IDENTIFYING DISEASE-RELATED GENE-ENVIRONMENT INTERACTIONS BASED ON METHOD OF MOMENTS

By

Linchuan Shen

Bachelor of Economics - Financial Engineering
Sichuan University
2014

Master of Science - Statistics
University of Science and Technology of China
2017

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Mathematical Sciences

Department of Mathematical Sciences
College of Sciences
The Graduate College

University of Nevada, Las Vegas
May 2024

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

May 1, 2024

This dissertation prepared by

Linchuan Shen

entitled

Identifying Disease-Related Gene-Environment Interactions Based on Method of Moments

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Mathematical Sciences
Department of Mathematical Sciences

Amei Amei, Ph.D.
Examination Committee Chair

Malwane Ananda, Ph.D.
Examination Committee Member

Farhad Shokoohi, Ph.D.
Examination Committee Member

Mira Han, Ph.D.
Graduate College Faculty Representative

Alyssa Crittenden, Ph.D.
Vice Provost for Graduate Education & Dean of the Graduate College

ABSTRACT

IDENTIFYING DISEASE-RELATED GENE-ENVIRONMENT INTERACTIONS BASED ON METHOD OF MOMENTS APPROACHES

by

Linchuan Shen

Dr. Amei Amei, Examination Committee Chair
Professor of Mathematics
University of Nevada, Las Vegas, USA

Human diseases are often caused by a complex interplay of multiple factors, including genetics and environmental factors. These factors can play critical roles in the development and progression of diseases. Although genome-wide association studies (GWAS) have successfully identified many genetic variants associated with human diseases, the estimated effects of these variants are small and can explain only a relatively small portion of the heritability of the underlying diseases.

Detecting gene-environment interactions ($G \times E$) can shed light on the biological mechanisms of diseases. However, most existing methods that investigate $G \times E$ only look at how one environmental factor interacts with either common or rare genetic variants, not both. In this study, we propose two approaches to detect interaction effects of an environmental factor and a set of genetic markers containing both rare and common variants.

The first approach is derived from the MinQue for Summary statistics (MQS) method and has been adapted in our study to develop two sub-methods: the MArginal Gene-Environment Interaction Test with RANdom or FIXed genetic effects (MAGEIT_RAN or MAGEIT_FIX). Our second approach leverages the Generalized Method of Moments (GMM), leading to the Gene-Environment Interaction Test based on GMM (GEITGMM). Through simulation studies and real data analysis, we evaluate the performance of these methods. Both the MQS-based MAGEIT_RAN and MAGEIT_FIX, and the GMM-based GEITGMM are grounded in moment estimation and offer analytical tools for examining gene-environment interactions.

ACKNOWLEDGEMENTS

As I sit down, attempting to articulate my appreciation for all those who have influenced my journey at UNLV, I am overwhelmed by the flood of memories that cascade through my mind. This not only marks the culmination of my six years of graduate study, but also signifies the conclusion of my life as a student spanning over two decades.

To begin with, I would like to pay tribute to my advisor, Dr. Amei Amei. Her persistent guidance, both in my academic pursuits and in navigating life beyond the academia, has been invaluable. She devoted countless hours and relentless energy to every detail of my research - from theoretical derivation to application in real datasets, and even manuscript drafting. Although I am far from being a perfect student, her patience, understanding, and mentorship helped me correct my mistakes and grow. More than just an academic counselor, she has been a beacon in my life. I feel lucky to have her as my mentor, and her contributions to my academic trajectory and future career are beyond measure.

Simultaneously, my gratitude extends to Dr. Zuoheng Wang. She opened up an array of research opportunities and ushered me into the domain of genome-wide association studies and biology, which have since become my academic focal points. Beyond that, Dr. Zuoheng Wang played an integral role in every step of my research, guiding me when necessary and expanding my viewpoints. Her influence extended far beyond the realms of academia, nurturing not only a scholar but a lifelong learner, and the lessons I've learned under her tutelage will continue to guide my path as

I move forward.

Deep appreciation is due to Dr. Malwane Ananda. His Mathematical Statistics course over two semesters laid a solid foundation for my future exploration and learning. Additionally, his willingness to serve on my dissertation committee is something I hold in high regard. In a similar vein, my heartfelt thanks go to Dr. Farhad Shokoohi and Dr. Mira Han, who provided invaluable suggestions and feedback for my dissertation. Their involvement in my dissertation committee is an act of generosity that I am deeply appreciative of.

I also owe a great deal of gratitude to numerous professors, classmates, friends, and all others who contributed to sculpting my dissertation. Specifically, my profound appreciation goes to Chong Chen, who not only offered tremendous assistance with supercomputing but was also a crucial resource when it came to navigating the challenges of programming. His timely and effective responses were instrumental in resolving many problems, and his course on parallel computing significantly broadened my programming proficiency. Special recognition goes to Dr. Edwin Oh for his extensive contributions to the pathway analysis of my project. His insightful commentary on biology were truly illuminating and helped steer the course of my research. I extend my gratitude to Dr. Qing Wu, who endowed me with a wealth of background knowledge on bone density, and provided crucial research data, thus enriching my understanding of the field. I am also thankful to Dr. Xiting Yan, whose detailed explanations on cell-cell communication and invaluable advice significantly improved my project. My thanks also go to Dr. Bowen Liu, whose help with C++ coding tremendously accelerated my programming efforts. I am indebted to Jongyun Jung for his patient with handling SNP data. Similarly, my sincere thanks go to Nating Wang for her assistance in processing raw biological data into a format ready for analysis, significantly reducing my workload

and saving valuable time. I also want to thank Ji Qi for the fruitful discussions and brainstorming sessions that we've had regarding the challenges encountered during the research process. Yunqing Liu's assistance with pathway analysis on numerous occasions also deserves acknowledgment. In addition, I am deeply appreciative of all my professors at UNLV, including Dr. Hokwon Cho, Dr. Petros Hadjicostas, Dr. Kaushik Ghosh, Dr. Peter Shiue, and Dr. Hossein Tehrani, among others. Their tutelage in statistics and mathematics provided the essential building blocks of this dissertation and will continue to inspire my future academic pursuits.

I would like to express my sincere gratitude to UNLV, the Graduate College, and the Math Department for granting me the opportunity to pursue my PhD degree in Statistics, and for their generous TA funding and scholarships. These past six years have been a joyous journey, brimming with invaluable academic, teaching, and life experiences, all of which I will cherish for a lifetime.

Last but not least, I want to express my deepest gratitude to my family. Especially my mother, an extraordinary woman, who has offered her unconditional understanding and support throughout the years. Every stride I have made carries the imprint of her love, efforts and contributions. I also extend my profound thanks to my boyfriend, Dr. Gang Xu. He has been not only a constant academic companion but also a pillar of support in my personal life. He is always there for me through thick and thin. I also want to acknowledge my circle of friends including Dr. Jiacheng Cai, Dr. Libo Zhou, Dr. Chengcheng Li, Dr. Li Zhu, Dr. Md Nahid Hasan, Shen Chan Huang, Jeong Jun Lee, Keoni Castellano and Mi Xia, among others, as well as all the other department friends. Their invaluable aid, wisdom, and camaraderie through both joyful and challenging times have enriched my life and ensured my mental well-being amidst academic pressure.

In closing, I feel blessed to have intersected paths with each one of you. Your presence and contributions, whether direct or indirect, have shaped me into the person I am today and have been pivotal to the creation of this dissertation and my subsequent life. The beautiful memories created during my time at UNLV are ones I will treasure. To all of you, I extend my deepest thanks!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

Over almost two decades, genome-wide association studies (GWAS) have significantly revolutionized the field of genetic research [1, 2] and have successfully detected hundreds of thousands to millions of genetic variants across a multitude of genomes that are associated with specific diseases or traits [1]. Despite these promising capabilities, GWAS are not without their inherent constraints and have left certain aspects of genetic research unresolved, highlighting the necessity for further research methodologies to deeply understand the intricate relationships between genetics and disease [1, 3]. A notable challenge, often referred to as the "missing heritability" problem, continues to persist. This issue underscores the fact that the identified variants only explain a minor portion of the estimated heritability for most complex disease [4, 5]. The observed discrepancy might be ascribed to the stringent significance threshold applied in these studies, which potentially overlooks single nucleotide polymorphisms (SNPs) with relatively moderate effects [6, 7]. The inception of methodologies involving larger sample sizes and novel research designs could pave the way for GWAS results to account for a larger fraction of heritability in a variety of complex diseases in future studies [2, 3, 8, 9, 10, 11].

Furthermore, it is critical to understand that the risk of an individual to a certain disease cannot be evaluated in the absence of environmental risk factors, as these factors exhibit a complex interaction with genetic elements. The expectation of gene-gene and gene-environment interactions may help bridge some knowledge gaps in "missing heritability" [12, 13]. Moreover, the practical

utility of GWAS is faced with limitations of interpreting their results. For example, linkage disequilibrium (LD)—a scenario where neighboring genetic variants tend to be inherited together—poses challenges in pinpointing the actual causal variants [14, 15]. All of the aforementioned factors call for further in-depth research [15, 16, 17].

Considering these limitations of GWAS, this dissertation proposes two complementary research approaches to address one of the issues focusing on gene-environment interactions. The interaction between genetic elements and the environment plays significant roles in the development and progression of numerous complex diseases. Statistical models considering gene-environment interactions provide a potential answer to the enduring problem of "missing heritability".

1.2 Literature review

1.2.1 Gene-environment interaction

The causes of human complex diseases are multifactorial, involving a complex interplay between genetic factors and the environment. The impact of environment exposures on disease outcomes may differ among genotypic groups. In many complex diseases, individuals with specific profiles exhibit an increased disease risk only when exposed to a particular environmental factor [18]. For example, many environmental factors, such as smoking, drinking, diet, stress, air quality, influence disease risk, progression and severity [19, 20]. As a result, incorporating gene-environment interactions ($G \times E$) has become crucial in the study of complex traits. Genome-wide association studies (GWAS) have successfully identified many genetic variants associated with human diseases. Nevertheless, the estimated effects of these variants are modest and account for only a small portion of the heritability observed in complex diseases [21]. Several studies have indicated that $G \times E$ might partially account for the missing heritability. Detecting such interactions could offer meaningful

implication in the field of public health and personalized medicine [21, 22].

Traditional $G \times E$ analyses have typically involved evaluating interactions with genetic individually [23, 24, 25]. Potential limitations in such approaches include the burden of multiple hypothesis testing and a failure to account for joint effects shared by multiple variants with similar biological functions, resulting in decreased statistical power [18]. In recent years, genome-wide search for $G \times E$ has been emerging [22, 26]. Several studies have explored $G \times E$ using multiple genetic variants within a marker set [18, 27, 28, 29, 30, 31, 32, 33, 34]. For common genetic variants, a gene-environment set association test (GESAT) was developed using a generalized linear model and ridge regression [18]. For rare variants, Chen *et al.* proposed INT-FIX and INT-RAN to test $G \times E$ effect, along with a joint test, JOINT, to detect the effects of a set of genetic variants and their interactions with an environmental factor simultaneously [29]. Genetic effects were modeled using a beta density function to account for larger contributions from rare genetic variants. In their tests, the genetic main effects when the environmental factor is absent were treated as fixed in INT-FIX and as random in INT-RAN. To assess rare variants by environment interactions, Lin *et al.* developed the interaction sequence kernel association test (iSKAT), where the main effects of rare variants were modeled using weighted ridge regression, and the interactions with the environment across genetic variants were considered to be correlated [28]. These tests are all variance component-based, ensuring robustness when many variants in a genetic region are non-causal and/or exhibit mixed beneficial and detrimental variants [30, 35, 36, 37]. Subsequently, Su *et al.* developed a mixed effects score test for interaction (MiSTi), providing a unified regression framework for testing interaction effects between a set of rare variants and an environmental factor [30]. Many existing methods can be derived from MiSTi by constraining certain parameters to be zero. Additionally, apart from the regression-based $G \times E$ tests mentioned above, Lin *et al.* proposed

the adaptive combination of Bayes factors method (ADABF), a polygenic test of $G \times E$ effect using Bayes factors [27]. This method assumes that $G \times E$ effects follow a normal distribution. Variants in a genetic region were ranked by Bayes factors, and p-values were calculated using a resampling procedure. While ADABF considers both common and rare variants within a genetic region, it does not distinguish between the effects of these two types of variants in model fitting, potentially overlooking the relatively larger contribution from rare variants [35, 38].

