1%	0.011
0.01	[0.01,0.05]	a=0.75,b=1	8%	0%	0.187
0.01	[0.0,0.5]	a=b=0	4%	0%	0.388
0.01	[0.0,0.5]	a=0,b=1	8%	0%	0.415
0.01	[0.0,0.5]	a=0.75,b=0	4%	0%	0.593
0.01	[0.0,0.5]	a=0.75,b=1	2%	0%	0.367
0.01	[0.05,0.5]	a=b=0	7%	0%	0.046
0.01	[0.05,0.5]	a=0,b=1	4%	0%	0.813
0.01	[0.05,0.5]	a=0.75,b=0	6%	1%	0.105
0.01	[0.05,0.5]	a=0.75,b=1	1%	0%	0.855
1.0	[0.0,0.5]	a=b=0	2%	0%	0.196
1.0	[0.0,0.5]	a=0,b=1	5%	0%	0.298
1.0	[0.0,0.5]	a=0.75,b=0	7%	0%	0.522
1.0	[0.0,0.5]	a=0.75,b=1	2%	0%	0.130

Table 4.1: **Calibration of tests of dominance heritability**. We assess the false positive rate of tests of dominance heritability based on our method in the absence of dominance deviation effects under 16 different genetic architectures. We varied the MAF range of causal variants (MAF of CV), the coupling of MAF with effect size (a), and the effect of local LD on effect size ($b = 0$ indicates no LDAK weights and $b = 1$ indicates LDAK weights. Probability of rejection is computed from 100 replicates. We report p-value of a test of the null hypothesis of no bias in the estimates of h^2_D.

Genetic architecture		Power		$\hat{h}^2_D$		Test of bias
Additive	Dominance	$t = 0.05$	$t = 10^{-3}$	Mean	SE	p-value
$h^2_A = 0.5$	$p_{causal}(D) = 1$, $h^2_D = 0.05$	100%	100%	0.05	0.003	0.432
$h^2_A = 0.5$	$p_{causal}(D) = 0.01$, $h^2_D = 0.05$	100%	100%	0.049	0.003	0.596
$h^2_A = 0.5$	$p_{causal}(D) = 1$, $h^2_D = 0.02$	100%	100%	0.02	0.002	0.351
$h^2_A = 0.5$	$p_{causal}(D) = 0.01$, $h^2_D = 0.02$	100%	100%	0.02	0.002	0.869
$h^2_A = 0.5$	$p_{causal}(D) = 1$, $h^2_D = 0.01$	97%	68%	0.01	0.002	0.901
$h^2_A = 0.5$	$p_{causal}(D) = 0.01$, $h^2_D = 0.01$	98%	67%	0.0099	0.002	0.730
$h^2_A = 0.5$	$p_{causal}(D) = 1$, $h^2_D = 0.002$	11%	2%	0.0018	0.0025	0.738
$h^2_A = 0.5$	$p_{causal}(D) = 0.01$, $h^2_D = 0.002$	10%	1%	0.0019	0.0027	0.590
$h^2_A = 0.25$	$p_{causal}(D) = 1$, $h^2_D = 0.05$	100%	100%	0.049	0.003	0.434
$h^2_A = 0.25$	$p_{causal}(D) = 0.01$, $h^2_D = 0.05$	100%	100%	0.048	0.003	2.5e-06
$h^2_A = 0.25$	$p_{causal}(D) = 1$, $h^2_D = 0.02$	100%	100%	0.02	0.002	0.889
$h^2_A = 0.25$	$p_{causal}(D) = 0.01$, $h^2_D = 0.02$	100%	100%	0.02	0.002	0.476
$h^2_A = 0.25$	$p_{causal}(D) = 1$, $h^2_D = 0.01$	93%	73%	0.01	0.002	0.744
$h^2_A = 0.25$	$p_{causal}(D) = 0.01$, $h^2_D = 0.01$	93%	66%	0.0098	0.002	0.632
$h^2_A = 0.25$	$p_{causal}(D) = 1$, $h^2_D = 0.002$	9%	0%	0.0017	0.0024	0.373
$h^2_A = 0.25$	$p_{causal}(D) = 0.01$, $h^2_D = 0.002$	12%	1%	0.0017	0.0026	0.292

Table 4.2: **Accuracy and power to detect dominance heritability in simulations ($N = 291,273$ unrelated individuals, $M = 459,792$ array SNPs)**. We assess power, bias, and SE of our method in the presence of dominance and additive heritability under 16 different genetic architectures. Power, mean and SE are computed from 100 replicates. We report p-value of a test of the null hypothesis of no bias in the estimates of h^2_D . Here, $p_{causal}(A) = 1$ and $p_{causal}(D)$ denote the proportion of additive and dominance causal variants respectively. h^2_A and h^2_D denotes total additive and dominance heritabilities. For both components, we assumed GCTA model which is defined as setting $a = b = 0$ in Equation 4.7. Power is reported for p-value threshold of $t \in \{0.05, 0.001\}$.

CHAPTER 5

Epistasis effects

5.1 Background

Genome-wide association studies (GWAS), the dominant approach to identify genetic variants that modulate a trait, have successfully identified hundreds of thousands of associations across thousands of traits primarily by testing a linear additive model. Approaches that attempt to quantify the aggregate effects of genetic variants across the genome (using variance components analysis) have shown that additive effects explain a substantial proportion of trait variation for many complex traits, *i.e.*, the narrow-sense heritability of complex traits is substantial. Interactions in the effect of genes or genetic variants on a trait (*epistasis*) [8] have been hypothesized to play an important role in human complex trait variation and disease risk [84, 123].

Understanding the nature and contribution of epistasis is important for elucidating the genetic architecture of complex traits and disease etiology and improving the accuracy of genetic prediction. Epistasis is one of the factors that could explain the gap between the accuracy of linear models for predicting traits from genetic variants and the expected accuracy based on family-based studies (termed the *missing heritability* problem). Recent studies analyzing the estimates of genetic effects across ancestral populations [78] and the transferability of genetic predictors both within [75] and across ancestries [68, 69] suggest that genetic interactions could explain why genetic effects differ across ancestral populations and that lack of transferability of genetic predictors both within and across ancestries. Nevertheless, our understanding of the role of epistasis in human traits is limited [5].

Over the past decade, a number of methods to detect epistasis have been developed. The first class of methods explicitly search for pairs of genetic variants (usually single nucleotide polymorphisms or SNPs) that have a non-linear effect on a trait. While allowing for an unbiased search for epistasis (analogous to GWAS enabling an unbiased approach to detect associations), these methods pose serious challenges. Exhaustively searching all pairs of SNPs is computationally difficult (scaling quadratically in the number of SNPs). Further, testing such a large number of hypotheses while controlling the false positive rate requires imposing stringent significance thresholds (scaling quadratically in the number of SNPs if a Bonferroni correction were to be used) which, in turn, reduces power. Efforts to solve this problem have involved the use of statistical techniques [144, 143, 114, 121, 36, 142], algorithmic innovations [85] or hardware infrastructure [50, 35, 62, 100, 140, 40, 122, 46, 126]. Alternate strategies have attempted to reduce the set of SNPs analyzed either restricting to analysis to SNPs identified in GWAS [111, 17, 58] or that are biologically functional [65, 6]. An alternate approach to detect epistasis aims to test for the aggregate epistatic effect across SNPs. These approaches parallel the development of variance components analysis (also known as mixed models) that have improved power to detect additive genetic effects in aggregate (in contrast to GWAS, which aims to identify individual effects). In this framework, it is of interest to test if the effect of a SNP on a trait is modulated by an individual's genetic background. Such tests of *marginal epistasis* [9, 10] can improve power on account of the reduced multiple testing burden (that now scales with the number of SNPs) and due to the aggregation of a number of weak epistatic signals.

Even with the potential improvements in power, it is likely the case that tests of marginal epistasis need to be applied to datasets with large samples to identify robust signals of epistasis [27, 123, 43]. The availability of datasets that contain genetic and phenotypic information across hundreds of thousands of individuals offers an opportunity to detect epistasis with confidence. Estimating marginal epistasis from large data sets such as the UK Biobank consisting of $\approx 500,000$ individuals genotyped at nearly one million SNPs is computationally intractable.

We study the problem of testing whether the effect of a SNP on a trait is modulated by the individual's genotype at the remaining SNPs. Given genotypes collected from N individuals across M SNPs, we consider a model that aims to estimate and test the marginal epistatic effect defined as the combined pairwise interaction effects between a given SNP and all other SNPs while controlling for linear, additive effects [9]. We present the **FA**st **M**arginal **E**pistasis test, to test for ME of a SNP on a trait. Our algorithm is a streaming randomized method-of-moments estimator that has a runtime sub-linear in the size of the genotype matrix thereby able to test for epistasis of a single SNP with a background of half a million SNPs across $\approx 300K$ individuals in a few hours.

5.2 Materials and Methods

5.2.1 Marginal epistasis model

To identify SNPs involved in epistasis while retaining statistical power, we focus on identifying SNPs that have non-zero interaction effect with any other variants based on the following model :

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{E\alpha} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \frac{\sigma_g^2}{M} \boldsymbol{I}_M) \\ \boldsymbol{\alpha} &\sim \mathcal{N}(\mathbf{0}, \frac{\sigma_{gxg}^2}{M-1} \boldsymbol{I}_{M-1}) \end{aligned} \tag{5.1}$$

Where $\boldsymbol{X}$ denotes a $N \times M$ genotype matrix, $\boldsymbol{y}$ denotes a N-vector of phenotypes and $\boldsymbol{E}$ denotes a $N \times M - 1$ gene-by-gene interaction matrix defined as $\boldsymbol{E} = \boldsymbol{X_{-i}} \odot \boldsymbol{X_{:i}}$ where $\boldsymbol{X_{:i}}$ is the i-th column of $\boldsymbol{X}$ and $\boldsymbol{X_{-i}}$ is formed by excluding vector $\boldsymbol{X_{:i}}$ from $\boldsymbol{X}$. Here $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In this model, σ_e^2, σ_g^2, and σ_{gxg}^2 are the residual variance, genetic variance and gene-by-gene variance components respectively. Here $\boldsymbol{\beta}$ denotes M-vector of SNPs effect sizes, and $\boldsymbol{\alpha}$ denotes $M-1$-vector of GxG effect sizes.

We assume without loss of generality that $\boldsymbol{y}$ is centered and the columns of $\boldsymbol{X}$ are standardized. To estimate the variance components of our LMM, we use a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments. Since $\mathbb{E}\left[\boldsymbol{y}\right] = 0$, we derived the MoM estimates by equating the population covariance to the empirical covariance. The population covariance is given by:

$$\boldsymbol{\Sigma} = cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sigma_g^2 \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \sigma_{gxg}^2 \frac{1}{M-1}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} + \sigma_e^2 \boldsymbol{I} \tag{5.2}$$

Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to estimate the variance parameters :

$$(\tilde{\sigma}_g^2, \tilde{\sigma}_{gxg}^2, \tilde{\sigma}_e^2) = argmin_{(\sigma_g^2, \sigma_{gxg}^2, \sigma_e^2)} ||\boldsymbol{y}\boldsymbol{y}^T - \left(\sigma_g^2 \boldsymbol{K}_1 + \sigma_{gxg}^2 \boldsymbol{K}_2 + \sigma_e^2 \boldsymbol{K}_3\right)||_F^2 \tag{5.3}$$

where $\boldsymbol{K}_1 = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, $\boldsymbol{K}_2 = \frac{1}{M-1}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$ and $\boldsymbol{K}_3 = \boldsymbol{I}_N$. The MoM estimator satisfies the following normal equations:

$$\boldsymbol{T} = \boldsymbol{\sigma^2}\boldsymbol{q} \tag{5.4}$$

where $\boldsymbol{T}$ is a 3×3 matrix with entries $T_{kl} = tr(\boldsymbol{K}_k \boldsymbol{K}_l), k, l \in \{1, 2, 3\}$, and $\boldsymbol{q}$ is a 3-vector with entries $c_k = \boldsymbol{y}^T \boldsymbol{K}_k \boldsymbol{y}$. To compute the variance of the $\tilde{\boldsymbol{\sigma^2}} = (\tilde{\sigma}_g^2, \tilde{\sigma}_{gxg}^2, \tilde{\sigma}_e^2)^T$, we have:

$$\tilde{\boldsymbol{\sigma^2}} = \boldsymbol{T}^{-1}\boldsymbol{q} \tag{5.5}$$

therefore, the covariance matrix of $\tilde{\sigma^2}$ is:

$$\mathrm{Cov}\left[\tilde{\boldsymbol{\sigma^2}}\right] = \boldsymbol{T}^{-1}\mathrm{Cov}\left[\boldsymbol{q}\right]\boldsymbol{T}^{-1}$$

where

$$\mathrm{Cov}\left[\boldsymbol{q}\right] = E[\boldsymbol{q}\boldsymbol{q}^T] - E[\boldsymbol{q}]E[\boldsymbol{q}]^T$$

such that

$$\begin{aligned} \text{Cov}\,[\boldsymbol{q}]_{ij} &= E[\boldsymbol{y}^T\boldsymbol{K}_i\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{K}_j\boldsymbol{y}] - E[\boldsymbol{y}^T\boldsymbol{K}_i\boldsymbol{y}]E[\boldsymbol{y}^T\boldsymbol{K}_j\boldsymbol{y}] \\ &= 2tr(\boldsymbol{\Sigma}\boldsymbol{K}_i\boldsymbol{\Sigma}\boldsymbol{K}_j) + tr(\boldsymbol{K}_i\boldsymbol{\Sigma})tr(\boldsymbol{K}_j\boldsymbol{\Sigma}) - (tr(\boldsymbol{K}_i\boldsymbol{\Sigma})tr(\boldsymbol{K}_j\boldsymbol{\Sigma})) \\ &= 2tr(\boldsymbol{\Sigma}\boldsymbol{K}_i\boldsymbol{\Sigma}\boldsymbol{K}_j) \end{aligned} \tag{5.6}$$

In the derivation 5.6, the vector $\boldsymbol{y}$ is assumed to be a zero-mean Gaussian random vector (see Chapter 3 for the details of derivation).

5.2.2 Computational challenges

Computing the coefficients $tr(\boldsymbol{K}_k\boldsymbol{K}_l)$ of the system of linear equation 5.4 and $\text{Cov}\,[\boldsymbol{q}]_{ij}$ require $\mathcal{O}(N^2M)$ time complexity and $\mathcal{O}(NM)$ memory usage imposing challenging memory and computation requirements for Biobank-scale data (N in the hundreds of thousands, M in the millions). To perform hypothesis testing, we are required to compute p-values. There are two well-known approaches; the first is a normal test requiring the point estimate and standard error, which is appropriate when we have large sample sizes N. The second approach is called the Davies method [13], valid in small sample sizes but its time complexity is $\mathcal{O}(N^3M)$, which is not feasible when we have large sample sizes. Therefore, existing methods for estimating and testing epistatic effects can not be applied to Biobank-scale data sets.

As demonstrated in Chapter 3, similar to our methodology for estimating GxE effect, we can efficiently estimate the variance components and their associated standard errors in the model 5.1.

5.3 Results

5.3.1 Calibration

First, we assessed FAME in terms of controlling type I error by applying it to simulated data in the absence of gene-gene interaction effects. We simulated phenotypes based on genotypes from unrelated white British individuals in the UK Biobank ($M = 459,792$ SNPs, $N = 291,273$ individuals). We simulated phenotypes by varying the additive heritability $\sigma_g^2 \in \{0.25, 0.5\}$, the proportion of variants $p \in \{0.01, 0.10\}$ that have non-zero additive effects (causal variants). We assessed the calibration of FAME when applied to two sets of target SNPs. The first set consists of randomly chosen SNPs. The second set consists of SNPs that were identified to have a significant additive effect from a GWAS ($p < 5 \times 10^{-8}$) and were chosen to mirror our real data analyses.

While FAME is calibrated when the target ME SNPs are selected at random, the p-values tend to be inflated when the target SNPs are selected from a GWAS, we believe this is partially due to the correlation between the linear component and ME component due to the linkage disequilibrium around the target SNP. To address this issue, we exclude the SNPs within the LD block around the target SNP when constructing the matrix $\boldsymbol{E}_t$. This approach effectively controlled the type-I error across different settings with no significant ME signal detected across any of the settings (Figure 5.1).

5.3.2 Estimating marginal epistasis in the UK Biobank

We attempted to test for marginal epistasis in 53 quantitative traits measured across $N = 291,273$ unrelated white British individuals with genotypes measured across $M = 459,792$ SNPs on the UK Biobank array. To identify sets of target SNPs to test for ME, we chose SNPs that were found to be associated with the trait in a GWAS. Specifically, we ran GWAS on each trait, including as covariates sex, age, and the top 20 genetic PCs. For each trait, we selected SNPs with p-value $p < \frac{5\times10^{-8}}{53}$ that we then pruned to remove high-LD SNPs (using a window size of 500 SNPs, we computed r^2 between each pair and removed one of them if $r^2 > 0.1$, shifting the window by 1 SNP, and repeating the process). We tested for ME at each remaining SNP using the model defined in Equation 5.1 in which we accounted for the linear additive effect of genome-wide SNPs and included age, sex, and the top 20 PCs as fixed effects. We excluded SNPs in the LD block surrounding the target SNP in these tests.

We applied FAME to test for ME at SNPs associated with 53 quantitative traits across

≈ 300 K unrelated white British individuals in the UK Biobank. Testing $16,914$ trait-loci associations, we identified 23 trait-loci pairs across traits demonstrating strong evidence of ME signals ($p < \frac{5\times10^{-8}}{53}$). We further partitioned the ME signals across the genome to identify 6 trait-loci pairs with strong evidence of local (within-chromosome) ME while 20 show evidence of distal (cross-chromosome) ME. Our results provide evidence for epistatic effects underlying complex traits that can now be interrogated in large sample sizes (Figure 5.2).

5.3.3 Studying loci with significant ME effects

We observe the largest ratio of h^2_{gxg} to h^2_{gwas} at SNP rs628031 (chr6:160,560,845) that shows significant ME for serum lipoprotein A levels (lipoA). This variant is a non-synonymous polymorphism that changes methionine to valine in the protein product of the organic cation transporter gene OCT1 (also known as SLC22A1). OCT1 mediates the uptake and efflux of cationic metabolites in the liver that includes as its substrates a variety of drugs including metformin that is widely used to treat type 2 diabetes [102]. Genetic variation in OCT1 has been shown to modulate the response to metformin and to other drugs [102].

SNP rs964184 (chr11:116,648,917) shows significant ME for multiple traits: Apolipoprotein B, cholesterol, and triglycerides with substantial ME effects ($\frac{h^2_{gxg}}{h^2_{gwas}}$ = 5.14, 7.83, and 0.59 respectively). This variant lies in the 3' UTR region of the ZPR1 gene (also referred to as ZNF259) that encodes a zinc finger protein that is known to play a regulatory role in cell proliferation and signal transduction [26]. The promoter region of ZPR1 is known to be bound by transcription factors that play a role in insulin sensitivity, cholesterol metabolism, and obesity. rs964184, as well as other variants in ZPR1, have been found to be associated with serum LDL-C [115], HDL-C [51], triglyceride levels [115, 77, 92] and risk for coronary artery disease (CAD) [99] in diverse populations. A regulatory role for rs964184 has been suggested based on its location in a DNaseI hypersensitive region and its overlap with an enhancer that is active in tissues relevant for lipid biology [77]. Further, rs964184 has been association with DNA methylation of a CpG site in the promoter region of the

APOA5 gene [72], potentially explaining the association between DNA methylation level at this site and triglyceride levels [83]. Integrative analyses of genotype and gene expression data have shown rs964184 to play a regulatory role: being a cis-eQTL for genes PCSK7, SIDT2, TAGLN, and BUD13 while also a trans-eQTL for TMEM165, YPEL5, PPM1B, and OBFC2A [137]. Further, mediation analyses revealed that a substantial proportion of the effect of rs964184 on HDL-C and triglycerides is mediated through its trans association with PPM1B and YPEL5 [137].