1.2.2 Statistical methods based on MoM and GMM

The analysis of GWAS and $G \times E$ have traditionally relied on linear mixed models (LMMs), utilizing maximum likelihood estimation (MLE) and restricted maximum likelihood estimation (REML) for parameter estimation. Recently, the MoM and the GMM have emerged as robust alternatives, offering novel approaches to parameter estimation [39, 40, 41, 42, 43]. In recent years, statistical methodologies in GWAS have increasingly leverage estimation based on the method of moments and LD score regression (LDSC) [44] is a notable example. Rooted in the method of moments, LDSC addresses the challenges associated with the lack of availability of individual-level genotype data and widespread sample overlap among meta-analyses by requiring only GWAS summary statistics and it is shown to be not biased by sample overlap. Similarly, GNOVA [45] employs a MoM framework to estimate annotation-stratified genetic covariance between traits, also using GWAS summary statistics. A benchmark study [42] compares their performance, demonstrating that GNOVA performs similarly to LDSC in terms of effectiveness and robustness. Additionally, the MQS method [46] represents a significant advancement in MoM-based approaches by integrating the Haseman–Elston regression and LDSC into a unified framework. This integration not only broadens the application of summary statistics but also yields more efficient statistical estimates than traditional LDSC.

Compared to MoM, GMM distinguishes itself in its approach to solving parameter estimation problems, which minimizes a quadratic form of the differences between theoretical population moments and empirical sample moments [47, 48]. This distinction allows GMM to provide a more flexible and reliable tool for analyzing complex interactions within GWAS and $G \times E$ studies. Recent studies have introduced methods such as the penalized LMM with generalized method of moments pLMMGMM [41] and MpLMMGMM [49] which are based on the GMM. These methods are designed for high-dimensional data analysis, efficiently detecting predictive markers and offering improved prediction accuracy across various disease models. Compared to existing penalized linear mixed models, these methods adopt GMM estimators, making them more computationally efficient. Compared to pLMMGMM, MpLMMGMM is designed for multi-omics data and is equivalent to pLMMGMM when only genomic or methylation data are considered. Similarly, multi-trait analysis of GWAS (MTAG) [50], another efficient GMM-based method, facilitates the joint analysis of GWAS summary statistics from different traits, even those from overlapping samples. Furthermore, the GMM has been applied to solve overly identified estimating equations, with simulation studies highlighting the efficiency of this approach [51].

1.3 Outline of the dissertation

This dissertation delves into the complex interaction between genetic and environmental factors, a key driver in the development of human diseases. It seeks to shed light on the unresolved issue of unexplained heritability. The organization of the dissertation is as follows:

Chapter 1 begins with the background and literature review, focusing on gene-environment inter-

actions and statistical methods based on the Method of Moments (MoM) and GMM. This chapter presents an overview of the existing research landscape, enumerating existing methodologies and underscoring their shortcomings, thus laying the foundation for the innovative techniques proposed in the subsequent chapters.

In Chapter 2, we propose two groups of novel tests: the first group consists of MAGEIT_RAN and MAGEIT_FIX, which are based on the MQS method, and the second group includes GEITGMM, developed using the GMM method.

In Chapter 3, we conduct simulation studies for MAGEIT_RAN, MAGEIT_FIX, and GEITGMM. Through these studies, we validate the robustness of MAGEIT_RAN and MAGEIT_FIX, demonstrate their effective control of type I error, and reveal that MAGEIT_RAN exhibits higher power than other compared methods in certain scenarios. The performance of GEITGMM is also assessed through simulation studies, offering insights into its effectiveness.

In Chapter 4, we apply MAGEIT_RAN and MAGEIT_FIX to the MESA dataset, conducting a genome-wide analysis to investigate gene-alcohol interactions on hypertension. Employing a suggestive significance threshold for the genome-wide scan, we identify two genes, *CCNDBP1* and *EPB42*. Furthermore, we identify two signal transduction pathways associated with hypertension.

Finally, Chapter 5 emphasizes the novel contributions made in the field of disease-related gene-environment interactions. We summarize the methods proposed in Chapters 2 and discuss potential future research directions in this field.

CHAPTER 2
GENE-ENVIRONMENT INTERACTION TEST

2.1 Marginal gene-environment interaction test with random or fixed genetic effects

Suppose a phenotype of interest, an environmental variable and genome-wide genetic variants are available on n subjects. The genotype of a variant can be directly measured or can consist of imputed values. Let $y_k, E_k, \boldsymbol{G}_k = (G_{k1},\ G_{k2},\ \ldots, G_{kp})^T$ and $\boldsymbol{X}_k = (X_{k1},\ X_{k2},\ \ldots, X_{km})^T$ denote the phenotype, environmental variable, genotypes of p variants in a genomic region, and m non-genetic covariates for the kth subject, respectively, for $k = 1,\ 2, \ldots, n$, where $G_{kj} = 0$, 1 or 2 depending on whether subject k has 0, 1 or 2 copies of minor allele at the jth variant. We use $\boldsymbol{S}_k = \big(E_kG_{k1},\ E_kG_{k2},\ \ldots, E_kG_{kp}\big)^T$ to denote the genetic variants by environment interaction for the kth subject. Our goal is to test whether there are interactions between the variant set and environment that influence the phenotype of interest.

2.1.1 Model for continuous phenotype

Let $\boldsymbol{y} = (y_1,\ y_2,\ \ldots,\ y_n)^T$, $\boldsymbol{E} = (E_1,\ E_2,\ \ldots, E_n)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ denote vectors of the phenotype, environmental variable, and error term of length n. We further define an $n \times m$ covariate matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{\cdots}, \boldsymbol{X}_n]^T$, an $n \times p$ genotype matrix $\boldsymbol{G} = [\boldsymbol{G}_1,\ \boldsymbol{G}_2,\ \boldsymbol{\cdots}, \boldsymbol{G}_n]^T$, and an $n \times p$ matrix $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_n]^T$ of the $G \times E$. Then, the following model specifies the relationship between a continuous phenotype $\boldsymbol{Y}$ and $\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{G}$ and $\boldsymbol{S}$

$$\boldsymbol{y} = \alpha_0\boldsymbol{1} + \boldsymbol{X}\boldsymbol{\alpha}_1 + \alpha_2\boldsymbol{E} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{1}$ is an $n \times 1$ vector of 1, α_0 is an intercept term, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1m})^T$, α_2, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$ are regression coefficients for the covariates, environmental factor, genetic variants, and $G \times E$ terms. We further assume that $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ follow multivariate normal distributions with $\boldsymbol{\gamma} \sim \text{MVN}(\mathbf{0}, \frac{\sigma^2}{p}\boldsymbol{W}_2^2)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \tau^2\boldsymbol{I}_n)$, where $\boldsymbol{W}_2 = \text{diag}(w_{21}, w_{22}, \cdots, w_{2p})$ contains weights of the p $G \times E$ terms and $\boldsymbol{I}_n$ is an identity matrix of dimension n.

2.1.2 Marginal gene-environment interaction test

We are interested in testing genetic variants by environment interactions in a genomic region, i.e., testing the null hypothesis $H_0 : \boldsymbol{\gamma} = \mathbf{0}$, which is equivalent to testing $H_0 : \sigma^2 = 0$. We develop two $G \times E$ tests, in which the genetic main effects $\boldsymbol{\beta}$ are modeled as random or fixed effects, respectively.

When we treat the genetic main effects $\boldsymbol{\beta}$ as random, we assume that $\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \frac{\omega^2}{p}\boldsymbol{W}_1^2)$, where $\boldsymbol{W}_1 = \text{diag}(w_{11}, w_{12}, \cdots, w_{1p})$ are weights of the p variants. We use the MQS method [46] to estimate the three variance components ω^2, σ^2 and τ^2. In order to eliminate the fix effects $\alpha_0, \boldsymbol{\alpha}_1$ and α_2 in Model (2.1), we multiply both sides of the model, from left, by a projection matrix $\boldsymbol{M}$, where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{b}(\boldsymbol{b}^T\boldsymbol{b})^{-1}\boldsymbol{b}^T$ with $\boldsymbol{b} = [\mathbf{1}, \boldsymbol{X}, \boldsymbol{E}]$. Then Model (2.1) becomes

$$\boldsymbol{y}^* = \boldsymbol{g}^* + \boldsymbol{s}^* + \boldsymbol{\varepsilon}^*,$$

where $\boldsymbol{y}^* = \boldsymbol{M}\boldsymbol{y}$, $\boldsymbol{g}^* = \boldsymbol{M}\boldsymbol{G}\boldsymbol{\beta}$, $\boldsymbol{s}^* = \boldsymbol{M}\boldsymbol{S}\boldsymbol{\gamma}$, and $\boldsymbol{\varepsilon}^* = \boldsymbol{M}\boldsymbol{\varepsilon}$. It follows that $\boldsymbol{g}^* \sim \text{MVN}(\mathbf{0}, \omega^2\boldsymbol{G}^*)$ with $\boldsymbol{G}^* = \frac{(\boldsymbol{M}\boldsymbol{G}\boldsymbol{W}_1)(\boldsymbol{M}\boldsymbol{G}\boldsymbol{W}_1)^T}{p}$, $\boldsymbol{s}^* \sim \text{MVN}(\mathbf{0}, \sigma^2\boldsymbol{S}^*)$ with $\boldsymbol{S}^* = \frac{(\boldsymbol{M}\boldsymbol{S}\boldsymbol{W}_2)(\boldsymbol{M}\boldsymbol{S}\boldsymbol{W}_2)^T}{p}$, and $\boldsymbol{\varepsilon}^* \sim \text{MVN}(\mathbf{0}, \tau^2\boldsymbol{M})$. Consequently, we have $\boldsymbol{y}^* \sim \text{MVN}\left(\mathbf{0}, \omega^2\boldsymbol{G}^* + \sigma^2\boldsymbol{S}^* + \tau^2\boldsymbol{M}\right)$.

We estimate the variance components using the method of moments based on the following set of

second moment matching equations,

$$E\left(\boldsymbol{y}^{*T}\boldsymbol{A}\boldsymbol{y}^{*}\right)=\mathrm{tr}\left(\boldsymbol{A}\left(\omega^{2}\boldsymbol{G}^{*}+\sigma^{2}\boldsymbol{S}^{*}+\tau^{2}\boldsymbol{M}\right)\right)=\omega^{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{G}^{*}\right)+\sigma^{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{S}^{*}\right)+\tau^{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{M}\right), \quad (2.2)$$

where $\boldsymbol{A}$ is an arbitrary symmetric non-negative definite matrix [46]. Since there are three unknown parameters $(\omega^2, \sigma^2, \tau^2)$, three different $\boldsymbol{A}$'s are required to obtain parameter estimates. In the method of moments, the expectation of Eq. (2.2) is usually replaced with the realized value $\boldsymbol{y}^{*T}\boldsymbol{A}\boldsymbol{y}^{*}$. Let $\boldsymbol{A}_1=\boldsymbol{G}^{*}$, $\boldsymbol{A}_2=\boldsymbol{S}^{*}$ and $\boldsymbol{A}_3=\boldsymbol{M}$ [46], then, the resulting estimates of the variance components are given in a matrix form as

$$\begin{bmatrix} \widehat{\omega}^2 \\ \widehat{\sigma}^2 \\ \widehat{\tau}^2 \end{bmatrix} = \boldsymbol{\Lambda}^{-1}\begin{bmatrix} \boldsymbol{y}^{*T}\boldsymbol{G}^{*}\boldsymbol{y}^{*} \\ \boldsymbol{y}^{*T}\boldsymbol{S}^{*}\boldsymbol{y}^{*} \\ \boldsymbol{y}^{*T}\boldsymbol{y}^{*} \end{bmatrix} = \begin{bmatrix} \mathrm{tr}(\boldsymbol{G}^{*}\boldsymbol{G}^{*}) & \mathrm{tr}(\boldsymbol{G}^{*}\boldsymbol{S}^{*}) & tr(\boldsymbol{G}^{*}) \\ \mathrm{tr}(\boldsymbol{S}^{*}\boldsymbol{G}^{*}) & \mathrm{tr}(\boldsymbol{S}^{*}\boldsymbol{S}^{*}) & tr(\boldsymbol{S}^{*}) \\ \mathrm{tr}(\boldsymbol{G}^{*}) & \mathrm{tr}(\boldsymbol{S}^{*}) & n-(m+2) \end{bmatrix}^{-1}\begin{bmatrix} \boldsymbol{y}^{*T}\boldsymbol{G}^{*}\boldsymbol{y}^{*} \\ \boldsymbol{y}^{*T}\boldsymbol{S}^{*}\boldsymbol{y}^{*} \\ \boldsymbol{y}^{*T}\boldsymbol{y}^{*} \end{bmatrix},$$

where we used $\mathrm{tr}\left(\boldsymbol{G}^{*}\boldsymbol{M}\right)=\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{G}^{*}\right)=\mathrm{tr}\left(\boldsymbol{G}^{*}\right)$, $\mathrm{tr}\left(\boldsymbol{S}^{*}\boldsymbol{M}\right)=\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{S}^{*}\right)=\mathrm{tr}\left(\boldsymbol{S}^{*}\right)$, $\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\right)=\mathrm{tr}\left(\boldsymbol{M}\right)=n-(m+2)$, and $\boldsymbol{y}^{*T}\boldsymbol{M}\boldsymbol{y}^{*}=\boldsymbol{y}^{*T}\boldsymbol{y}^{*}$. The variance component estimator $\widehat{\sigma}^2$ is considered as the test statistic, which we named as MArginal Gene-Environment Interaction Test with RANdom genetic main effects (MAGEIT_RAN). Specifically, the MAGEIT_RAN test statistic is

$$\widehat{\sigma}^2=\boldsymbol{y}^{*T}\left\{\left(\boldsymbol{\Lambda}^{-1}\right)_{21}\boldsymbol{G}^{*}+\left(\boldsymbol{\Lambda}^{-1}\right)_{22}\boldsymbol{S}^{*}+\left(\boldsymbol{\Lambda}^{-1}\right)_{23}\boldsymbol{I}\right\}\boldsymbol{y}^{*}=\boldsymbol{y}^{*T}\boldsymbol{H}\boldsymbol{y}^{*}, \quad (2.3)$$

where $\boldsymbol{H}=\left(\boldsymbol{\Lambda}^{-1}\right)_{21}\boldsymbol{G}^{*}+\left(\boldsymbol{\Lambda}^{-1}\right)_{22}\boldsymbol{S}^{*}+\left(\boldsymbol{\Lambda}^{-1}\right)_{23}\boldsymbol{I}$.