5.3.3.1 Robustness

Population stratification in GWAS is commonly accounted for by including principal components (PCs) computed from genotype data as covariates in the analysis [86, 87]. A concern is that this approach might not adequately correct for the confounding effects of population stratification on tests of ME effects. To explore the effect of population stratification, we reran our analyses on trait-loci pairs previously discovered as significant with the number of PCs included as covariates increased to 40 (from 20). We observe a high correlation in the p-values when using 40 vs 20 PCs (Figure 5.3; $\rho = 0.973$). Importantly, most of the significant trait-loci pairs remain significant after including the top 40 PCs, indicating that our findings are robust to population stratification.

A second concern with our analyses arises from the fact that the UK Biobank array includes only a subset of all SNPs, which could lead to the inference of spurious non-linear effects [38, 16, 128].

To address this concern, we analyzed our significant ME signals on imputed genotypes across unrelated white British individuals and $4,824,392$ SNPs (MAF $> 1\%$). We observed 19 out of the 23 significant trait-loci pairs detected on the array SNPs also demonstrate significance ($p \leq \frac{5\times10^{-8}}{53}$) on the imputed SNPs (the remaining four loci had p-values $p \leq 10^{-6}$ on the imputed SNPs). We observed considerable p-value correlation $\rho = 0.613$ of ME between array SNPs and imputed SNPs (Figure 5.3).

5.4 Discussion

We have presented a method for testing marginal epistasis (ME) in Biobank-scale data. FAME yields calibrated results in simulations. Applying FAME to 53 quantitative phenotypes in the UK Biobank, we found 16 trait-loci pairs with significant signals of ME, a vast majority of which remain significant after testing with additional PCs to correct for population stratification, and on imputed genotypes to reduce the impact of missing causal SNPs.

While these observations provide strong evidence for epistatic effects in genetic variants that have been discovered to be associated with complex traits using GWAS, our results have several limitations. Firstly, it is plausible that the impact of population structure on epistatic effects might not be well-modeled by approaches, such as including principal components based on common genetic variants as covariates, that have successfully tested additive effects. Secondly, prior studies have shown that epistasis tests can have inflated false-positive rates due to imperfect tagging of causal variants that have large additive effects [39]. Our simulations show that FAME is robust to imperfect tagging of causal variants. Further, the replication of signals discovered using array SNPs on imputed SNPs, which are unlikely to miss causal variants common in the population, makes the issue of missing causal variants less likely. Nevertheless, it is plausible that the distributions of causal variants and the LD patterns between causal and genotyped variants could be complex, impacting our method's calibration. While the number of loci showing ME effects is small (in part due to the stringent p-value threshold that we impose and the GWAS selection strategy that we used to identify target SNPs), we observe that the proportion of variance explained by ME is comparable to, and sometimes substantially larger than, the proportion of variance explained by GWAS. These results suggest that the polygenic background can substantially modulate the effect of genetic variants on traits and has implications for efforts to annotate genetic variants and to understand how genetic effects vary across populations [78]. We caution, however, that since we are analyzing SNPs with significant ME and GWAS signals, the effect size estimates are likely to be biased upwards due to winner's curse [130].

We further partitioned the ME signal within and across chromosomes to detect both within and cross-chromosomal signals and found 6 within chromosomal signals, a strict subset of the 20 cross-chromosomal signals. This observation suggests that the epistatic signal that we detect is likely to be polygenic so the approach of testing for the aggregate effects as we do here is likely to be more powerful than an approach that aims to identify specific pairs of SNPs. While our current application of FAME has focused on genome-wide signals of ME where we test a single target SNP against a background set consisting of SNPs across the genome (excluding those in the LD block as the target), the model underlying FAME is flexible, and can be applied to test for epistasis in other settings. For example, FAME can be extended to test for interactions of a target SNP or other covariates (such as polygenic scores) with a background set of SNPs where the set is defined based on functional annotation, such as genes or pathways. The ideas underlying FAME allow such tests to be applied to biobank-scale data. Such an approach can improve our understanding by localizing the ME signal. Despite its scalability, FAME is still not efficient enough to perform genome-wide scans of ME, which, in turn, led us to focus on testing for ME at GWAS loci. Extending the scope and efficiency of FAME presents important directions for future work.

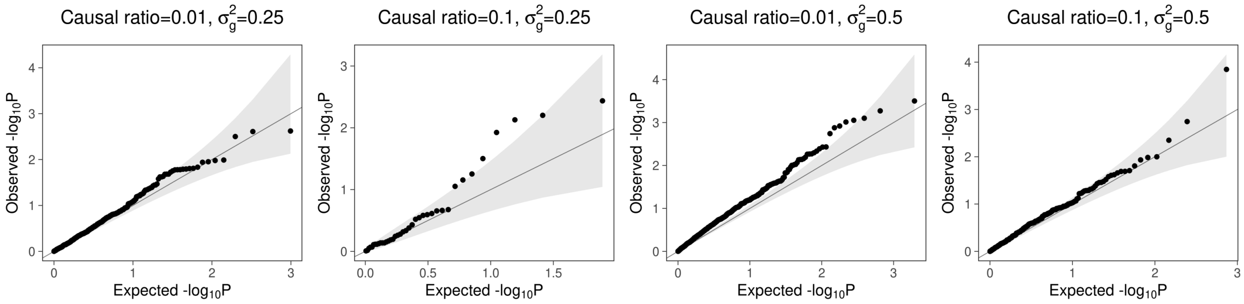
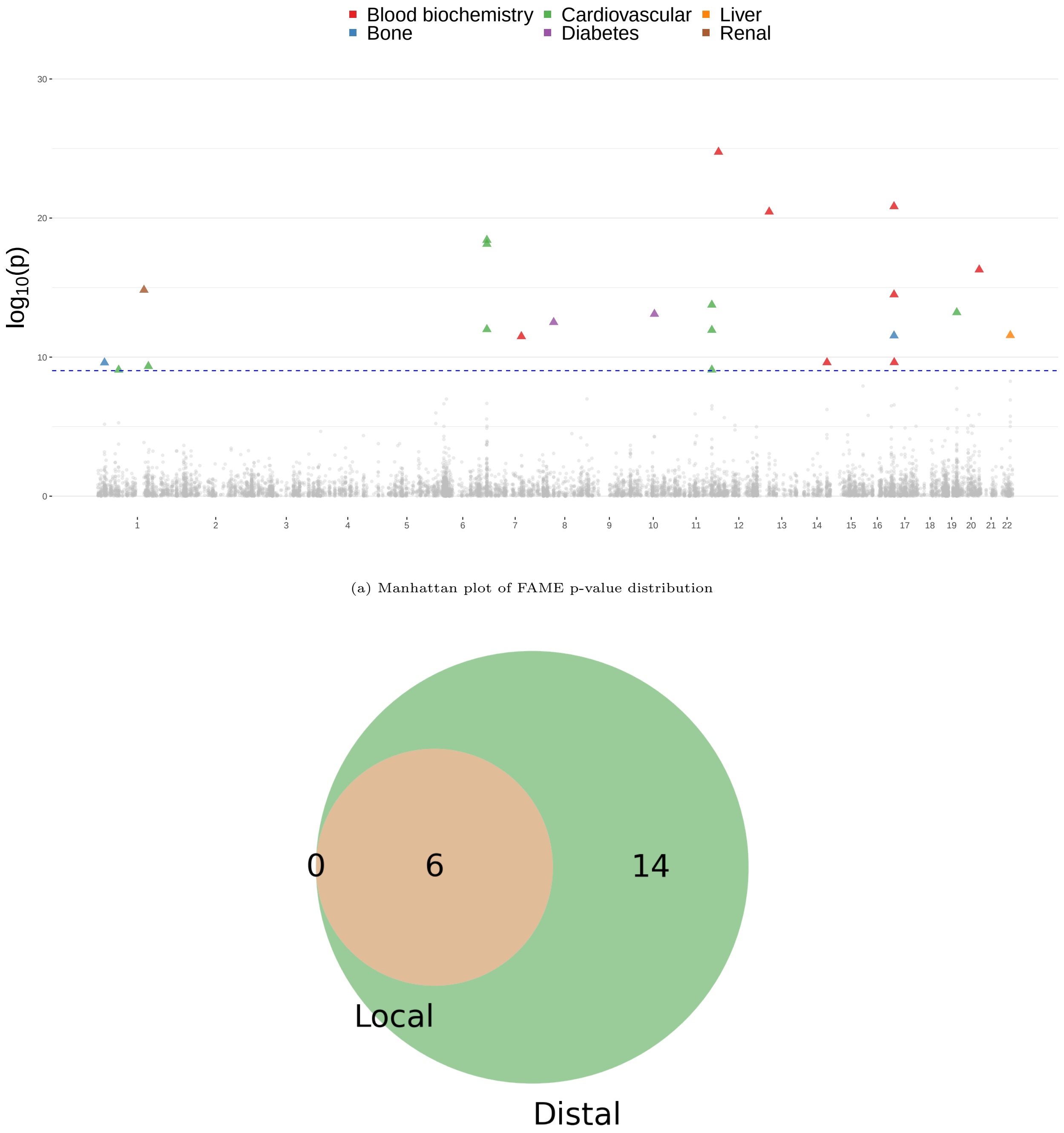


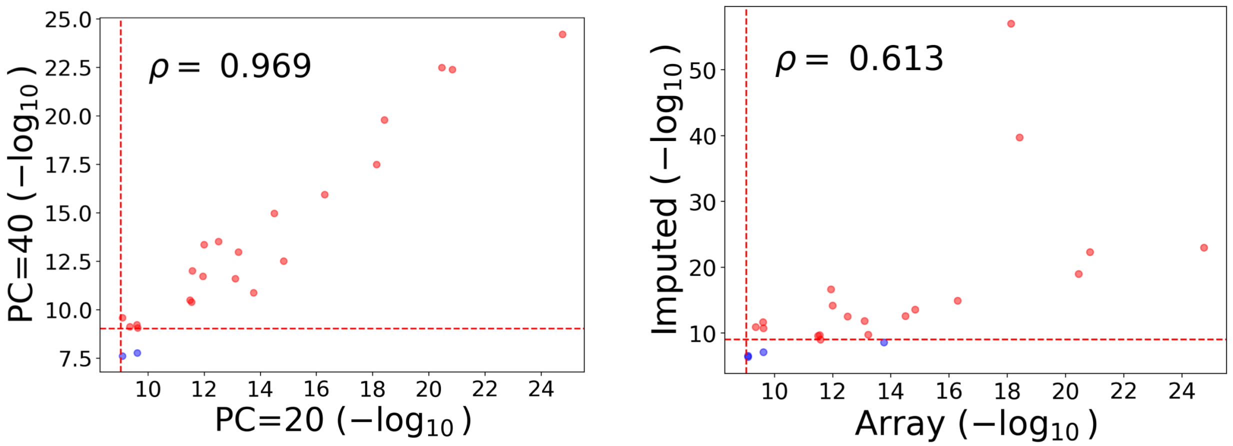
Figure 5.1: **Calibration in simulations.** We applied FAME to phenotypes simulated from genotypes with linear additive effects but no marginal epistatic (ME) effects. Phenotypes were simulated using genotypes measured on $\approx 300K$ unrelated white-British individuals in the UK Biobank, with varying ratios of causal SNPs (Causal ratio) and heritability (h^2). We first ran GWAS to identify significant SNPs, which were then used as target SNPs in a test of ME. We detected no significant ME signals ($p \leq 5 \times 10^{-8}$) across all the settings.



(a) Manhattan plot of FAME p-value distribution

(b) Localization of ME signals

Figure 5.2: **Significant marginal epistasis (ME) loci.** (a) Manhattan plot of the ME loci across 53 complex traits in UKBB. Colored shapes denote significant trait-loci pairs at $p \leq \frac{5\times10^{-8}}{53}$; (b) Localization of ME signals. For each of 23 trait-loci pairs, we tested whether the ME signals remained significant when testing against all SNPs on the same chromosome as the target SNP (after removing SNPs in the same LD block as the target SNP), which we term *local*, and against all SNPs on chromosomes different from the chromosome containing the target SNP, which we term *distal*. We then compared the overlap between the *local* and *distal* significant signals ($p \leq \frac{5\times10^{-8}}{53}$).



(a)

(b)

Figure 5.3: **Robustness in real data analysis** (a) We assessed the robustness of ME signals to population stratification. We tested the trait-loci pairs, which were significant for ME signals, and repeated the test by varying the number of principal components (PC=20 vs. PC=40). We plot the p-values from both analyses. (b) We assessed the robustness of ME signals to the missingness of features. We tested the trait-loci pairs, which were significant for ME signals in whole genome array data, and repeated the test by switching to the imputed dataset. We plot the p-values from both analyses.

CHAPTER 6

Conclusions

In this thesis, our primary focus was on the development of scalable and robust statistical models within the framework of variance component analysis. Our objective was to study important questions in human genetics through the analysis of large-scale genotype-phenotype data.

Chapter 2 introduced the RHE-mc method, which aimed to quantify the proportion of phenotypic variation that can be explained by a linear function of genotypes. We estimated the heritability of a set of complex traits and partitioned the heritability across the genomes, taking into account functional annotations, genes, pathways, as well as minor allele frequency (MAF) and linkage disequilibrium (LD).

In Chapter 3, we presented a novel method called GENIE, which explored the influence of genetic background on the effect of an environmental variable on phenotype. Specifically, we investigated the presence of a Gene-Environment interaction (GxE) effect and examined how this interaction was distributed across the genomes, considering tissue-specific genes, functional annotations, MAF, and LD.

Chapter 4 focused on assessing the contribution of interactions between alleles at a specific locus to phenotypic variations, beyond the additive effects. We quantified the magnitude of these interactions and examined their impact on phenotype variations.

In Chapter 5, we introduced the FAME method, which aimed to delve into the exploration of interactions between different SNPs contributing to phenotypic variance. We investigated the presence of SNPs interactions and their genomic locations.

CHAPTER 7

Future directions

Looking toward future works, several potential directions exist to extend our proposed methods. First, we can apply our proposed methods to binary traits available in the UK Biobank, treating these traits as continuous. Developing methods that explicitly model binary traits and the underlying ascertainment involved in case-control studies is likely to lead to more accurate heritability estimates [33, 125]. For example, the phenotype correlation-genotype correlation (PCGC) method [33] presents an extension of HE regression that could be of interest to explore, aiming to develop a scalable randomized PCGC estimator.

Secondly, jointly analyzing multiple genomics datasets is critical for detecting weak yet important genetic signals. However, privacy concerns often restrict data sharing between different sources. To address this bottleneck, developing secure and scalable methods to analyze genomic databases without explicit data sharing becomes imperative. This problem poses significant challenges as existing methods for privacy-preserving computations often introduce computational bottlenecks, hindering their application to large-scale genomic datasets and statistical models. Leveraging ideas from our current work, we can explore the development of cryptographic protocols for genomic analysis that preserve privacy while maintaining scalability.

7.1 Randomized PCGC

Assume the following linear mixed model, which relates phenotypes to genotypes.

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \frac{\sigma_g^2}{M} \boldsymbol{I}_M) \end{aligned} \tag{7.1}$$

$$cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sum_k \sigma_k^2 \boldsymbol{K}_k + \sigma_e^2 \boldsymbol{I}_N$$

Here $\boldsymbol{K} = \frac{\boldsymbol{X}\boldsymbol{X}^T}{M}$ is the genetic relatedness matrix (GRM) computed from all SNPs. Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to find the variance components.

$$\begin{aligned} (\tilde{\sigma_e^2}, \tilde{\sigma_g^2}) &= argmin_{(\sigma_g^2,, \sigma_e^2)} ||\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I})||_F^2 \\ &= argmin_{(\sigma_g^2, \sigma_e^2)} \sum_{i=1}^{N} \sum_{j=1}^{N} |y_i y_j - (\sigma_g^2 \boldsymbol{K}_{i,j} + \sigma_e^2)|^2 \end{aligned} \tag{7.2}$$

In this model, we assume that the individuals comprise a random sample from the population and the phenotype is an additive polygenic quantitative trait ($E[y_i y_j] = \sigma_g^2 \boldsymbol{K}_{i,j} + \sigma_e^2)$). PCGC regression is based on the simple idea that the heritability of a trait controls the strength of the relationship between genotype and phenotype. In the general case, the relationship among genetic correlation ($\boldsymbol{K}_{i,j}$), phenotypic correlation ($y_i y_j$), and the genetic variance component σ_g^2 can be expressed as

$$E[y_i y_j] = f(\sigma_g^2, \boldsymbol{K}_{i,j}) \tag{7.3}$$

where the function f depends on (i) the design of the study and (ii) the properties of the phenotype. Following [33], f can be approximated by a Taylor series at $\boldsymbol{K}_{i,j} = 0$ as follows:

$$f(\sigma_g^2, \boldsymbol{K}_{i,j}) = c_{i,j} \sigma_g^2 \boldsymbol{K}_{i,j} + o(\boldsymbol{K}_{i,j}) \tag{7.4}$$

where

$$c_{i,j} = \frac{\phi(t_i)\phi(t_j)[1-(P_i+P_j)\alpha + P_iP_j\alpha^2)]}{d_id_j} \tag{7.5}$$

where $\frac{P-K}{P(1-K)} = \alpha$,$d_i = \sqrt{P_i(1-P_i)}(K_i + (1-K_i)\frac{K(1-P)}{P(1-K)})$, K is the fraction of cases in the population, P is the probability that an individual in the study is affected, t_i is individual-specific liability threshold, K_i is an individual-specific of being effected condition on her/his specific covariates, and P_i is an individual-specific of being effected condition on both her/his specific covariates and the fact is selected for study. In this situation, we need to solve the following least squares problem to find the variance components

$$\begin{aligned}(\tilde{\sigma_e^2}, \tilde{\sigma_g^2}) &= argmin_{(\sigma_g^2,\sigma_e^2)} \sum_{i=1}^{N}\sum_{j=1}^{N} |y_iy_j - f(\sigma_g^2, \boldsymbol{K}_{i,j})|^2 \\ &= argmin_{(\sigma_g^2,\sigma_e^2)} ||\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K}\circ\boldsymbol{C} + \sigma_e^2\boldsymbol{I})||_F^2\end{aligned} \tag{7.6}$$

where C is a $N \times N$ matrix where $c_{i,j} = \frac{\phi(t_i)\phi(t_j)[1-(P_i+P_j)\alpha+P_iP_j\alpha^2)]}{d_id_j}$

The MoM estimator satisfies the following normal equations :

$$\begin{bmatrix} tr((\boldsymbol{K}\circ\boldsymbol{C})(\boldsymbol{K}\circ\boldsymbol{C})) & tr(\boldsymbol{K}\circ\boldsymbol{C}) \\ tr(\boldsymbol{K}\circ\boldsymbol{C}) & N \end{bmatrix} \begin{bmatrix} \tilde{\sigma_g^2} \\ \tilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^T(\boldsymbol{K}\circ\boldsymbol{C})\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix} \tag{7.7}$$

In the following, we discuss the efficient computation of all elements of the normal equations.