Under $H_0:\ \sigma^2=0$, $\boldsymbol{y}^{*}\ \sim\ \mathrm{MVN}\left(\boldsymbol{0},\ \omega^{2}\boldsymbol{G}^{*}+\tau^{2}\boldsymbol{M}\right)$, suggesting that $\boldsymbol{y}^{*}$ has the same distribution as $(\omega^{2}\boldsymbol{G}^{*}+\tau^{2}\boldsymbol{M})^{\frac{1}{2}}\boldsymbol{Z}$ with $\boldsymbol{Z}\ \sim\ \mathrm{MVN}(\boldsymbol{0},\boldsymbol{I}_n)$. Therefore, the method of moments estimator $\widehat{\sigma}^2$ follows the same distribution as $\boldsymbol{Z}^{T}((\widehat{\omega}_0^{2}\boldsymbol{G}^{*}+\widehat{\tau}_0^{2}\boldsymbol{M})^{\frac{1}{2}})^{T}\boldsymbol{H}(\widehat{\omega}_0^{2}\boldsymbol{G}^{*}+\widehat{\tau}_0^{2}\boldsymbol{M})^{\frac{1}{2}}\boldsymbol{Z}$, which has a mixture of χ^2distribution $\widehat{\sigma}^2\sim\sum_{i=1}^{n}\lambda_i\chi^2_{1,i}$. Here, $(\widehat{\omega}_0^2,\ \widehat{\tau}_0^2)$ are estimates of (ω^2,τ^2) under the null hypothesis, $(\lambda_1,\cdots,\lambda_n)$ are eigenvalues of the matrix $((\widehat{\omega}_0^{2}\boldsymbol{G}^{*}+\widehat{\tau}_0^{2}\boldsymbol{M})^{\frac{1}{2}})^{T}\boldsymbol{H}\left(\widehat{\omega}_0^{2}\boldsymbol{G}^{*}+\widehat{\tau}_0^{2}\boldsymbol{M}\right)^{\frac{1}{2}}$, and $\chi^2_{1,i}$ are independent χ^2_1 variables [46]. The p-value of $\widehat{\sigma}^2$ can be evaluated by the Davies method [35, 52] and Liu-Tang-Zhang approximation [53].

If we treat the genetic main effects $\boldsymbol{\beta}$ as fixed, we use the MQS method [46] to estimate the two variance components σ^2 and τ^2. To eliminate the fix effect terms $\alpha_0, \boldsymbol{\alpha}_1$, α_2 and $\boldsymbol{\beta}$ in Model (2.1), we left multiply the model by a projection matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{b}(\boldsymbol{b}^T\boldsymbol{b})^{-1}\boldsymbol{b}^T$ with $\boldsymbol{b} = [\boldsymbol{1}, \boldsymbol{X}, \ \boldsymbol{E}, \boldsymbol{G}]$. Then the model becomes $\boldsymbol{y}^* = \boldsymbol{s}^* + \ \boldsymbol{\varepsilon}^*$ and it contains two variance components σ^2 and τ^2. Using the method of moments, we obtain the following estimates of the variance components,

$$\begin{bmatrix} \widehat{\sigma}^2 \\ \widehat{\tau}^2 \end{bmatrix} = \begin{bmatrix} tr(\boldsymbol{S}^*\boldsymbol{S}^*) & tr(\boldsymbol{S}^*) \\ tr(\boldsymbol{S}^*) & n-(m+p+2) \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y}^{*T}\boldsymbol{S}^*\boldsymbol{y}^* \\ \boldsymbol{y}^{*T}\boldsymbol{y}^* \end{bmatrix}.$$

The variance component estimator $\widehat{\sigma}^2$ is considered as the test statistic, which we named as MArginal Gene-Environment Interaction Test with FIXed genetic main effects (MAGEIT_FIX). Specifically, the MAGEIT_FIX test statistic is

$$\widehat{\sigma}^2 = \frac{\boldsymbol{y}^{*T}\{(n-(m+p+2))\,\boldsymbol{S}^* - tr(\boldsymbol{S}^*)\boldsymbol{I}\}\boldsymbol{y}^*}{(n-(m+p+2))\,tr\,(\boldsymbol{S}^*\boldsymbol{S}^*) - tr(\boldsymbol{S}^*)^2}. \tag{2.4}$$

Under $H_0: \ \sigma^2 = 0$, $\widehat{\sigma}^2$ follows a mixture of χ^2 distribution $\widehat{\sigma}^2 \sim \sum_{i=1}^n \lambda_i \chi^2_{1,i}$ with $(\lambda_1, \cdots, \lambda_n)$ being the eigenvalues of the matrix $((\widehat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}})^T \boldsymbol{H}(\widehat{\tau}_0^2\boldsymbol{M})^{\frac{1}{2}}$.

2.1.3 Model for binary phenotype

We consider a liability threshold model and assume the binary outcome y_k of the kth subject is determined by an unobserved continuous liability variable z_k, i.e.,

$$y_k = \begin{cases} 1, & z_k \geq 0 \\ 0, & z_k < 0 \end{cases} \qquad \text{for } k = 1, \ \ldots, \ n, \tag{2.5}$$

where the underlying liability vector $\boldsymbol{z} = (z_1, \ z_2, \cdots, z_n)^T$ is specified using Model (2.1). The full likelihood of the liability threshold mixed effects model is intractable due to an n-dimensional integration over the liability variable $\boldsymbol{z}$. Following the previous studies [54, 55, 56, 57, 58], the liability threshold mixed effects model can be approximated by a linear mixed effects model on

$\widehat{\boldsymbol{z}} = E(\boldsymbol{z}|\boldsymbol{y})$, an estimated posterior mean of the liabilities,

$$\widehat{\boldsymbol{z}} = \alpha_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\alpha}_1 + \alpha_2 \boldsymbol{E} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{2.6}$$

The posterior mean $\widehat{\boldsymbol{z}}$ can be obtained by approximation under certain assumptions based on the properties of GWAS data [58]. Specifically, we assume that (i) subjects are unrelated, and (ii) both the genetic main effects and interaction effects are small such that the terms $\boldsymbol{G}\boldsymbol{\beta}$ and $\boldsymbol{S}\boldsymbol{\gamma}$ can be ignored. Under these assumptions, the distribution of the liability variable can be approximated by $\boldsymbol{z} \sim \text{MVN}(\alpha_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\alpha}_1 + \alpha_2 \boldsymbol{E},\ \boldsymbol{I}_n)$ and $\widehat{\boldsymbol{z}}$ is computed as the mean of the following truncated normal distribution [58]:

$$z_k|y_k \ \sim \ \begin{cases} N\left(\alpha_0 + \boldsymbol{X}_k^T\boldsymbol{\alpha}_1 + \alpha_2 E_k, 1\right) & \text{with } z_k \geq 0 \quad \text{if } y_k = 1 \\ N\left(\alpha_0 + \boldsymbol{X}_k^T\boldsymbol{\alpha}_1 + \alpha_2 E_k, 1\right) & \text{with } z_k < 0 \quad \text{if } y_k = 0 \end{cases} \qquad \text{for } k = 1,\ 2,\ \ldots,\ n\ .$$

The parameters $\alpha_0, \boldsymbol{\alpha}_1$ and α_2 are estimated using a probit model on the phenotype $\boldsymbol{y}$.

To test the interaction effects between a set of genetic variants and an environmental variable on the binary phenotype $\boldsymbol{y}$, we implement MAGEIT_RAN and MAGEIT_FIX on the estimate of the liability variable $\widehat{\boldsymbol{z}}$. To construct MAGEIT_RAN, the liability threshold mixed effects model specified in Eqs (2.5) and (2.6) contains three variance components $(\omega^2, \sigma^2, \tau^2)$, where σ^2 represents a measure of interactions between the p genetic variants and the environmental variable. In order for the model to be identifiable, we put a constraint on the variance of $\boldsymbol{z}$, e.g., $\omega^2 + \sigma^2 + \tau^2 = 1$ [59]. Similarly, we set $\sigma^2 + \tau^2 = 1$ for MAGEIT_FIX.

2.2 Gene-environment interaction test based on GMM

In last section, we employ the MQS method [46], which, despite its advantages, occasionally produces negative estimates for variance components and necessitates methodologies capable of gener-

ating non-negative estimates. This issue aligns with ongoing research efforts that investigate testing variance components on the boundary of the parameter space [60, 61]. To address this crucial need, our study progresses by incorporating Wang's innovative approach [41] into the development of GEITGMM. This method, rooted in the principles of GMM, is applied to gene-environment interaction analyses, ensuring nonnegative variance component estimates.

In Model (2.1), we center the response variable $\boldsymbol{y}$ by subtracting its sample mean from every observation. This adjustment aligns the mean of $\boldsymbol{y}$ to zero. We assume that the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and the error term $\boldsymbol{\varepsilon}$ follow multivariate normal distributions with $\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \frac{\omega^2}{p}\boldsymbol{W}_1)$, $\boldsymbol{\gamma} \sim \text{MVN}(\mathbf{0}, \frac{\sigma^2}{p}\boldsymbol{W}_2^2)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \tau^2\boldsymbol{I}_n)$. Here, $\boldsymbol{W}_1 = \text{diag}(w_{11}, w_{12}, \cdots, w_{1p})$ and $\boldsymbol{W}_2 = \text{diag}(w_{21}, w_{22}, \cdots, w_{2p})$ represent the diagonal matrices containing weights for the p terms associated with $\boldsymbol{G}$ and $\boldsymbol{G} \times \boldsymbol{E}$, respectively, and $\boldsymbol{I}_n$ is the identity matrix of dimension n. As in Section 2.1, we define the projection matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{b}\big(\boldsymbol{b}^T\boldsymbol{b}\big)^{-1}\boldsymbol{b}^T$ with $\boldsymbol{b} = [\mathbf{1}, \boldsymbol{X}, \boldsymbol{E}]$. This matrix $\boldsymbol{M}$ possesses several key properties: (1) $\boldsymbol{M}^T = \boldsymbol{M}$, (2) $\boldsymbol{M}^2 = \boldsymbol{M}$, and (3) its eigenvalues are either 0 or 1. We then perform the decomposition of $\boldsymbol{M} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$, where $\boldsymbol{E} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n-(m+2)}, \ldots, \boldsymbol{v}_n]$ is orthogonal, and $\boldsymbol{D}$ is a diagonal matrix with $n-(m+2)$ 1s and $(m+2)$ 0s on its diagonal. We define matrix $\boldsymbol{A}$ as the first $n-(m+2)$ columns of $\boldsymbol{E}$, which leads to $\boldsymbol{A}$ having the properties: (1) $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_{n-(m+2)}$, (2) $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{M}$, and (3) $\boldsymbol{A}^T\boldsymbol{b} = \mathbf{0}$. For Model (2.1), we left multiply the model by $\boldsymbol{A}^T$, then the model becomes:

$$\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{A}^T\boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{A}^T\boldsymbol{\varepsilon}$$