Lemma 8 *$tr[\boldsymbol{X}\boldsymbol{X}^T \circ \boldsymbol{C}]$ can be estimated in $\mathcal{O}(\frac{NMB}{max(\log_3 N, \log_3 M)})$.*

Proof:

$$\begin{aligned} tr[\boldsymbol{X}\boldsymbol{X}^T\circ\boldsymbol{C}] &= \sum_{i=1}^{N} c_{i,i}\sum_{j=1}^{M} x_{i,j}^2 && (7.8) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M}(\sqrt{c_{i,i}}x_{i,j})^2 && (7.9) \\ &= tr[(\sqrt{diag(\boldsymbol{C})}\boldsymbol{X})(\sqrt{diag(\boldsymbol{C})}\boldsymbol{X})^T] \end{aligned}$$

$diag(\boldsymbol{C})$ can be computed in $\mathcal{O}(N)$. $tr[(\sqrt{diag(\boldsymbol{C})}\boldsymbol{X})(\sqrt{diag(\boldsymbol{C})}\boldsymbol{X})^T]$ can be estimated in $\mathcal{O}(\frac{NMB}{max(\log_3 N, \log_3 M)})$ by using Hutchinson estimator and mailman algorithm described in Chapter 2 and 3.

□

Lemma 9 *Let $\boldsymbol{X}$ be a $N \times M$ matrix. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be N-vectors. Then we have the following :*

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T \circ \boldsymbol{\alpha}\boldsymbol{\beta}^T = (diag(\boldsymbol{\alpha})\boldsymbol{X})(diag(\boldsymbol{\beta})\boldsymbol{X})^T \tag{7.10}$$

Proof: Every element of matrix $\boldsymbol{K}$ can be written as :

$$\begin{aligned} \boldsymbol{K}_{i,j} &= \alpha_i\beta_j X_{i,:}^T X_{j,:} && (7.11)\\ &= \alpha_i\beta_j \sum_{k=1}^{M} x_{i,k}x_{j,k} && (7.12)\\ &= \sum_{k=1}^{M}(\alpha_i x_{i,k})(\beta_j x_{j,k}) && (7.13) \end{aligned}$$

Therefore, we have $\boldsymbol{K} = (diag(\boldsymbol{\alpha})\boldsymbol{X})(diag(\boldsymbol{\beta})\boldsymbol{X})^T$

□

Lemma 10 *$tr[(\boldsymbol{K} \circ \boldsymbol{C})(\boldsymbol{K} \circ \boldsymbol{C})]$ can be computed in $\mathcal{O}(\frac{NMB}{max(\log_3 N, \log_3 M)})$.*

Proof: Let C be a $N \times N$ matrix where $c_{i,j} = \frac{\phi(t_i)\phi(t_j)[1-(P_i+P_j)\alpha+P_iP_j\alpha^2)]}{d_i d_j}$. Assume that $\boldsymbol{\psi}$ is a N-vector such that $\boldsymbol{\psi}_i = \frac{\phi(t_i)}{d_i}$, and $\boldsymbol{p}$ is a N-vector such that $\boldsymbol{p}_i = P_i$. Matrix C can be expressed as follows:

$$C = \boldsymbol{\psi}\boldsymbol{\psi}^T + (\alpha\boldsymbol{\psi} \circ \boldsymbol{P})(\alpha\boldsymbol{\psi} \circ \boldsymbol{P})^T - \alpha(\boldsymbol{\psi} \circ \boldsymbol{P})\boldsymbol{\psi}^T - \alpha\boldsymbol{\psi}(\boldsymbol{\psi} \circ \boldsymbol{P})^T \tag{7.14}$$

let $\boldsymbol{\gamma} = \boldsymbol{\psi} \circ \boldsymbol{P}$, then we have

$$C = \boldsymbol{\psi}\boldsymbol{\psi}^T + \alpha^2\boldsymbol{\gamma}\boldsymbol{\gamma}^T - \alpha\boldsymbol{\gamma}\boldsymbol{\psi}^T - \alpha\boldsymbol{\psi}\boldsymbol{\gamma}^T \tag{7.15}$$

Let $\boldsymbol{A}_1 = \boldsymbol{K} \circ (\boldsymbol{\psi}\boldsymbol{\psi}^T),\ \boldsymbol{A}_2 = \boldsymbol{K} \circ (\alpha^2\boldsymbol{\gamma}\boldsymbol{\gamma}^T), \boldsymbol{A}_3 = \boldsymbol{K} \circ (\alpha\boldsymbol{\gamma}\boldsymbol{\psi}^T),\ \boldsymbol{A}_4 = \boldsymbol{K} \circ (\alpha\boldsymbol{\psi}\boldsymbol{\gamma}^T)$

We have :

$$\begin{aligned} tr[(\boldsymbol{K} \circ \boldsymbol{C})(\boldsymbol{K} \circ \boldsymbol{C})] &= \sum_{i=1}^{4} tr[\boldsymbol{A}_i\boldsymbol{A}_i] \qquad (7.16) \\ &+ 2(tr[\boldsymbol{A}_1\boldsymbol{A}_2] + tr[\boldsymbol{A}_3\boldsymbol{A}_4] - tr[\boldsymbol{A}_1\boldsymbol{A}_3] - tr[\boldsymbol{A}_1\boldsymbol{A}_4] - tr[\boldsymbol{A}_2\boldsymbol{A}_4]) \end{aligned}$$

Each term in 7.16 can be estimated efficiently using the Hutchinson trace estimator. For instance:

$$\begin{aligned} tr[\boldsymbol{A}_1\boldsymbol{A}_4] &= tr[\boldsymbol{K} \circ (\boldsymbol{\psi}\boldsymbol{\psi}^T)\boldsymbol{K} \circ (\alpha\boldsymbol{\psi}\boldsymbol{\gamma}^T)] \qquad (7.17) \\ &= tr[(diag(\boldsymbol{\psi})\boldsymbol{X})(diag(\boldsymbol{\psi})\boldsymbol{X})^T(diag(\alpha\boldsymbol{\psi})\boldsymbol{X})(diag(\boldsymbol{\gamma})\boldsymbol{X})^T] \end{aligned}$$

□

Lemma 11 $\boldsymbol{y}^T(\boldsymbol{K} \circ \boldsymbol{C})\boldsymbol{y}$ *can be computed in* $\mathcal{O}(\frac{NMB}{max(\log_3 N, \log_3 M)})$.

Proof:

$$\begin{aligned} \boldsymbol{y}^T(\boldsymbol{K} \circ \boldsymbol{C})\boldsymbol{y} &= \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T \circ (\boldsymbol{\psi}\boldsymbol{\psi}^T + \alpha^2\boldsymbol{\gamma}\boldsymbol{\gamma}^T - \alpha\boldsymbol{\gamma}\boldsymbol{\psi}^T - \alpha\boldsymbol{\psi}\boldsymbol{\gamma}^T))\boldsymbol{y} \qquad (7.18) \\ &= \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T \circ (\boldsymbol{\psi}\boldsymbol{\psi}^T))\boldsymbol{y} + \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T \circ (\alpha^2\boldsymbol{\gamma}\boldsymbol{\gamma}^T))\boldsymbol{y} \qquad (7.19) \\ &- \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T \circ (\alpha\boldsymbol{\gamma}\boldsymbol{\psi}^T))\boldsymbol{y} \\ &- \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T \circ (\alpha\boldsymbol{\psi}\boldsymbol{\gamma}^T))\boldsymbol{y} \end{aligned}$$

Each term in can be estimated efficiently using the Hutchinson trace estimator.

□

7.2 Secure variance component analysis

Here we discuss the challenge of variance component analysis when the data are distributed among multiple parties. Our goal is to propose a secure and scalable approach that allows us to estimate variance components on a combined database without explicitly merging the information sources. The problem becomes particularly relevant when the data involve genomic data, and the parties involved are data centers or biobanks. In such cases, privacy barriers prevent the direct sharing of data among parties, and database owners may have restrictions or preferences in sharing their data. Nevertheless, fitting an LMM on the combined dataset may yield better statistical properties compared to analyzing individual, incomplete datasets. To address this, we propose a protocol that enables the computation of necessary quantities for fitting an LMM on the merged data. This protocol involves a sequence of steps where each party performs local computations and transmits messages to other parties. Our aim is to achieve cryptographic security by following a ”semi-honest” model. This model assumes that each party will adhere to the protocol and use their true input values but also maintains curiosity about the secret inputs of other parties. The security of the protocol relies on ensuring that the messages exchanged during its execution do not reveal information about the secret inputs belonging to each party.

Suppose that we have P parties and each party owns $\boldsymbol{y_p}$ which is an outcome N_p-vector (e.g. phenotypes), and $\boldsymbol{X}_p$ which is a $N_p \times M$ design matrix (e.g. genotype matrix), where N_p is the number of individuals in p-th data set and M is the number of features(e.g. SNPs), $N \ll M$. The goal is to estimate the variance components of the following LMMs:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{D}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N) \\
\boldsymbol{\beta} &\sim \mathcal{D}(\boldsymbol{0}, \frac{\sigma_g^2}{M} \boldsymbol{I}_M)
\end{aligned} \tag{7.20}$$

where $\boldsymbol{y} = (\boldsymbol{y_1}...\boldsymbol{y_P})^T$ is a N-vector, and $\boldsymbol{X} = (\boldsymbol{X}_1...\boldsymbol{X}_P)^T$ is a $N \times M$ matrix where $N = \sum_{p=1}^{P} N_p$.

To estimate the variance components of the LMM, we use a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments. According to the above model, the theoretical covariance of $\boldsymbol{y}$ is :

$$cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \frac{\sigma_g^2}{M}\boldsymbol{X}\boldsymbol{X^T} + \sigma_e^2\boldsymbol{I}_N \tag{7.21}$$

Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to estimate the variance components.

$$(\tilde{\sigma_e^2}, \tilde{\sigma_g^2}) = argmin_{(\sigma_g^2,,\sigma_e^2)}||\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I})||_F^2 \tag{7.22}$$

It turns out that the MoM estimator satisfies the following normal equations:

$$\begin{bmatrix} \frac{1}{M^2}tr(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T) & \frac{1}{M}tr(\boldsymbol{X}\boldsymbol{X}^T) \\ \frac{1}{M}tr(\boldsymbol{X}\boldsymbol{X}^T) & N \end{bmatrix} \begin{bmatrix} \tilde{\sigma_g^2} \\ \tilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{M}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix} \tag{7.23}$$

As we discussed the Chapter 2, instead of computing the exact value of $tr(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T)$, we use Hutchinson's estimator of the trace [48] as follows:

$$tr(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T) \approx \frac{1}{B}\sum_b \boldsymbol{z}_b^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b \tag{7.24}$$

Here $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$ are B independent random vectors with zero mean and covariance $\boldsymbol{I}_N$. Our method draws these random vectors independently from a standard normal distribution. Therefore, the total running time will be $\mathcal{O}(NMB)$, which is linear in the size of the design matrix. It turns out that $B \approx 10$ is sufficient.

Without loss of generality, suppose we have two parties. Fist party owns $(\boldsymbol{y}_1, \boldsymbol{X}_1)$ and second party owns $(\boldsymbol{y}_2, \boldsymbol{X}_2)$. To solve the corresponding normal equation 7.23 , we need secure and efficient computation of the elements of 7.23.

First, we start with $tr(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T)$:

$$tr(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T) \;\approx\; \sum_b \boldsymbol{z}_b^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b \tag{7.25}$$

we have $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)^T$. Let $\boldsymbol{w}_b = \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b$, then $\boldsymbol{w_b}$ can be partitioned to $\boldsymbol{w_b} = (\boldsymbol{w_{b1}}, \boldsymbol{w_{b2}})^T$ where

$$\boldsymbol{w}_{b1} = \boldsymbol{X_1}\boldsymbol{X_1}^T\boldsymbol{z}_{b1} + \boldsymbol{X_1}\boldsymbol{X_2}^T\boldsymbol{z}_{b2} = \boldsymbol{v}_{b11} + \boldsymbol{v}_{b12} \tag{7.26}$$

$$\boldsymbol{w}_{b2} = \boldsymbol{X_2}\boldsymbol{X_2}^T\boldsymbol{z}_{b2} + \boldsymbol{X_2}\boldsymbol{X_1}^T\boldsymbol{z}_{b1} = \boldsymbol{v}_{b22} + \boldsymbol{v}_{b21} \tag{7.27}$$

here $\boldsymbol{z_b} = (\boldsymbol{z_{b1}}, \boldsymbol{z_{b2}})^T$ where $\boldsymbol{z_{bi}}$ is a N_i-vector. Therefore, we have

$$\begin{aligned} tr(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T) &\approx \sum_b \boldsymbol{z}_b^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b = \sum_b \boldsymbol{w_b}^T\boldsymbol{w_b} && (7.28)\\ &= \sum_b \boldsymbol{w_{b1}}^T\boldsymbol{w_{b1}} + \boldsymbol{w_{b2}}^T\boldsymbol{w_{b2}} && (7.29)\\ &= \sum_b (\boldsymbol{v}_{b11}^T\boldsymbol{v}_{b11} + 2\boldsymbol{v}_{b11}^T\boldsymbol{v}_{b12} + \boldsymbol{v}_{b12}^T\boldsymbol{v}_{b12}) + (\boldsymbol{v}_{b22}^T\boldsymbol{v}_{b22} + 2\boldsymbol{v}_{b22}^T\boldsymbol{v}_{b21} + \boldsymbol{v}_{b21}^T\boldsymbol{v}_{b21}) \end{aligned}$$

Scalars $\boldsymbol{v}_{bii}^T\boldsymbol{v}_{bii}$ and vectors $\boldsymbol{v}_{bii}$ for $i \in \{1,2\}$ can be computed locally by the respective party without the need for interaction between the parties. Computing the other terms (e.g. $\boldsymbol{v}_{b12}^T\boldsymbol{v}_{b12}, \boldsymbol{v}_{b12}$) needs interactions between two parties. For example, to compute $\boldsymbol{v}_{b12} = \boldsymbol{X_1}\boldsymbol{X_2}^T\boldsymbol{z}_{b2}$, second party can compute $\boldsymbol{u}_{b2} = \boldsymbol{X_2}^T\boldsymbol{z}_{b2}$ locally, and computing $\boldsymbol{X}_1\boldsymbol{u}_{b2}$ needs interactions between two parties.

Outcome vector can be decomposed as $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2})^T$. Let $\boldsymbol{r}_i = \boldsymbol{X}_i^T\boldsymbol{y_i}$, then we have

$$\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{r}_1^T\boldsymbol{r}_1 + \boldsymbol{r}_2^T\boldsymbol{r}_2 \tag{7.30}$$

Each of these terms can be computed locally by the respective party, enabling efficient computation of the overall expression. Importantly, this protocol allows for data transmission between parties that is independent of the size of the dataset, denoted by N and M. Thus, the efficiency and scalability of the approach are maintained regardless of the dataset's dimensions. Furthermore, it is worth noting that this protocol can be extended to involve more than two parties, expanding its applicability to a broader range of scenarios involving distributed data analysis.

Bibliography

[1] François Aguet, Andrew A Brown, Stephane E Castel, Joe R Davis, Yuan He, Brian Jo, Pejman Mohammadi, Yoson Park, Princy Parsana, Ayellet V Segre, et al. Genetic effects on gene expression across human tissues. *Nature*, 2017.

[2] Elena Bernabeu, Oriol Canela-Xandri, Konrad Rawlik, Andrea Talenti, James Prendergast, and Albert Tenesa. Sex differences in genetic architecture in the uk biobank. *Nature genetics*, 53(9):1283–1289, 2021.

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

[4] Bycroft C et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 2018.

[5] Örjan Carlborg and Chris S Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.

[6] Gary K Chen and Duncan C Thomas. Using biological knowledge to discover higher order interactions in genetic association studies. *Genetic epidemiology*, 34(8):863–878, 2010.

[7] James M Cheverud and Eric J Routman. Epistasis and its contribution to genetic variance components. *Genetics*, 139(3):1455–1461, 1995.

[8] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

[9] Lorin Crawford, Ping Zeng, Sayan Mukherjee, and Xiang Zhou. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS genetics*, 13(7):e1006869, 2017.

[10] Lorin Crawford and Xiang Zhou. Genome-wide marginal epistatic association mapping in case-control studies. *bioRxiv*, page 374983, 2018.

[11] James F Crow. On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1241–1244, 2010.

[12] Andy Dahl, Khiem Nguyen, Na Cai, Michael J Gandal, Jonathan Flint, and Noah Zaitlen. A robust method uncovers significant context-specific heritability in diverse complex traits. *The American Journal of Human Genetics*, 106(1):71–91, 2020.

[13] Robert B Davies. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, pages 323–333, 1980.

[14] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

[15] Angela Döring, Christian Gieger, Divya Mehta, Henning Gohlke, Holger Prokisch, Stefan Coassin, Guido Fischer, Kathleen Henke, Norman Klopp, Florian Kronenberg, et al. Slc2a9 influences uric acid concentrations with pronounced sex-specific effects. *Nature genetics*, 40(4):430–436, 2008.

[16] Frank Dudbridge and Olivia Fletcher. Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, 95(3):301–307, 2014.

[17] David M Evans, Chris CA Spencer, Jennifer J Pointon, Zhan Su, David Harvey, Grazyna Kochan, Udo Oppermann, Alexander Dilthey, Matti Pirinen, Millicent A Stone, et al. Interaction between erap1 and hla-b27 in ankylosing spondylitis implicates peptide handling in the mechanism for hla-b27 in disease susceptibility. *Nature genetics*, 43(8):761–767, 2011.

[18] Luke M Evans, Rasool Tahmasbi, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Steven Gazal, Douglas W Bjelland, Teresa R Candia, Michael E Goddard, Benjamin M Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics*, 50(5):737, 2018.

[19] Marie-Julie Favé, Fabien C Lamaze, David Soave, Alan Hodgkinson, Héloïse Gauvin, Vanessa Bruat, Jean-Christophe Grenier, Elias Gbeha, Kimberly Skead, Audrey Smargiassi, et al. Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nature communications*, 9(1):1–12, 2018.

[20] Anthony S Findley, Alan Monziani, Allison L Richards, Katherine Rhodes, Michelle C Ward, Cynthia A Kalita, Adnan Alazizi, Ali Pazokitoroudi, Sriram Sankararaman, Xiaoquan Wen, et al. Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions. *Elife*, 10:e67077, 2021.

[21] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228, 2015.

[22] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228, 2015.

[23] Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, 50(4):621–629, 2018.

[24] Ronald Aylmer Fisher et al. 009: The correlation between relatives on the supposition of mendelian inheritance. 1918.

[25] Emily Flynn, Yosuke Tanigawa, Fatima Rodriguez, Russ B Altman, Nasa Sinnott-Armstrong, and Manuel A Rivas. Sex-specific genetic effects across biomarkers. *European Journal of Human Genetics*, 29(1):154–163, 2021.

[26] Zoya Galcheva-Gargova, Konstantin N Konstantinov, I-Huan Wu, F George Klier, Tamera Barrett, and Roger J Davis. Binding of zinc finger protein zpr1 to the epidermal growth factor receptor. *Science*, 272(5269):1797–1802, 1996.

[27] W James Gauderman. Sample size requirements for association studies of gene-gene interaction. *American journal of epidemiology*, 155(5):478–484, 2002.

[28] Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10):1421, 2017.

[29] Steven Gazal, Po-Ru Loh, Hilary K Finucane, Andrea Ganna, Armin Schoech, Shamil Sunyaev, and Alkes L Price. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet*, 50:1600–1607, 2018.

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

[31] Arthur R Gilmour, Robin Thompson, and Brian R Cullis. Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450, 1995.

[32] Daniel Glass, Ana Viñuela, Matthew N Davies, Adaikalavan Ramasamy, Leopold Parts, David Knowles, Andrew A Brown, Åsa K Hedman, Kerrin S Small, Alfonso Buil, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome biology*, 14(7):R75, 2013.

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

[34] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[35] Attila Gyenesei, Jonathan Moody, Asta Laiho, Colin AM Semple, Chris S Haley, and Wen-Hua Wei. Biforce toolbox: powerful high-throughput computational analysis of gene–gene interactions in genome-wide association studies. *Nucleic acids research*, 40(W1):W628–W632, 2012.