where $\boldsymbol{G}\boldsymbol{\beta} \sim MVN(\mathbf{0}, \frac{\omega^2}{p}\boldsymbol{G}\boldsymbol{W}_1\boldsymbol{G}^T)$, $\text{var}(\boldsymbol{G}\boldsymbol{\beta}) = \omega^2\frac{\boldsymbol{G}\boldsymbol{W}_1\boldsymbol{G}^T}{p} = \omega^2\boldsymbol{K}_{\boldsymbol{G}}$. Similarly, $\boldsymbol{S}\boldsymbol{\gamma} \sim MVN(\mathbf{0}, \frac{\sigma^2}{p}\boldsymbol{S}\boldsymbol{W}_2\boldsymbol{S}^T)$, $\text{var}(\boldsymbol{S}\boldsymbol{\gamma}) = \sigma^2\frac{\boldsymbol{S}\boldsymbol{W}_2\boldsymbol{S}^T}{p} = \sigma^2\boldsymbol{K}_{\boldsymbol{S}}$. Consequently, the variance of $\boldsymbol{A}^T\boldsymbol{y}$ is independent of the covariates $\boldsymbol{X}$, and we express it as: $\text{var}\big(\boldsymbol{A}^T\boldsymbol{y}\big) = \text{var}\big(\boldsymbol{A}^T\boldsymbol{G}\boldsymbol{\beta}\big) + \text{var}\big(\boldsymbol{A}^T\boldsymbol{S}\boldsymbol{\gamma}\big) + \text{var}(\boldsymbol{A}^T\boldsymbol{\varepsilon})$. Given that $\boldsymbol{y}$ is centered, $\text{var}(\boldsymbol{y}) = E(\boldsymbol{y}\boldsymbol{y}^T)$. We use $\boldsymbol{y}\boldsymbol{y}^T$ to replace $E(\boldsymbol{y}\boldsymbol{y}^T)$, leading to $\text{var}\left(\boldsymbol{A}^T\boldsymbol{y}\right) = \boldsymbol{A}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{A}$.

Following previous work [41], we formulate the GMM estimator by minimizing the squared Frobenius norm of the difference between the observed and expected variance matrices:

$$\begin{bmatrix} \widehat{\omega}^2 \\ \widehat{\sigma}^2 \\ \widehat{\tau}^2 \end{bmatrix} = \underset{\boldsymbol{\delta}^2 \geq 0}{\operatorname{argmin}} \, ||\boldsymbol{A}^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{A} - \omega^2 \boldsymbol{A}^T \boldsymbol{K_G} \boldsymbol{A} - \sigma^2 \boldsymbol{A}^T \boldsymbol{K_S} \boldsymbol{A} - \tau^2 \boldsymbol{I}_{n-(m+2)}||_F^2,$$

where $|| \bullet ||_F^2$ denotes the square of Frobenius norm and $\boldsymbol{\delta}^2 = \left(\omega^2, \sigma^2, \tau^2\right)^T$. Similarly as in [41], to facilitate estimation, we re-parameterize the optimization problem as: $\underset{\boldsymbol{\delta}^2 \geq 0}{\operatorname{argmin}} \; ||\boldsymbol{V} - \boldsymbol{T}\boldsymbol{\delta}^2||_F^2$, $\boldsymbol{V} = vec(\boldsymbol{A}^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{A})$, $\boldsymbol{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_3)$, and $\boldsymbol{T}_1 = vec(\boldsymbol{A}^T \boldsymbol{K_G} \boldsymbol{A})$, $\boldsymbol{T}_2 = vec(\boldsymbol{A}^T \boldsymbol{K_S} \boldsymbol{A})$, $\boldsymbol{T}_3 = vec(\boldsymbol{I}_{n-(m+2)})$, $vec(\bullet)$ is a vectorization of a matrix. This optimization problem can be solved using standard statistical software packages [41].

For hypothesis testing of $H_0 : \sigma^2 = 0$, we employ resampling techniques to compute p-values, utilizing the efficient sequential resampling procedures [27, 62, 63] to enhance computational speed. This process involves permuting the $\boldsymbol{G}$ matrix for each individual to compute null statistics. The p-value is then calculated as the proportion of null statistics that are greater than or equal to the observed statistic. Additionally, we adapt an alternative approach for calculating the p-value: when a null statistic is exactly equal to the observed statistic, we count it as half. The number of resamplings is dynamically decided. Initially, we permute $\boldsymbol{G}$ for 10^3 times. If the p-value derived from these 10^3 null statistics falls below 0.1, the number of resamplings is increased 10-fold to enhance the p-value's accuracy. This procedure is iterated until the p-value exceeds $100/B$, where B represents the total number of resamplings, or until the maximum number of resamplings 10^7 is reached, ensuring a balance between precision and computational feasibility.

CHAPTER 3
SIMULATION STUDIES

3.1 Simulation studies for MAGEIT_RAN and MAGEIT_FIX

We conducted simulation studies to evaluate the performance of MAGEIT_RAN and MAGEIT_FIX to detect set-based $G \times E$ effects for both continuous and binary phenotypes, where the variant set contains both common and rare variants. We assessed type I error and empirical power of MAGEIT_RAN and MAGEIT_FIX, and compared them with three set-based $G \times E$ tests, GESAT-W [18], aMiSTi [30], and ADABF [27]. These three existing methods are popular for $G \times E$ analysis and have well-developed R packages. For fair comparisons, the same weights for rare and common variants were used in all methods except ADABF which does not distinguish common and rare variants and hence no weights were used in the implementation.

3.1.1 Simulation settings

To generate genotypes, we first simulated 100,000 chromosomes over a 5 Kb region using a coalescent model that mimics the linkage disequilibrium (LD) structure and recombination rates of the European population [64, 65]. Then we randomly selected 10 common variants with minor allele frequency (MAF) > 0.05 and 40 rare variants with $0.005 <$ MAF < 0.05 to compose a set of 50 genetic variants.

We simulated a continuous phenotype using the following trait model,

$$y_k = 0.05X_{k1} + 0.057X_{k2} + 0.64E_k + \sum_{j=1}^{10} w_{1j}\beta_j G_{kj} + \sum_{l=1}^{10} w_{2l}\gamma_l E_k G_{kl} + \varepsilon_k,$$

where $X_{k1} \sim \mathrm{N}\left(62.4, 11.5^2\right)$ mimicking age and $X_{k2} \sim \mathrm{Bernoulli}\,(0.52)$ mimicking sex [18]. The 10 genetic variants with main effects and the 10 variants with interaction effects were randomly selected from the set of the 50 variants, independent of E. The environmental variable E is a Bernoulli random variable taking values of 0 or 1 with a probability of 0.5. The weight of a rare variant in w_{1j} or w_{2l} is set to $\mathrm{Beta}\,(\mathrm{MAF}; 1,\ 25)$, the beta density function with parameters 1 and 25 evaluated at the variant's MAF, and the weight of a common variant in w_{1j} or w_{2l} is set to $c\mathrm{Beta}(\mathrm{MAF}; 0.5,\ 0.5)$ with $c = \frac{\mathrm{Beta}(0.05;\ 1,\ 25)}{\mathrm{Beta}(0.05;\ 0.5,\ 0.5)}$ [66, 67]. The error term $\varepsilon_k \sim \mathrm{N}(0,\ 1.5^2)$ indicates independent noise.

For a binary trait, we use the following logistic regression model,

$$\mathrm{logit}\,(P\,(y_k = 1)) = -6.2 + 0.05X_{k1} + 0.057X_{k2} + 0.64E_k + \sum_{j=1}^{10} w_{1j}\beta_j G_{kj} + \sum_{l=1}^{10} w_{2l}\gamma_l E_k G_{kl},$$

where all parameters are the same as those used in the continuous phenotype model. In all simulation settings, each simulated dataset contains 5,000 subjects (2,500 cases and 2,500 controls for binary phenotype).

In the Type I error assessment, we set all γ_l to be 0, i.e., no $G \times E$ effects, and generated 10^6 datasets each containing 50 genetic variants (10 common and 40 rare variants) randomly picked for each dataset. We considered three scenarios as stated in Table 3.1: (1) no genetic main effect, i.e., $\beta_j = 0$ for $j = 1,\ 2,\ \ldots,\ 10$; (2) for continuous/binary phenotype, assigning $\beta_j \sim \mathrm{U}\,(0.07,\ 0.11)/\mathrm{U}\,(0.08,\ 0.12)$ to two randomly selected common variants and $\beta_j \sim \mathrm{U}(0.15,\ 0.19)/\mathrm{U}(0.18,\ 0.22)$ to eight randomly selected rare variants; (3) similar to scenario (2) except that half of the common/rare variants have negative effects.

Table 3.1: The 3 simulation scenarios for Type I error assessment

Scenario	SNP main effects				$G \times E$ effects			
	Common SNPs		Rare SNPs		Common SNPs		Rare SNPs	
	# +	# -	# +	# -	# +	# -	# +	# -
(1)	0	0	0	0	0	0	0	0
(2)	2	0	8	0	0	0	0	0
(3)	1	1	4	4	0	0	0	0

In the power comparison, we designed eight simulation scenarios (Table 3.2) that differ in three key factors that represent different considerations in the simulation design. The first factor pertains to the presence or absence of genetic main effects; the second factor focuses on the allocation of contributions from common and rare variants; and the third factor considers the direction of genetic main effects and $G \times E$ effects, either all positive effects or half positive and half negative effects. We considered ten variants with $G \times E$ effects, either two common and eight rare variants, or four common and six rare variants. The $G \times E$ effect γ_l was generated from $\mathrm{U}\,(0.17,\ 0.21)$ and $\mathrm{U}\,(0.57,\ 0.61)$ for common and rare variants, respectively, for continuous phenotype; and from $\mathrm{U}\,(0.28,\ 0.32)$ and $\mathrm{U}\,(0.86,\ 0.90)$ for common and rare variants, respectively, for binary phenotype. The first four simulation scenarios have no genetic main effect and they are as follow: (1) two common and eight rare variants with positive $G \times E$ effects; (2) two common and eight rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$; (3) four common and six rare variants with positive $G \times E$ effects; and (4) four common and six rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$. The remaining four simulation scenarios have two common and eight rare variants with genetic main effects: (5) β_j was specified the same as in scenario (2) in the type I error assessment, two common and eight rare variants with positive $G \times E$ effects; (6) β_j was specified the same as in scenario (3) in the type I error assessment, two common and eight rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$; (7) β_j was specified the same

as in scenario (2) in the type I error assessment, four common and six rare variants with positive $G \times E$ effects; and (8) β_j was specified the same as in scenario (3) in the type I error assessment, four common and six rare variants with $G \times E$ effects, 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$. Power was evaluated using 1,000 simulated datasets in each scenario.

Table 3.2: The 8 simulation scenarios for power comparison

Scenario	**SNP main effects**				**$G \times E$ effects**			
	Common SNPs		**Rare SNPs**		**Common SNPs**		**Rare SNPs**	
	# +	# -	# +	# -	# +	# -	# +	# -
(1)	0	0	0	0	2	0	8	0
(2)	0	0	0	0	1	1	4	4
(3)	0	0	0	0	4	0	6	0
(4)	0	0	0	0	2	2	3	3
(5)	2	0	8	0	2	0	8	0
(6)	1	1	4	4	1	1	4	4
(7)	2	0	8	0	4	0	6	0
(8)	1	1	4	4	2	2	3	3

3.1.2 Simulation results

Empirical type I error rate was calculated at the nominal level α, for $\alpha = 0.01$, 0.001 and 0.0001, based on 10^6 replicates, under three simulation scenarios, for both continuous and binary phenotypes (Table 3.3). In most simulations, the type I error of MAGEIT_FIX was within the 95% confidence interval of the nominal level, while the type I error of MAGEIT_RAN was lower than the nominal level in all simulation settings, especially for binary phenotype, suggesting that the MQS-based testing procedure tends to produce conservative p-values due to the approximation we used to handle binary phenotype [58, 68].