[36] Attila Gyenesei, Jonathan Moody, Colin AM Semple, Chris S Haley, and Wen-Hua Wei. High-throughput analysis of epistasis in genome-wide association studies with biforce. *Bioinformatics*, 28(15):1957–1964, 2012.

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

[38] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, et al. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249–253, 2014.

[39] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, et al. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249–253, 2014.

[40] Gibran Hemani, Athanasios Theocharidis, Wenhua Wei, and Chris Haley. Epigpu: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, 27(11):1462–1465, 2011.

[41] Charles R Henderson. Estimation of variance and covariance components. *Biometrics*, 9(2):226–252, 1953.

[42] William G Hill, Michael E Goddard, and Peter M Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*, 4(2):e1000008, 2008.

[43] Valentin Hivert, Julia Sidorenko, Florian Rohart, Michael E Goddard, Jian Yang, Naomi R Wray, Loic Yengo, and Peter M Visscher. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *bioRxiv*, 2020.

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

[45] Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature genetics*, 51(8):1244–1251, 2019.

[46] Jie Kate Hu, Xianlong Wang, and Pei Wang. Testing gene–gene interactions in genome wide association studies. *Genetic epidemiology*, 38(2):123–134, 2014.

[47] Wen Huang and Trudy FC Mackay. The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS genetics*, 12(11):e1006421, 2016.

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

[49] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

[50] Tony Kam-Thong, Benno Pütz, Nazanin Karbalai, Bertram Müller-Myhsok, and Karsten Borgwardt. Epistasis detection on quantitative phenotypes by exhaustive enumeration using gpus. *Bioinformatics*, 27(13):i214–i221, 2011.

[51] Sekar Kathiresan, Cristen J Willer, Gina M Peloso, Serkalem Demissie, Kiran Musunuru, Eric E Schadt, Lee Kaplan, Derrick Bennett, Yun Li, Toshiko Tanaka, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics*, 41(1):56–65, 2009.

[52] Matthew Kerin and Jonathan Marchini. Inferring gene-by-environment interactions with a bayesian whole-genome regression model. *The American Journal of Human Genetics*, 107(4):698–713, 2020.

[53] David A Knowles, Joe R Davis, Hilary Edgington, Anil Raj, Marie-Julie Favé, Xiaowei Zhu, James B Potash, Myrna M Weissman, Jianxin Shi, Douglas F Levinson, et al. Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods*, 14(7):699–702, 2017.

[54] Melanie Kolz, Toby Johnson, Serena Sanna, Alexander Teumer, Veronique Vitart, Markus Perola, Massimo Mangino, Eva Albrecht, Chris Wallace, Martin Farrall, et al. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS genetics*, 5(6):e1000504, 2009.

[55] Anthony YC Kuk and Yuk W Cheng. The monte carlo newton-raphson algorithm. *Journal of Statistical Computation and Simulation*, 59(3):233–250, 1997.

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

[57] S Hong Lee, Teresa R DeCandia, Stephan Ripke, Jian Yang, Patrick F Sullivan, Michael E Goddard, Matthew C Keller, Peter M Visscher, Naomi R Wray, Schizophrenia Psychiatric Genome-Wide Association Study Consortium, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nature genetics*, 44(3):247, 2012.

[58] Juan Pablo Lewinger, John L Morrison, Duncan C Thomas, Cassandra E Murcray, David V Conti, Dalin Li, and W James Gauderman. Efficient two-step testing of gene-gene interactions in genome-wide association studies. *Genetic epidemiology*, 37(5):440–451, 2013.

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

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

[61] Jun S Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

[62] Yang Liu, Haiming Xu, Suchao Chen, Xianfeng Chen, Zhenguo Zhang, Zhihong Zhu, Xueying Qin, Landian Hu, Jun Zhu, Guo-Ping Zhao, et al. Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS genetics*, 7(3):e1001338, 2011.

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

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

[65] Li Ma, Ariel Brautbar, Eric Boerwinkle, Charles F Sing, Andrew G Clark, and Alon Keinan. Knowledge-driven analysis identifies a gene–gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS genetics*, 8(5):e1002714, 2012.

[66] Asko Mäki-Tanila and William G Hill. Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–367, 2014.

[67] Andrew R Marderstein, Emily R Davenport, Scott Kulm, Cristopher V Van Hout, Olivier Elemento, and Andrew G Clark. Leveraging phenotypic variability to identify genetic interactions in human phenotypes. *The American Journal of Human Genetics*, 108(1):49–67, 2021.

[68] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

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

[70] Kaarina Matilainen, Esa A Mäntysaari, Martin H Lidauer, Ismo Strandén, and Robin Thompson. Employing a monte carlo algorithm in newton-type methods for restricted maximum likelihood estimation of genetic parameters. *PloS one*, 8(12):e80821, 2013.

[71] Charles E McCulloch and Shayle R Searle. *Generalized, linear, and mixed models.* John Wiley & Sons, 2004.

[72] Erika L Moen, Xu Zhang, Wenbo Mu, Shannon M Delaney, Claudia Wing, Jennifer McQuade, Jamie Myers, Lucy A Godley, M Eileen Dolan, and Wei Zhang. Genome-wide variation of cytosine modifications between european and african populations and the implications for complex traits. *Genetics*, 194(4):987–996, 2013.

[73] Rachel Moore, Francesco Paolo Casale, Marc Jan Bonder, Danilo Horta, Lude Franke, Inês Barroso, and Oliver Stegle. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics*, 51(1):180–186, 2019.

[74] Hakhamanesh Mostafavi, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*, 9:e48376, 2020.

[75] Hakhamanesh Mostafavi, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*, 9:e48376, 2020.

[76] Guiyan Ni, Julius Van Der Werf, Xuan Zhou, Elina Hyppönen, Naomi R Wray, and S Hong Lee. Genotype–covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. *Nature communications*, 10(1):2239, 2019.

[77] Esteban J Parra, Andrew Mazurek, Christopher R Gignoux, Alexandra Sockell, Michael Agostino, Andrew P Morris, Lauren E Petty, Craig L Hanis, Nancy J Cox, Adan Valladares-Salgado, et al. Admixture mapping in two mexican samples identifies significant associations of locus ancestry with triglyceride levels in the bud13/znf259/apoa5 region and fine mapping points to rs964184 as the main driver of the association signal. *PLoS One*, 12(2):e0172880, 2017.

[78] Roshni A Patel, Shaila A Musharoff, Jeffrey P Spence, Harold Pimentel, Catherine Tcheandjieu, Hakhamanesh Mostafavi, Nasa Sinnott-Armstrong, Shoa L Clarke, Courtney J Smith, VA Million Veteran Program, et al. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *The American Journal of Human Genetics*, 109(7):1286–1297, 2022.

[79] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[80] Ali Pazokitoroudi, Alec M Chiu, Kathryn S Burch, Bogdan Pasaniuc, and Sriram Sankararaman. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *The American Journal of Human Genetics*, 108(5):799–808, 2021.

[81] Ali Pazokitoroudi, Yue Wu, Kathryn S. Burch, Kangcheng Hou, Aaron Zhou, B. Pasaniuc, and S. Sankararaman. Efficient variance components analysis across millions of genomes. *Nature Communications*, 11, 2020.

[82] Ali Pazokitoroudi, Yue Wu, Kathryn S. Burch, Kangcheng Hou, Aaron Zhou, B. Pasaniuc, and S. Sankararaman. Efficient variance components analysis across millions of genomes. *Nature Communications*, 11, 2020.

[83] L Pfeiffer, S Wahl, LC Pilling, E Reischl, JK Sandling, S Kunze, et al. Dna methylation of lipid-related genes affects blood lipid levels. *Circ Cardiovasc Genet*, 8(2):334–42, 2015.

[84] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[85] Snehit Prabhu and Itsik Pe'er. Ultrafast genome-wide scan for snp–snp interactions in common complex disease. *Genome research*, 22(11):2230–2240, 2012.

[86] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[87] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics*, 11(7):459–463, 2010.

[88] Sara L Pulit, Charli Stoneman, Andrew P Morris, Andrew R Wood, Craig A Glastonbury, Jessica Tyrrell, Loïc Yengo, Teresa Ferreira, Eirini Marouli, Yingjie Ji, et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of european ancestry. *Human molecular genetics*, 28(1):166–174, 2019.

[89] Joshua C Randall, Thomas W Winkler, Zoltán Kutalik, Sonja I Berndt, Anne U Jackson, Keri L Monda, Tuomas O Kilpeläinen, Tõnu Esko, Reedik Mägi, Shengxu Li, et al. Sex-stratified genome-wide association studies including 270,000 individuals

show sexual dimorphism in genetic loci for anthropometric traits. *PLoS genetics*, 9(6):e1003500, 2013.

[90] Mathias Rask-Andersen, Torgny Karlsson, Weronica E Ek, and Åsa Johansson. Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects. *Nature communications*, 10(1):339, 2019.

[91] Konrad Rawlik, Oriol Canela-Xandri, and Albert Tenesa. Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome biology*, 17:1–8, 2016.

[92] Robert W Read, Karen A Schlauch, Vincent C Lombardi, Elizabeth T Cirulli, Nicole L Washington, James T Lu, and Joseph J Grzymski. Genome-wide identification of rare and common variants driving triglyceride levels in a nevada population. *Frontiers in Genetics*, 12:639418, 2021.

[93] Matthew R Robinson, Geoffrey English, Gerhard Moser, Luke R Lloyd-Jones, Marcus A Triplett, Zhihong Zhu, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, Harold Snieder, Tonu Esko, et al. Genotype–covariate interaction effects and the heritability of adult body mass index. *Nature genetics*, 49(8):1174, 2017.

[94] James Niels Rosenquist, Steven F Lehrer, A James O'Malley, Alan M Zaslavsky, Jordan W Smoller, and Nicholas A Christakis. Cohort of birth modifies the association between fto genotype and bmi. *Proceedings of the National Academy of Sciences*, 112(2):354–359, 2015.