Table 3.3: Empirical type I error of MAGEIT_RAN and MAGEIT_FIX, based on 10^6 replicates

		Continuous			Binary		
Test	**Level**	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 1**	**Scenario 2**	**Scenario 3**
MAGEIT_RAN	0.01	$\mathbf{9.66 \times 10^{-3}}$	$\mathbf{8.85 \times 10^{-3}}$	$\mathbf{8.62 \times 10^{-3}}$	$\mathbf{9.51 \times 10^{-3}}$	$\mathbf{7.77 \times 10^{-3}}$	$\mathbf{8.47 \times 10^{-3}}$
	0.001	$\mathbf{8.17 \times 10^{-4}}$	$\mathbf{6.09 \times 10^{-4}}$	$\mathbf{5.30 \times 10^{-4}}$	$\mathbf{7.90 \times 10^{-4}}$	$\mathbf{4.92 \times 10^{-4}}$	$\mathbf{5.04 \times 10^{-4}}$
	0.0001	$\mathbf{6.70 \times 10^{-5}}$	$\mathbf{2.90 \times 10^{-5}}$	$\mathbf{3.40 \times 10^{-5}}$	$\mathbf{6.20 \times 10^{-5}}$	$\mathbf{2.80 \times 10^{-5}}$	$\mathbf{2.20 \times 10^{-5}}$
MAGEIT_FIX	0.01	9.87×10^{-3}	9.98×10^{-3}	1.02×10^{-2}	$\mathbf{9.68 \times 10^{-3}}$	$\mathbf{9.67 \times 10^{-3}}$	$\mathbf{9.70 \times 10^{-3}}$
	0.001	9.89×10^{-4}	9.99×10^{-4}	9.57×10^{-4}	9.38×10^{-4}	9.58×10^{-4}	$\mathbf{8.99 \times 10^{-4}}$
	0.0001	1.01×10^{-4}	9.80×10^{-5}	8.80×10^{-5}	9.00×10^{-5}	8.70×10^{-5}	9.20×10^{-5}

The 95% confidence interval of a nominal level α was calculated as $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/10^6}$. Specifically, the 95% confidence intervals are $\left(9.80 \times 10^{-3},\ 1.02 \times 10^{-2}\right)$ for $\alpha = 0.01$, $(9.38 \times 10^{-4},\ 1.06 \times 10^{-3})$ for $\alpha = 0.001$, and $(8.04 \times 10^{-5},\ 1.20 \times 10^{-4})$ for $\alpha = 0.0001$. Rates outside of the 95% confidence interval are in bold.

Empirical power was calculated at the significant level of 10^{-4}, based on 1,000 simulation replicates. Figures 3.3 and 3.2 demonstrate the power results of the five methods, MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi and ADABF, under eight simulation scenarios, for continuous and binary phenotypes, respectively. MAGEIT_RAN had comparable to higher power than the other methods across all simulation scenarios. The high power of MAGEIT_RAN may attribute to its unbiased and statistically efficient estimates of the variance component. Additionally, the genetic effects are treated as random in MAGEIT_RAN, which aligns with a more realistic assumption when the genetic region consists of both common and rare variants. We observed similar patterns for continuous and binary phenotypes. MAGEIT_RAN was much more powerful than other tests when there was no genetic main effect (Scenarios 1-4). For continuous traits, MAGEIT_FIX had comparable power to GESAT-W and higher power than aMiSTi in all simulation scenarios. For binary phenotypes, GESAT-W was comparable or more powerful than MAGEIT_FIX and ADABF. When the $G \times E$ effects had mixed positive and negative directions (Scenarios 2, 4, 6, 8), aMiSTi had the lowest power for both continuous and binary phenotypes. Since aMiSTi is a combination of burden and variance component test, it loses power when there are both protective and detrimental variants in the genomic region being tested [69].

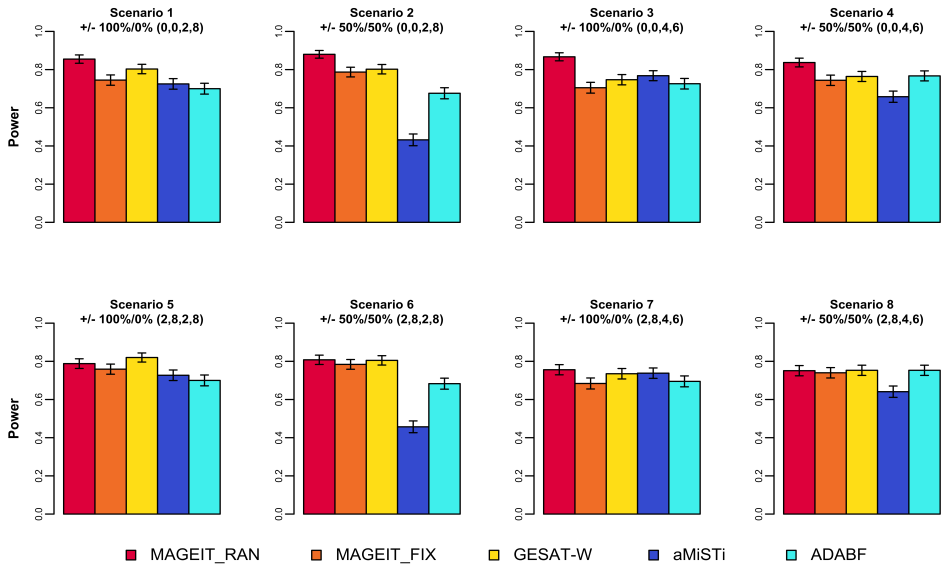

Figure 3.1: Empirical power of MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi and ADABF for a continuous phenotype. Error bars show the approximated 95% confidence interval for the empirical power β, which is calculated as $\beta \pm 1.96\sqrt{\beta(1-\beta)/1000}$.

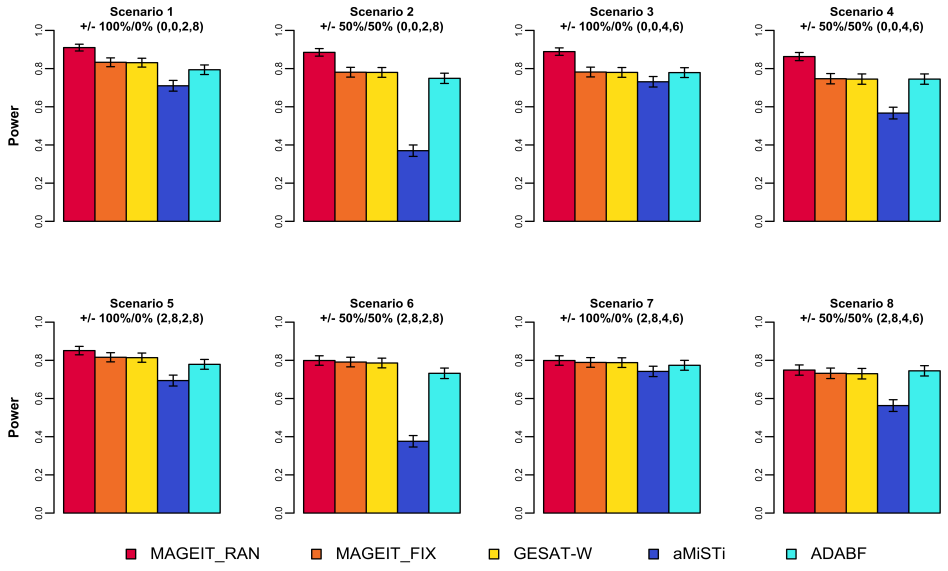

Figure 3.2: Empirical power of MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi and ADABF for a binary phenotype. Error bars show the approximated 95% confidence interval for the empirical power β, which is calculated as $\beta \pm 1.96\sqrt{\beta(1-\beta)/1000}$.

3.2 Simulation studies for MAGEITGMM

We conducted simulation studies to evaluate the performance of GEITGMM in identifying set-based $G \times E$ effects for continuous traits, incorporating both common and rare genetic variants within the vairant set. We assessed the type I error rates and empirical power of GEITGMM, benchmarking its performance against five alternative set-based $G \times E$ tests, MAGEIT_RAN, MAGEIT_FIX, GESAT-W [18], aMiSTi [30], and ADABF [27]. As in Section 3.1, we ensured fair comparisons across all methods being compared by applying the same weighting scheme for rare and common variants except for ADABF.

3.2.1 Simulation settings

Following the procedures outlined in Section 3.1, we simulated 100,000 chromosomes within a 5 Kb region using a coalescent model that mimics the linkage disequilibrium (LD) structure and recombination rates of the European population [64, 65]. Then, we randomly selected 10 common variants, each with MAF > 0.05 and 40 rare variants with $0.005 <$ MAF < 0.05 to compose a set of 50 genetic variants.

We simulated a continuous phenotype using the same trait model as in Section 3.1; i.e.,

$$y_k = 0.05X_{k1} + 0.057X_{k2} + 0.64E_k + \sum_{j=1}^{10} w_{1j}\beta_j G_{kj} + \sum_{l=1}^{10} w_{2l}\gamma_l E_k G_{kl} + \varepsilon_k.$$

Here, X_{k1} follows a normal distribution with a mean of 62.4 and a standard deviation of 11.5, X_{k2} is distributed according to a Bernoulli process with a success probability of 0.52, E adheres to a Bernoulli distribution with a probability of 0.5, and ε_k is normally distributed with a mean of 0 and a standard deviation of 1.5. Consistent with Section 3.1, we assign weights to variants using the same criteria: rare variants in w_{1j} or w_{2l} are weighted by Beta(MAF; 1, 25), the beta density

function with parameters 1 and 25 evaluated at the variant's MAF. In contrast, common variants receive a weight of $c\text{Beta}(\text{MAF}; 0.5,\ 0.5)$, where c is calculated as the ratio of Beta(0.05; 1, 25) to Beta(0.05; 0.5, 0.5), following the previous approach [66, 67]. As in Section 3.1, we randomly selected 10 genetic variants to model main effects and another 10 for interaction effects from a pool of 50 variants, ensuring that this selection was independent of E.

In evaluating type I error, we set all interaction effect coefficients, γ_l, as 0, indicating an absence of $G \times E$ effects. This setup involved generating 5,000 unique datasets, each dataset containing 50 randomly selected genetic variants (10 common and 40 rare variants). We explored two distinct scenarios (Table 3.4) for our analysis: (1) no genetic main effect, i.e., $\beta_j = 0$ for $j = 1,\ 2,\ \ldots,\ 10$; (2) assigning $\beta_j \sim \text{U}\,(0.07,\ 0.11)$ to two randomly selected common variants and $\beta_j \sim \text{U}(0.15,\ 0.19)$ to eight randomly selected rare variants.

Table 3.4: The 2 simulation scenarios for Type I error assessment

Scenario	**SNP main effects**				**$G \times E$ effects**			
	Common SNPs		**Rare SNPs**		**Common SNPs**		**Rare SNPs**	
	# +	# -	# +	# -	# +	# -	# +	# -
(1)	0	0	0	0	0	0	0	0
(2)	2	0	8	0	0	0	0	0

In the power analysis, we evaluated the statistical power of GEITGMM across four distinct simulation scenarios (Table 3.5), delineated by two pivotal factors: the presence or absence of genetic main effects and the direction of $G \times E$ effects, categorized either as all positive or as a mix of half positive and half negative effects. Our simulations focused on a set of ten variants exhibiting $G \times E$ effects, comprised of two common and eight rare variants. The $G \times E$ effect γ_l was generated from $\text{U}\,(0.113,\ 0.153)$ and $\text{U}\,(0.393,\ 0.433)$ for common and rare variants, respectively. The first

two simulation scenarios were constructed without any genetic main effects: (1) two common and eight rare variants, all exhibiting positive $G \times E$ effects, i.e., all $\gamma_j > 0$; (2) two common and eight rare variants with 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$. The remaining two simulation scenarios incorporate two common and eight rare variants with genetic main effects alongside the $G \times E$ effects: (3) β_j was specified same as in scenario (2) in the type I error assessment, two common and eight rare variants with positive $G \times E$ effects; (4) β_j was specified same as in scenario (2) in the type I error assessment, two common and eight rare variants with 50% of $\gamma_j > 0$ and 50% of $\gamma_j < 0$. To assess power, 200 simulated datasets were analyzed for each scenario, allowing for a comparison across different simulation conditions.

Table 3.5: The 4 simulation scenarios for power comparison

Scenario	**SNP main effects**				***G* × *E* effects**			
	Common SNPs		**Rare SNPs**		**Common SNPs**		**Rare SNPs**	
	# +	# -	# +	# -	# +	# -	# +	# -
(1)	0	0	0	0	2	0	8	0
(2)	0	0	0	0	1	1	4	4
(3)	2	0	8	0	2	0	8	0
(4)	1	1	4	4	1	1	4	4

3.2.2 Simulation results

The empirical type I error rate was calculated at the nominal levels of $\alpha = 0.05$ and $\alpha = 0.01$, utilizing 5000 replicates across two distinct simulation scenarios. As detailed in Table 3.6, the method denoted as GEITGMM_full calculates the p-value by determining the proportion of null statistics that are equal to or exceed the observed statistic. In contrast, the GEITGMM_half method employs an alternative p-value calculation strategy, wherein a null statistic precisely matching the observed statistic is counted as half. In Scenario 1, the GEITGMM method has the type I error rate under control. However, in Scenario 2, there is an observed inflation in the type I error rate, indicating a

potential issue with the method's performance under certain conditions.

Table 3.6: Empirical type I error of GEITGMM based on 5000 replicates

		GEITGMM_full		**GEITGMM_half**	
Test	**Level**	**Scenario 1**	**Scenario 2**	**Scenario 1**	**Scenario 2**
GEITGMM	0.05	0.046	**0.088**	0.041	**0.076**
	0.01	0.007	**0.019**	0.003	**0.015**

The 95% confidence interval of a nominal level α was calculated as $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/5000}$. Specifically, the 95% confidence intervals are $(0.044,\ 0.056)$ for $\alpha = 0.05$, and $(0.007,\ 0.013)$ for $\alpha = 0.01$. Rates outside of the 95% confidence interval are in bold.

Empirical power was assessed at a significant level of 0.01, utilizing 200 simulation replicates. The power outcomes for various methods: GEITGMM_full, GEITGMM_half, MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi and ADABF across four distinct simulation scenarios are illustrated in Figures 3.3. Notably, GEITGMM (both GEITGMM_full and GEITGMM_half versions) consistently exhibited superior power across all simulation scenarios. This robust performance potentially attributable to the strong consistency of the GMM estimator [47]. Particularly, in Scenarios 3 and 4, which included genetic main effects, both GEITGMM_full and GEITGMM_half outperformed the alternatives, however, this advantage may come at the cost of an inflated type I error rate, hinting at a trade-off between power and Type I error rate control that warrants further scrutiny. The performance trends for the other methods mirrored those discussed in Section 3.1. MAGEIT_RAN exhibits higher power than other methods compared after GEITGMM. In contrast, aMiSTi's performance was notably weaker in scenarios where the $G \times E$ have opposite directions, which is likely because aMiSTi combines burden and variance component tests, leading to less power in genomic regions harboring both protective and detrimental variants [69].

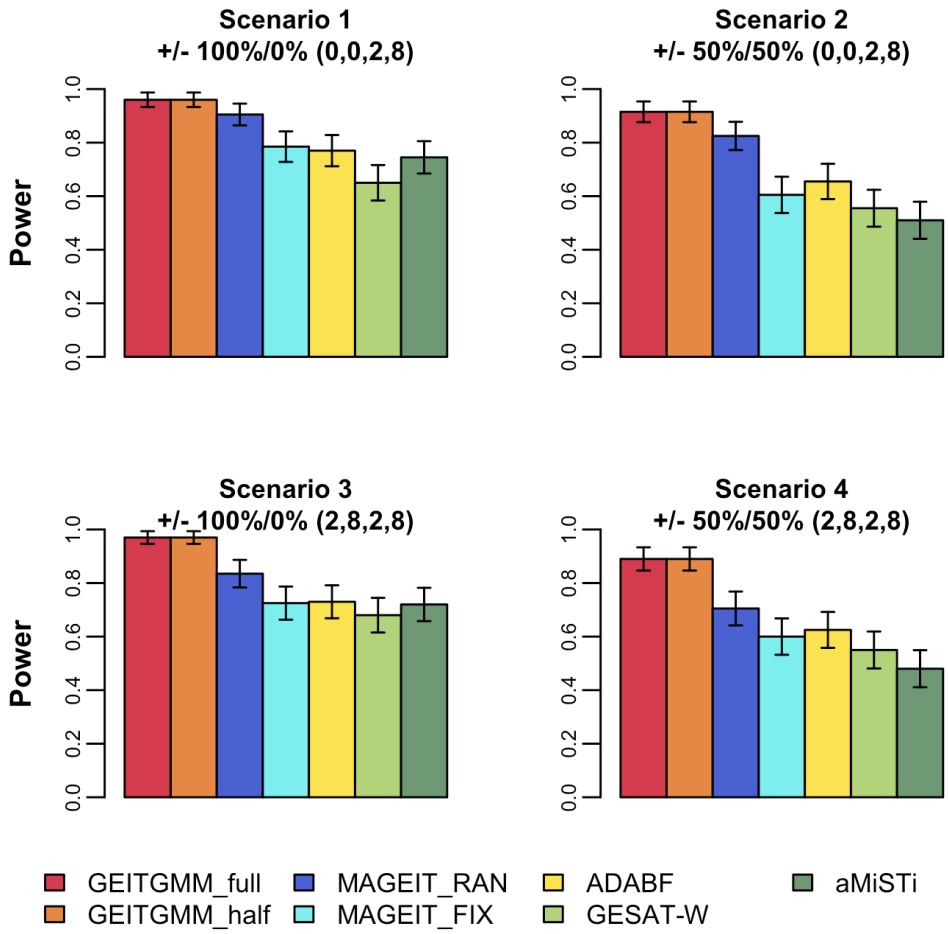

Figure 3.3: Empirical power of GEITGMM_full, GEITGMM_half, MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi and ADABF. Error bars show the approximated 95% confidence interval for the empirical power β, which is calculated as $\beta \pm 1.96\sqrt{\beta(1-\beta)/200}$.

CHAPTER 4
APPLICATION TO MESA DATA

To demonstrate the utility of our proposed methods MAGEIT_RAN and MAGEIT_FIX, we performed a genome-wide analysis of gene-alcohol interaction on hypertension in MESA [70]. MESA is a large longitudinal study of subclinical cardiovascular diseases including more than 6,800 participants. We analyzed the hypertension outcome measured at the first physical examination of 6,403 participants, consisting of 2,851 subjects with hypertension and 3,552 subjects without hypertension. The participants cover a diverse group of subjects including white (39.3%), African American (26.1%), Hispanic (22.5%), and Asian (12.1%). Alcohol usage (consumption of alcoholic beverages currently or formerly) was treated as an environmental variable, with 6,379 responses including 5,058 YESs and 1,321 NOs.

Samples were genotyped using the Affymetrix Human SNP Array 6.0. After data cleaning, the genotypes are then pre-phased using SHAPEIT [71], which estimates "best-guess" haplotypes by efficiently inferring the combination of alleles inherited together. These estimated haplotypes are subsequently imputed with IMPUTE2 [72], leveraging the comprehensive 1000 Genomes Project Phase 3 as a reference panel to infer missing genotype data accurately. We excluded subjects whose proportion of successfully imputed variants $< 5\%$ or empirical inbreeding coefficients > 0.05, resulting in 6,424 subjects for further analysis. The following quality-control criteria were applied: (1) call rate $> 95\%$, (2) MAF $> 0.5\%$, and (3) Hardy-Weinberg χ^2 statistic p-value $> 10^{-6}$, resulting in a final set of 8,540,864 variants. In the gene-based $G \times E$ analysis, we restricted analysis on protein-coding regions based on the reference genome GRCh37 [73]. In total, there were 18,977

genes on the 22 chromosomes and the number of variants in each gene region ranges from 2 to 5000, with a medium number of 383. Upon integrating the hypertension, alcohol usage and genotype data, a final set of 6,375 individuals are retained for downstream analyses.

4.1 Analysis of $G \times E$ effects

We performed genome-wide tests of gene-alcohol interaction effects on hypertension using all five methods, MAGEIT_RAN, MAGEIT_FIX, GESAT-W, aMiSTi, and ADABF. In the analysis, alongside age at the first exam and sex, we included the top ten principal components (PCs) of the genetic relationship matrix to account for population structure. The top ten PCs were calculated using the LD pruned variants with MAF > 0.05 to control for population structure.

MAGEIT_RAN and aMiSTi showed no evidence of inflation, with the genomic control inflation factors of 0.966 and 0.997, respectively. The $G \times E$ test assuming fixed genetic main effects, MAGEIT_FIX, and the Bayes factor-based test, ADABF, were conservative, with the genomic control inflation factors of 0.822 and 0.826, respectively. The genomic control inflation factor was 1.403 for GESAT-W. Therefore, we further adjusted the results of GESAT-W using genomic control.

No genes reached genome-wide significance at the p-value threshold of $\frac{0.05}{18{,}977} = 2.63 \times 10^{-6}$, commonly-used in gene-based analyses [74]. Table 4.1 lists the top genes for which at least one of the five tests gives a p-value $< 10^{-4}$. The gene *CCNDBP1* had the smallest p-value, detected by MAGEIT_RAN (p-value $= 2.80 \times 10^{-5}$) at a significance level of $\frac{1}{18{,}977} = 5.27 \times 10^{-5}$, a suggestive significance threshold in genome-wide scan [75]. The p-value of *EPB42* (p-value $= 5.98 \times 10^{-5}$) is close to the suggestive significance threshold, generated by MAGEIT_RAN. Both *CCNDBP1* and

EPB42 are located at 15q15.2. The cytogenetic region 15q15 has previously been reported to be associated with blood pressure [76]. Moreover, *EPB42* was shown to be significantly down-regulated in heavy drinkers after exposed to psychological stress [77, 78, 79].

Table 4.1: Genes with p-value $< 10^{-4}$ in at least one of the tests in the MESA data

Chr	Gene	# SNP	Region	MAGEIT_RAN	MAGEIT_FIX	GESAT-W	aMiSTi	ADABF
15	*CCNDBP1*	237	15q15.2	$\mathbf{2.80 \times 10^{-5}}$	2.03×10^{-3}	6.28×10^{-3}	3.32×10^{-2}	4.90×10^{-2}
	EPB42	269	15q15.2	$\mathbf{5.98 \times 10^{-5}}$	2.05×10^{-3}	1.12×10^{-2}	3.97×10^{-2}	1.00×10^{-1}

The smallest p-values among the five tests at the given genes are in bold.

4.2 Pathway analysis

Functional pathway analysis was conducted on genes that had $G \times E$ to identify enriched pathways related to hypertension, using MetaCore^{TM}. The top genes for which at least one of the five tests had a p-value < 0.001 were selected. Fisher's exact test was used to determine whether the gene list was enriched for a functional pathway. At the false discovery rate (FDR) < 0.01, there are two significant pathways that were reported to be related to hypertension (Table 4.2).The first pathway is related to ERK1/2 signaling (p-value $= 1.72 \times 10^{-5}$, FDR $= 2.38 \times 10^{-3}$). ERK1/2 is instrumental in transmitting signals from surface receptor to the nucleus. Once activated, it induces cell proliferation, differentiation, and other processes [80]. It has reported that the ERK1/2-RSK-nNOS might be crucial in the regulation of central blood pressure influenced by Ang II [81]. The second pathway is a signal transduction pathway related to Adenosine A1 receptor signaling (p-value $= 1.27 \times 10^{-4}$, FDR $= 7.83 \times 10^{-3}$). Adenosine modulates cardiovascular function and produces bradycardia and hypotension when mediated systematically [82, 83]. Activation of adenosine A1 receptor causes contraction of vascular smooth muscle and the adenosine A1 receptor agonists produce decreases in blood pressure and heart rate [84]. It has been observed that raised adenosine levels mediate

the ataxic and sedative/hypnotic effects of ethanol through activation of A1 receptors in the cerebellum, striatum, and cerebral cortex [85]. A1 agonists have been shown to decrease anxiety-like behavior, tremor, and seizures during acute ethanol withdrawal in mice [86].

Table 4.2: Pathways with FDR < 0.01 in the MESA data

Pathway	**P-value**	**FDR**
Signal transduction_ERK1/2 signaling pathway	1.72×10^{-5}	2.38×10^{-3}
Signal transduction_Adenosine A1 receptor signaling	1.27×10^{-4}	7.83×10^{-3}

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

5.1 Summary and discussion

Human complex diseases are influenced by a combination of genetic variation and interactions between genes and environmental factors. Many genes associated with diseases have been successfully identified. Therefore, the identification and comprehension of gene-environment interactions have become essential in predicting disease risk [87]. In this research, we have developed innovative statistical approaches to analyze gene-environment interactions, which play an important role in unraveling the complexities of human diseases and addressing the "missing heritability" challenge inherent in GWAS. Our contributions include development of three gene-environment interaction tests: MAGEIT_RAN and MAGEIT_FIX, which are built upon the MQS method, and GEITGMM, which is based on the GMM. These approaches are designed to capture the interplay between genes and environmental factors, accommodating both common and rare variants within gene sets, thereby providing new insights into the nature of gene-environment interactions.

In Chapter 1, we conduct a systematic review of the limitations and challenges encountered by traditional GWAS. We introduce the concept of $G \times E$ as a promising avenue to solve the "missing heritability" that GWAS has not yet fully explained. This chapter emphasizes the necessity of integrating $G \times E$ into genetic research, highlighting their potential to reveal the intricate relationships between genetic variants and environmental factors. Through literature review, we discuss the current research landscape of gene-environment interactions and statistical methods based on MoM and GMM. At the end of this chapter, we provid the outline of the dissertation.