[95] Daniel E Runcie and Lorin Crawford. Fast and flexible linear mixed models for genome-wide genetics. *PLoS genetics*, 15(2):e1007978, 2019.

[96] Katherine S Ruth, Felix R Day, Jessica Tyrrell, Deborah J Thompson, Andrew R Wood, Anubha Mahajan, Robin N Beaumont, Laura Wittemans, Susan Martin, Alexander S Busch, et al. Using human genetics to understand the disease impacts of testosterone in men and women. *Nature medicine*, 26(2):252–258, 2020.

[97] Naveed Sattar, David Preiss, Heather M Murray, Paul Welsh, Brendan M Buckley, Anton JM de Craen, Sreenivasa Rao Kondapally Seshasai, John J McMurray, Dilys J Freeman, J Wouter Jukema, et al. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. *The Lancet*, 375(9716):735–742, 2010.

[98] Armin P Schoech, Daniel M Jordan, Po-Ru Loh, Steven Gazal, Luke J O'Connor, Daniel J Balick, Pier F Palamara, Hilary K Finucane, Shamil R Sunyaev, and Alkes L Price. Quantification of frequency-dependent genetic architectures in 25 uk biobank traits reveals action of negative selection. *Nature communications*, 10(1):790, 2019.

[99] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre FR Stewart, Maja Barbalic, Christian Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338, 2011.

[100] Thierry Schüpbach, Ioannis Xenarios, Sven Bergmann, and Karen Kapur. Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, 26(11):1468–1469, 2010.

[101] Brian H Shirts, Sandra J Hasstedt, Paul N Hopkins, and Steven C Hunt. Evaluation of the gene–age interactions in hdl cholesterol, ldl cholesterol, and triglyceride levels: the impact of the sort1 polymorphism on ldl cholesterol levels is age dependent. *Atherosclerosis*, 217(1):139–141, 2011.

[102] Yan Shu, Steven A Sheardown, Chaline Brown, Ryan P Owen, Shuzhong Zhang, Richard A Castro, Alexandra G Ianculescu, Lin Yue, Joan C Lo, Esteban G Burchard, et al. Effect of genetic variation in the organic cation transporter 1 (oct1) on metformin action. *The Journal of clinical investigation*, 117(5):1422–1431, 2007.

[103] Dmitry Shungin, Thomas W Winkler, Damien C Croteau-Chonka, Teresa Ferreira, Adam E Locke, Reedik Mägi, Rona J Strawbridge, Tune H Pers, Krista Fischer, Anne E Justice, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, 2015.

[104] Jeannette Simino, Gang Shi, Joshua C Bis, Daniel I Chasman, Georg B Ehret, Xiangjun Gu, Xiuqing Guo, Shih-Jen Hwang, Eric Sijbrands, Albert V Smith, et al. Gene-age interactions in blood pressure regulation: a large-scale investigation with the charge, global bpgen, and icbp consortia. *The American Journal of Human Genetics*, 95(1):24–38, 2014.

[105] Nasa Sinnott-Armstrong, Sahin Naqvi, Manuel Rivas, and Jonathan K Pritchard. Gwas of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife*, 10:e58615, feb 2021.

[106] Kerrin S Small, Marijana Todorčević, Mete Civelek, Julia S El-Sayed Moustafa, Xiao Wang, Michelle M Simon, Juan Fernandez-Tajes, Anubha Mahajan, Momoko Horikoshi, Alison Hugill, et al. Regulatory variants at klf14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nature genetics*, 50(4):572–580, 2018.

[107] Doug Speed and David Balding. Better estimation of snp heritability from summary statistics provides a new understanding of the genetic architecture of complex traits. *bioRxiv*, page 284976, 2018.

[108] Doug Speed, Na Cai, Michael R Johnson, Sergey Nejentsev, David J Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature genetics*, 49(7):986, 2017.

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

[110] Camelia Stancu and Anca Sima. Statins: mechanism of action and effects. *Journal of cellular and molecular medicine*, 5(4):378–387, 2001.

[111] Amy Strange, Francesca Capon, Chris CA Spencer, Jo Knight, Michael E Weale, Michael H Allen, Anne Barton, Gavin Band, Celine Bellenguez, Judith GM Bergboer, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between hla-c and erap1. *Nature Genetics*, 42(11):985–990, 2010.

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

[113] Jae Hoon Sul, Michael Bilow, Wen-Yun Yang, Emrah Kostem, Nick Furlotte, Dan He, and Eleazar Eskin. Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS genetics*, 12(3):e1005849, 2016.

[114] Wanwan Tang, Xuebing Wu, Rui Jiang, and Yanda Li. Epistatic module detection for case-control studies: a bayesian model with a gibbs sampling strategy. *PLoS genetics*, 5(5):e1000464, 2009.

[115] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[116] Susan A Treloar and Nicholas G Martin. Age at menarche as a fitness trait: nonadditive genetic variance detected in a large twin sample. *American journal of human genetics*, 47(1):137, 1990.

[117] Peter M Visscher, Gibran Hemani, Anna AE Vinkhuyzen, Guo-Bo Chen, Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Jian Yang. Statistical power to detect genetic (co) variance of complex traits using snp data in unrelated samples. *PLoS genetics*, 10(4):e1004269, 2014.

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

[119] Zulma G Vitezica, Luis Varona, and Andres Legarra. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, 195(4):1223–1230, 2013.

[120] Pierrick Wainschtein, Deepti P Jain, Loic Yengo, Zhili Zheng, L Adrienne Cupples, Aladdin H Shadyab, Barbara McKnight, Benjamin M Shoemaker, Braxton D Mitchell, Bruce M Psaty, et al. Recovery of trait heritability from whole genome sequence data. *bioRxiv*, page 588020, 2019.

[121] Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson LS Tang, and Weichuan Yu. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, 2010.

[122] Zhengkui Wang, Yue Wang, Kian-Lee Tan, Limsoon Wong, and Divyakant Agrawal. eceo: an efficient cloud epistasis computing model in genome-wide association study. *Bioinformatics*, 27(8):1045–1051, 2011.

[123] Wen-Hua Wei, Gibran Hemani, and Chris S Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733, 2014.

[124] Xinzhu Wei, Christopher R Robles, Ali Pazokitoroudi, Andrea Ganna, Alexander Gusev, Arun Durvasula, Steven Gazal, Po-Ru Loh, David Reich, and Sriram Sankararaman. The lingering effects of neanderthal introgression on human complex traits. *eLife*, 12:e80757, mar 2023.

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

[126] Lars Wienbrandt, Jan Christian Kässens, Jorge González-Domínguez, Bertil Schmidt, David Ellinghaus, and Manfred Schimmler. Fpga-based acceleration of detecting statistical epistasis in gwas. *Procedia Computer Science*, 29:220–230, 2014.

[127] Thomas W Winkler, Anne E Justice, Mariaelisa Graff, Llilda Barata, Mary F Feitosa, Su Chu, Jacek Czajkowski, Tõnu Esko, Tove Fall, Tuomas O Kilpeläinen, et al. The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS genetics*, 11(10):e1005378, 2015.

[128] Andrew R Wood, Marcus A Tuke, Mike A Nalls, Dena G Hernandez, Stefania Bandinelli, Andrew B Singleton, David Melzer, Luigi Ferrucci, Timothy M Frayling, and Michael N Weedon. Another explanation for apparent epistasis. *Nature*, 514(7520):E3–E5, 2014.

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

[130] Rui Xiao and Michael Boehnke. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(5):453–462, 2009.

[131] Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John RB Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, 47(10):1114, 2015.

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

[133] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[134] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[135] Jian Yang, Ruth JF Loos, Joseph E Powell, Sarah E Medland, Elizabeth K Speliotes, Daniel I Chasman, Lynda M Rose, Gudmar Thorleifsson, Valgerdur Steinthorsdottir, Reedik Mägi, et al. Fto genotype is associated with phenotypic variability of body mass index. *Nature*, 490(7419):267–272, 2012.

[136] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza De Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature genetics*, 43(6):519, 2011.

[137] Chen Yao, Brian H Chen, Roby Joehanes, Burcak Otlu, Xiaoling Zhang, Chunyu Liu, Tianxiao Huan, Oznur Tastan, L Adrienne Cupples, James B Meigs, et al. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation*, 131(6):536–549, 2015.

[138] Alexander I Young, Fabian Wauthier, and Peter Donnelly. Multiple novel gene-by-environment interactions modify the effect of fto variants on body mass index. *Nature communications*, 7(1):1–12, 2016.

[139] Alexander I Young, Fabian L Wauthier, and Peter Donnelly. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nature genetics*, 50(11):1608–1614, 2018.

[140] Ling Sing Yung, Can Yang, Xiang Wan, and Weichuan Yu. Gboost: a gpu-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9):1309–1310, 2011.

[141] Jian Zeng, Ronald De Vlaming, Yang Wu, Matthew R Robinson, Luke R Lloyd-Jones, Loic Yengo, Chloe X Yap, Angli Xue, Julia Sidorenko, Allan F McRae, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics*, 50(5):746–753, 2018.

[142] Yu Zhang. A novel bayesian graphical model for genome-wide multi-snp association mapping. *Genetic epidemiology*, 36(1):36–47, 2012.

[143] Yu Zhang, Bo Jiang, Jun Zhu, and Jun S Liu. Bayesian models for detecting epistatic interactions from genetic data. *Annals of human genetics*, 75(1):183–193, 2011.

[144] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

[145] Xiang Zhou. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics*, 11(4):2027, 2017.

[146] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821, 2012.

[147] Carrie Zhu, Matthew J Ming, Jared M Cole, Michael D Edge, Mark Kirkpatrick, and Arbel Harpak. Amplification is the primary mode of gene-by-sex interaction in complex human traits. *Cell Genomics*, 2022.

[148] Zhihong Zhu, Andrew Bakshi, Anna AE Vinkhuyzen, Gibran Hemani, Sang Hong Lee, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, Harold Snieder, Tonu Esko, Lili Milani, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics*, 96(3):377–385, 2015.