In Chapter 2, we develop two groups of tests to detect interactions between an environmental factor and a set of genetic markers containing both rare and common variants. The first group of test include two tests and are based on the variance component method MinQue for Summary statistics (MQS) [46], which has been applied in MAPIT (Marginal ePIstasis Test) [88] and LT-MAPIT (liability threshold marginal epistasis test) [58] to detect gene-gene interactions. The advantage of MQS lies in its ability to provide unbiased and statistically efficient estimate using the method of moments and the minimal norm quadratic unbiased estimation criterion. We name these two tests as MArginal Gene-Environment Interaction Test with RANdom or FIXed genetic effects (MAGEIT_RAN or MAGEIT_FIX), where the genetic main effects are modeled as random or fixed, respectively. Both tests can be applied to continuous and binary phenotypes. Compared to existing methods, both MAGEIT_RAN and MAGEIT_FIX not only incorporate common and rare variants within a genetic region but also differentiate their effects during model fitting. Moreover, they produce unbiased estimators for the variance components. Our methods not only implement the MQS estimation to identify gene-environment interaction but also extend this approach by modeling genetic main effects as random in MAGEIT_RAN. Given that variants within a genomic region can have protective or detrimental effects and their effect sizes may differ, modeling genetic effects as random, such as in MAGEIT_RAN, enables the consideration of varying directions and magnitudes of the genetic effects. The second group of test is based on the Generalized Method of Moments (GMM) approach [41], leading to the Gene-Environment Interaction Test based on GMM (GEITGMM).

In Chapter 3, we evaluate the performance of these tests in detecting $G \times E$ for a set of genetic variants. We compare the performance of MAGEIT_RAN, MAGEIT_FIX, and three set-based $G \times E$

tests through simulation studies. Our findings demonstrate that MAGEIT_FIX maintains a well-controlled type I error rate, whereas MAGEIT_RAN is slightly conservative, especially for binary phenotypes, due to the approximations applied when dealing with binary phenotypes. However, across certain simulation settings, MAGEIT_RAN has higher statistical power than other compared methods. We also conduct simulation studies to check the type I error rate of GEITGMM and compare its power with some existing $G \times E$ tests. The simulation results show its ability to achieve higher power in detecting $G \times E$ in certain scenarios.

In Chapter 4, we conduct the genome-wide analysis investigating gene-alcohol interactions on hypertension in the MESA dataset and no genes reach genome-wide significance. However, employing a suggestive significance threshold in the genome-wide scan [75], we identify two genes, *CCNDBP1* and *EPB42*, located at the cytogenetic region 15q15.2, a region previously reported to be associated with blood pressure. The *EPB42* gene expression was found to be significantly downregulated in heavy drinkers following exposure to psychological stress. Furthermore, we identify two signal transduction pathways associated with hypertension, with one of them related to hypertension and alcohol usage. Considering the established roles of the identified genes and pathways, our findings suggest that MAGEIT effectively identifies biologically relevant genes that interact with environmental factors to influence complex traits.

5.2 Limitations and future work

This dissertation underscores the critical role of $G \times E$ in the complexity of disease and addresses the challenges inherent in their analysis. We introduce innovative methods based on moment estimation MAGEIT_RAN, MAGEIT_FIX, and GEITGMM, which provide tools for dissecting the

interplay between genes and environmental factors. While these introduced methods mark advancements in the analysis of $G \times E$, they are accompanied by certain limitations. Specifically, while moment-based estimators expedite the estimation process compared to likelihood-based methods, they present challenges during the inference step. MAGEIT_RAN and MAGEIT_FIX require eigen value decomposition of large matrices, which can be computationally intensive, and GEITGMM relies on permutation tests necessitating extensive resampling, which slows down the computation. Notably, existing literature on the asymptotic properties of GMM estimators [41, 89] suggests a potential way for statistical inference. Moreover, in MAGEIT_RAN, the regression coefficients of the genetic main effects, β_j, and the interaction effects, γ_j, are assumed to be independent. However, in reality, these effects may be correlated in a genomic region. This restricted assumption might result in a loss of statistical power, especially in scenarios where most variants in a gene interact with the environmental factor, and their interaction effects are in the same direction. Considering the complexities in linkage disequilibrium and haplotype effects, it becomes more appropriate to account for potential correlations among these coefficients. Also, a notable limitation of GEITGMM is the observed inflation of type I error rates in specific analytical scenarios when using GEITGMM. This challenge underscores the imperative for ongoing refinement to ensure the method's robustness and precision. Furthermore, our model lacks consideration for sparsity, which is particularly noticeable when dealing with long genes.

In our future work, we will focus on optimizing the computational efficiency of MAGEIT_RAN and MAGEIT_FIX. Additionally, efforts will be made to integrate asymptotic distributions of the GMM estimator and to account for variant correlations within genomic regions, aiming to enhance the power of these methods. Furthermore, we intend to employ variable selection techniques to more effectively manage scenarios involving long genes. We will also explore incorporating the

correlation coefficient of the interaction effects between genetic main effects, β_j, and interaction effects, γ_j, which have previously been assumed to be independent, to better capture the realities of genomic complexities where these effects are often correlated. By addressing these challenges, this dissertation sets the stage for further research that moves us closer to fully understanding the intricate relationship between genetics and environment in human health.

5.3 Code availability

R code implementing MAGEIT_RAN and MAGEIT_FIX can be found at https://github.com/ZWang-Lab/MAGEIT2. R code implementing GEITGMM can be found at https://github.com/slcxding/GEITGMM.

BIBLIOGRAPHY

[1] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma, "Genome-wide association studies," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 59, 2021.

[2] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of gwas discovery: biology, function, and translation," *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.

[3] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nature Reviews Genetics*, vol. 20, no. 8, pp. 467–484, 2019.

[4] T. A. Manolio, "Bringing genome-wide association findings into clinical use," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 549–558, 2013.

[5] A. I. Young, "Discovering missing heritability in whole-genome sequencing data," *Nature Genetics*, vol. 54, no. 3, pp. 224–226, 2022.

[6] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[7] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, "Common snps explain a large proportion of the heritability for human height," *Nature genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[8] L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, P. M. Visscher, *et al.*, "Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry," *Human molecular genetics*, vol. 27, no. 20, pp. 3641–3649, 2018.

[9] G. Pare, S. Asma, and W. Q. Deng, "Contribution of large region joint associations to complex traits genetics," *PLoS genetics*, vol. 11, no. 4, p. e1005103, 2015.

[10] M. Stephens and D. J. Balding, "Bayesian statistical methods for genetic association studies," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 681–690, 2009.

[11] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. González-Recio, I. R. König, H. Zhang, and Y. V. Sun, "Machine learning in genome-wide association studies," *Genetic epidemiology*, vol. 33, no. S1, pp. S51–S57, 2009.

[12] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, "Human genetic variation and its contribution to complex traits," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 241–251, 2009.

[13] H. Aschard, J. Chen, M. C. Cornelis, L. B. Chibnik, E. W. Karlson, and P. Kraft, "Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases," *The American Journal of Human Genetics*, vol. 90, no. 6, pp. 962–972, 2012.

[14] M. Slatkin, "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future," *Nature Reviews Genetics*, vol. 9, no. 6, pp. 477–485, 2008.

[15] E. Cano-Gamez and G. Trynka, "From gwas to function: using functional genomics to identify the mechanisms underlying complex diseases," *Frontiers in genetics*, vol. 11, p. 424, 2020.

[16] K. Musunuru, A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt, K. V. Sachs, X. Li, H. Li, N. Kuperwasser, V. M. Ruda, *et al.*, "From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus," *Nature*, vol. 466, no. 7307, pp. 714–719, 2010.

[17] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, *et al.*, "Systematic localization of common disease-associated variation in regulatory dna," *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[18] X. Lin, S. Lee, D. C. Christiani, and X. Lin, "Test for interactions between a genetic marker set and environment in generalized linear models," *Biostatistics*, vol. 14, no. 4, pp. 667–681, 2013.

[19] A. Bhatnagar, "Environmental determinants of cardiovascular disease," *Circulation research*, vol. 121, no. 2, pp. 162–180, 2017.

[20] K. E. Cosselman, A. Navas-Acien, and J. D. Kaufman, "Environmental factors in cardiovascular disease," *Nature Reviews Cardiology*, vol. 12, no. 11, pp. 627–642, 2015.

[21] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature reviews genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[22] D. Thomas, "Gene–environment-wide association studies: emerging approaches," *Nature Reviews Genetics*, vol. 11, no. 4, pp. 259–272, 2010.

[23] P. Kraft, Y.-C. Yen, D. O. Stram, J. Morrison, and W. J. Gauderman, "Exploiting gene-environment interaction to detect genetic associations," *Human heredity*, vol. 63, no. 2, pp. 111–119, 2007.

[24] A. K. Manning, M. LaValley, C.-T. Liu, K. Rice, P. An, Y. Liu, I. Miljkovic, L. Rasmussen-Torvik, T. B. Harris, M. A. Province, *et al.*, "Meta-analysis of gene-environment interaction: joint estimation of snp and snp× environment regression coefficients," *Genetic epidemiology*, vol. 35, no. 1, pp. 11–18, 2011.

[25] H. Aschard, D. B. Hancock, S. J. London, and P. Kraft, "Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects," 2011.

[26] M. J. Khoury and S. Wacholder, "Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities," *American journal of epidemiology*, vol. 169, no. 2, pp. 227–230, 2009.

[27] W.-Y. Lin, C.-C. Huang, Y.-L. Liu, S.-J. Tsai, and P.-H. Kuo, "Genome-wide gene-environment interaction analysis using set-based association tests," *Frontiers in genetics*, vol. 9, p. 715, 2019.

[28] X. Lin, S. Lee, M. C. Wu, C. Wang, H. Chen, Z. Li, and X. Lin, "Test for rare variants by environment interactions in sequencing association studies," *Biometrics*, vol. 72, no. 1, pp. 156–164, 2016.

[29] H. Chen, J. B. Meigs, and J. Dupuis, "Incorporating gene-environment interaction in testing for association with rare genetic variants," *Human heredity*, vol. 78, no. 2, pp. 81–90, 2014.

[30] Y.-R. Su, C.-Z. Di, L. Hsu, Genetics, and E. of Colorectal Cancer Consortium, "A unified powerful set-based test for sequencing data analysis of gxe interactions," *Biostatistics*, vol. 18, no. 1, pp. 119–131, 2017.

[31] J.-Y. Tzeng, D. Zhang, M. Pongpanich, C. Smith, M. I. McCarthy, M. M. Sale, B. B. Worrall, F.-C. Hsu, D. C. Thomas, and P. F. Sullivan, "Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression," *The American Journal of Human Genetics*, vol. 89, no. 2, pp. 277–288, 2011.

[32] S. Jiao, L. Hsu, S. Bézieau, H. Brenner, A. T. Chan, J. Chang-Claude, L. Le Marchand, M. Lemire, P. A. Newcomb, M. L. Slattery, *et al.*, "Sberia: set-based gene-environment interaction test for rare and common variants in complex diseases," *Genetic epidemiology*, vol. 37, no. 5, pp. 452–464, 2013.

[33] X. Wang, E. Lim, C.-T. Liu, Y. J. Sung, D. C. Rao, A. C. Morrison, E. Boerwinkle, A. K. Manning, and H. Chen, "Efficient gene–environment interaction tests for large biobank-scale sequencing studies," *Genetic epidemiology*, vol. 44, no. 8, pp. 908–923, 2020.

[34] J. T. Chi, I. C. Ipsen, T.-H. Hsiao, C.-H. Lin, L.-S. Wang, W.-P. Lee, T.-P. Lu, and J.-Y. Tzeng, "Seagle: A scalable exact algorithm for large-scale set-based gene-environment interaction tests in biobank data," *Frontiers in genetics*, vol. 12, p. 710055, 2021.

[35] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.

[36] S. Lee, M. C. Wu, and X. Lin, "Optimal tests for rare variant effects in sequencing association studies," *Biostatistics*, vol. 13, no. 4, pp. 762–775, 2012.

[37] S. A. Santorico and A. E. Hendricks, "Progress in methods for rare variant association," in *BMC genetics*, vol. 17, pp. 57–66, Springer, 2016.

[38] W. Chen, B. J. Coombes, and N. B. Larson, "Recent advances and challenges of rare variant association analysis in the biobank sequencing era," *Frontiers in Genetics*, vol. 13, p. 1014947, 2022.

[39] C. R. Rao, "Estimation of heteroscedastic variances in linear models," *Journal of the American Statistical Association*, vol. 65, no. 329, pp. 161–172, 1970.

[40] C. R. Rao, "Estimation of variance and covariance components—minque theory," *Journal of multivariate analysis*, vol. 1, no. 3, pp. 257–275, 1971.

[41] X. Wang and Y. Wen, "A penalized linear mixed model with generalized method of moments estimators for complex phenotype prediction," *Bioinformatics*, vol. 38, no. 23, pp. 5222–5228, 2022.

[42] Y. Zhang, Y. Cheng, W. Jiang, Y. Ye, Q. Lu, and H. Zhao, "Comparison of methods for estimating genetic correlation between complex traits using gwas summary statistics," *Briefings in bioinformatics*, vol. 22, no. 5, p. bbaa442, 2021.

[43] J. Yu, "Simulation-based estimation methods for financial time series models," in *Handbook of Computational Finance*, pp. 401–435, Springer, 2011.

[44] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, R. Consortium, P. G. Consortium, G. C. for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, L. Duncan, *et al.*, "An atlas of genetic correlations across human diseases and traits," *Nature genetics*, vol. 47, no. 11, pp. 1236–1241, 2015.

[45] Q. Lu, B. Li, D. Ou, M. Erlendsdottir, R. L. Powles, T. Jiang, Y. Hu, D. Chang, C. Jin, W. Dai, *et al.*, "A powerful approach to estimating annotation-stratified genetic covariance via gwas summary statistics," *The American Journal of Human Genetics*, vol. 101, no. 6, pp. 939–964, 2017.

[46] X. Zhou, "A unified framework for variance component estimation with summary statistics in genome-wide association studies," *The annals of applied statistics*, vol. 11, no. 4, p. 2027, 2017.

[47] L. P. Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the econometric society*, pp. 1029–1054, 1982.

[48] A. Hall, "Generalized method of moments advanced texts in econometrics (oxford: University press)," 2005.

[49] X. Wang and Y. Wen, "A penalized linear mixed model with generalized method of moments for prediction analysis on high-dimensional multi-omics data," *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac193, 2022.

[50] P. Turley, R. K. Walters, O. Maghzian, A. Okbay, J. J. Lee, M. A. Fontana, T. A. Nguyen-Viet, R. Wedow, M. Zacher, N. A. Furlotte, *et al.*, "Multi-trait analysis of genome-wide association summary statistics using mtag," *Nature genetics*, vol. 50, no. 2, pp. 229–237, 2018.

[51] P. Kundu and N. Chatterjee, "Logistic regression analysis of two-phase studies using generalized method of moments," *Biometrics*, vol. 79, no. 1, pp. 241–252, 2023.

[52] R. B. Davies, "The distribution of a linear combination of $\chi 2$ random variables," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 29, no. 3, pp. 323–333, 1980.

[53] H. Liu, Y. Tang, and H. H. Zhang, "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables," *Computational Statistics & Data Analysis*, vol. 53, no. 4, pp. 853–856, 2009.

[54] R. Tempelman and D. Gianola, "Marginal maximum likelihood estimation of variance components in poisson mixed models using laplacian integration," *Genetics Selection Evolution*, vol. 25, no. 4, pp. 305–319, 1993.

[55] B. Engel, W. Buist, and A. Visscher, "Inference for threshold models with variance components from the generalized linear mixed model perspective," *Genetics Selection Evolution*, vol. 27, no. 1, pp. 15–32, 1995.

[56] C. K. Williams and D. Barber, "Bayesian classification with gaussian processes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[57] M. Kuss, C. E. Rasmussen, and R. Herbrich, "Assessing approximate inference for binary gaussian process classification.," *Journal of machine learning research*, vol. 6, no. 10, 2005.

[58] L. Crawford and X. Zhou, "Genome-wide marginal epistatic association mapping in case-control studies," *bioRxiv*, p. 374983, 2018.

[59] S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher, "test for interaction heritability for disease from genome-wide association studies," *The American Journal of Human Genetics*, vol. 88, no. 3, pp. 294–305, 2011.

[60] R. Drikvandi, A. Khodadadi, and G. Verbeke, "Testing variance components in balanced linear growth curve models," *Journal of Applied Statistics*, vol. 39, no. 3, pp. 563–572, 2012.

[61] R. Drikvandi, G. Verbeke, A. Khodadadi, and V. Partovi Nia, "Testing multiple variance components in linear mixed-effects models," *Biostatistics*, vol. 14, no. 1, pp. 144–159, 2013.

[62] L. S. Chen, L. Hsu, E. R. Gamazon, N. J. Cox, and D. L. Nicolae, "An exponential combination procedure for set-based association tests in sequencing studies," *The American Journal of Human Genetics*, vol. 91, no. 6, pp. 977–986, 2012.

[63] Q. Liu, L. S. Chen, D. L. Nicolae, and B. L. Pierce, "A unified set-based test with adaptive filtering for gene–environment interaction analyses," *Biometrics*, vol. 72, no. 2, pp. 629–638, 2016.

[64] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler, "Calibrating a coalescent simulation of human genome sequence variation," *Genome research*, vol. 15, no. 11, pp. 1576–1583, 2005.

[65] I. Shlyakhter, P. C. Sabeti, and S. F. Schaffner, "Cosi2: an efficient simulator of exact and approximate coalescent with selection," *Bioinformatics*, vol. 30, no. 23, pp. 3427–3429, 2014.

[66] I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin, "Sequence kernel association tests for the combined effect of rare and common variants," *The American Journal of Human Genetics*, vol. 92, no. 6, pp. 841–853, 2013.

[67] B. E. Madsen and S. R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS genetics*, vol. 5, no. 2, p. e1000384, 2009.

[68] R. Schweiger, O. Weissbrod, E. Rahmani, M. Müller-Nurasyid, S. Kunze, C. Gieger, M. Waldenberger, S. Rosset, and E. Halperin, "Rl-skat: an exact and efficient score test for heritability and set tests," *Genetics*, vol. 207, no. 4, pp. 1275–1283, 2017.

[69] S. Basu and W. Pan, "Comparison of statistical tests for disease association with rare variants," *Genetic epidemiology*, vol. 35, no. 7, pp. 606–619, 2011.

[70] D. E. Bild, D. A. Bluemke, G. L. Burke, R. Detrano, A. V. Diez Roux, A. R. Folsom, P. Greenland, D. R. JacobsJr, R. Kronmal, K. Liu, *et al.*, "Multi-ethnic study of atherosclerosis: objectives and design," *American journal of epidemiology*, vol. 156, no. 9, pp. 871–881, 2002.

[71] O. Delaneau, J. Marchini, and J.-F. Zagury, "A linear complexity phasing method for thousands of genomes," *Nature methods*, vol. 9, no. 2, pp. 179–181, 2012.

[72] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS genetics*, vol. 5, no. 6, p. e1000529, 2009.

[73] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, *et al.*, "Gencode reference annotation for the human and mouse genomes," *Nucleic acids research*, vol. 47, no. D1, pp. D766–D773, 2019.

[74] M. P. Epstein, R. Duncan, E. B. Ware, M. A. Jhun, L. F. Bielak, W. Zhao, J. A. Smith, P. A. Peyser, S. L. Kardia, and G. A. Satten, "A statistical approach for rare-variant association testing in affected sibships," *The American Journal of Human Genetics*, vol. 96, no. 4, pp. 543–554, 2015.

[75] E. Lander and L. Kruglyak, "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results," *Nature genetics*, vol. 11, no. 3, pp. 241–247, 1995.

[76] A. T. Kraja, S. C. Hunt, J. S. Pankow, R. H. Myers, G. Heiss, C. E. Lewis, D. Rao, and M. A. Province, "Quantitative trait loci for metabolic syndrome in the hypertension genetic epidemiology network study," *Obesity research*, vol. 13, no. 11, pp. 1885–1890, 2005.

[77] X. Ma, Z. Liao, R. Li, W. Xia, H. Guo, J. Luo, H. Sheng, M. Tian, and Z. Cao, "Myocardial injury caused by chronic alcohol exposure—a pilot study based on proteomics," *Molecules*, vol. 27, no. 13, p. 4284, 2022.

[78] J. Chen, C. Yang, Y. Yang, Q. Liang, K. Xie, J. Liu, and Y. Tang, "Targeting dkk1 prevents development of alcohol-induced osteonecrosis of the femoral head in rats," *American Journal of Translational Research*, vol. 13, no. 4, p. 2320, 2021.

[79] R. D. Beech, J. J. Leffert, A. Lin, K. A. Hong, J. Hansen, S. Umlauf, S. Mane, H. Zhao, and R. Sinha, "Stress-related alcohol consumption in heavy drinkers correlates with expression of mi r-10a, mi r-21, and components of the tar-rna-binding protein-associated complex," *Alcoholism: Clinical and Experimental Research*, vol. 38, no. 11, pp. 2743–2753, 2014.

[80] I. Wortzel and R. Seger, "The erk cascade: distinct functions within various subcellular organelles," *Genes & cancer*, vol. 2, no. 3, pp. 195–209, 2011.

[81] W.-H. Cheng, P.-J. Lu, W.-Y. Ho, C.-S. Tung, P.-W. Cheng, M. Hsiao, and C.-J. Tseng, "Angiotensin ii inhibits neuronal nitric oxide synthase activation through the erk1/2-rsk signaling pathway to modulate central control of blood pressure," *Circulation research*, vol. 106, no. 4, pp. 788–795, 2010.

[82] R. A. Barraco, D. R. Marcantonio, J. W. Phillis, and W. R. Campbell, "The effects of parenteral injections of adenosine and its analogs on blood pressure and heart rate in the rat.," *General Pharmacology*, vol. 18, no. 4, pp. 405–416, 1987.

[83] G. Evoniuk, R. W. von Borstel, and R. J. Wurtman, "Antagonism of the cardiovascular effects of adenosine by caffeine or 8-(p-sulfophenyl) theophylline.," *Journal of Pharmacology and Experimental Therapeutics*, vol. 240, no. 2, pp. 428–432, 1987.

[84] C. W. Schindler, M. Karcz-Kubicha, E. B. Thorndike, C. E. Müller, S. R. Tella, S. Ferré, and S. R. Goldberg, "Role of central and peripheral adenosine receptors in the cardiovascular responses to intraperitoneal injections of adenosine a1 and a2a subtype receptor agonists," *British journal of pharmacology*, vol. 144, no. 5, pp. 642–650, 2005.

[85] C. L Ruby, C. A Adams, E. J Knight, H. Wook Nam, and D.-S. Choi, "An essential role for adenosine signaling in alcohol abuse," *Current drug abuse reviews*, vol. 3, no. 3, pp. 163–174, 2010.

[86] G. B. Kaplan, N. H. Bharmal, K. A. Leite-Morris, and W. R. Adams, "Role of adenosine a1 and a2a receptors in the alcohol withdrawal syndrome," *Alcohol*, vol. 19, no. 2, pp. 157–162, 1999.

[87] D. J. Hunter, "Gene–environment interactions in human diseases," *Nature reviews genetics*, vol. 6, no. 4, pp. 287–298, 2005.

[88] L. Crawford, P. Zeng, S. Mukherjee, and X. Zhou, "Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits," *PLoS genetics*, vol. 13, no. 7, p. e1006869, 2017.

[89] K. Hayakawa, "The asymptotic properties of the system gmm estimator in dynamic panel data models when both n and t are large," *Econometric Theory*, vol. 31, no. 3, pp. 647–667, 2015.

CURRICULUM VITAE

Linchuan Shen

slcxding@gmail.com

Degrees:

Master of Science, Statistics, 2017

University of Science and Technology of China, Hefei, China

Bachelor of Economics, Financial Engineering, 2014

Sichuan University, Chengdu, China

Honors and Awards:

- Mathematical Sciences Summer Fellowship Summer 2023
- Summer Doctoral Research Fellowship Summer 2022

Publications:

- Detection of interactions between genetic marker sets and environment in a genome-wide study of hypertension, **Linchuan Shen**, Amei Amei, Bowen Liu, Yunqing Liu, Gang Xu, Edwin C. Oh, Zuoheng Wang, 2023 (Submitted).
- Retrospective varying coefficient association analysis of longitudinal binary traits: Application to the identification of genetic loci associated with hypertension, Gang Xu, Amei Amei, Weimiao Wu, Yunqing Liu, **Linchuan Shen**, Edwin C. Oh, Zuoheng Wang, 2022 (Accepted).

Dissertation Title:

Identifying Disease-Related Gene-Environment Interactions Based on Method of Moments

Dissertation Examination Committee:

Chairperson, Amei Amei, Ph.D.

Committee Member, Malwane Ananda, Ph.D.

Committee Member, Farhad Shokoohi, Ph.D.

Graduate Faculty Representative, Mira Han, Ph.D.