12-1-2022

Retrospective Varying Coefficient Association Analysis of Longitudinal Binary Traits

Gang Xu

RETROSPECTIVE VARYING COEFFICIENT ASSOCIATION ANALYSIS OF LONGITUDINAL BINARY TRAITS

By

Gang Xu

Bachelor of Science - Statistics
University of Science and Technology of China
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
December 2022

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

November 9, 2022

This dissertation prepared by

Gang Xu

entitled

Retrospective Varying Coefficient Association Analysis of Longitudinal Binary Traits

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

Kaushik Ghosh, Ph.D.
Examination Committee Member

Edwin Oh, Ph.D.
Graduate College Faculty Representative

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

ABSTRACT

RETROSPECTIVE VARYING COEFFICIENT ASSOCIATION ANALYSIS OF LONGITUDINAL BINARY TRAITS

by

Gang Xu

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

Many genetic studies contain rich information on longitudinal phenotypes that require powerful analytical tools for optimal analysis. Genetic analysis of longitudinal data that incorporates temporal variation is important for understanding the genetic architecture and biological variation of complex diseases. Most of the existing methods assume that the contribution of genetic variants is constant over time and fails to capture the dynamic pattern of disease progression. However, the relative influence of genetic variants on complex traits fluctuates over time.

We developed several tests to fill the gap of analyzing time-varying genetic effects in longitudinal GWAS for binary traits. First, we propose a retrospective varying coefficient mixed model association test, RVMMAT, to detect time-varying genetic effect for common genetic variants. Second, we propose a group of retrospective variant set varying coefficient mixed model association tests, RSVMMATs, to detect time-varying effects of a set of rare genetic variants on a binary trait measured repeatedly over time. Through simulations, we illustrated that the retrospective varying-coefficient tests were robust to model misspecification under different ascertainment schemes and

gained power over the association methods assuming constant genetic effect. We applied RVMMAT and RSVMMATs to a genome-wide association analysis of longitudinal measure of hypertension in the Multi-Ethnic Study of Atherosclerosis (MESA). Our results demonstrated that the proposed methods could detect biologically relevant genetic variants and pathways in a genome-wide scan and provided insight into the genetic architecture of hypertension.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Amei Amei for her guidance, patience, and support. During the time I struggled to complete my reseach projects and dissertation, she has helped me not just academically, but also mentally. I have this invaluable chance to learn and grow in this process.

I really appreciate Dr. Zuoheng Wang for introducing me to genome-wide association studies for longitudinal data, as well as for her continuous supporting and guidance.

I want to thank Dr. Kaushik Ghosh, Dr. Malwane Ananda, and Dr. Edwin Oh for their kind instructions and inspirations. Many thanks to Chong Cheng for his assistance with supercomputing. Also thanks to my math department friends.

Finally, I want to thanks all my family members for supporting me and my girlfriend, Linchuan, for her care and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Genome wide association studies

Genome-wide association studies (GWASs) are designed for testing the association between genetic variants and complex traits. GWAS have successfully identified thousands of susceptible loci underlying human diseases and complex traits. In GWAS, single-nucleotide polymorphisms (SNPs) are commonly used as genetic variants [1]. SNPs can be categorized as common SNPs or rare SNPs depending on their minor allele frequencies (MAFs) being larger than 0.01 or not. A basic approach to analyze millions common SNPs is to test the association between a single genetic variant and a trait at a time, pioneered by [2]. Later, more than one thousand human GWASs for common SNPs was published [3].

With the development of high-throughput sequencing technologies, many high-throughput sequence data are generated, such as UK biobank. Those high-throughput sequencing data contain massive rare and low-frequency SNPs. However, most of the approaches for common SNPs test genetic variants one at a time and likely to be underpowered for rare SNPs. To identify rare variants that contribute to trait heritability, an increasing number of studies have considered testing for joint effect of multiple markers in a genomic region. The various gene-based association tests can be classified as burden tests [4], kernel tests [5, 6], and combination of the two methods. Burden tests collapse genetic variants in a genomic region into a single score and detect association between the score and an outcome. Such tests are powerful when most of the variants are causal with homogeneous effects in direction and magnitude. In contrast, kernel tests assume random effects for individual variants and tests the genotype-phenotype association via a variance com-

ponent score test. They are shown to be more powerful when the region contains both risk and protective variants or large proportion of the variants in the region is neutral. To avoid power loss in unknown underlying scenarios, several omnibus tests have been proposed to combine strength from both approaches, such as SKAT-O [7, 8], MONSTER [9], MiST [10], aSPU [11], SMMAT-E [12], and ACAT-O [13].

1.2 Statistical methods for longitudinal GWAS

Many epidemiological studies, such as Framingham Heart Study (FHS) and Women's Health Initiative (WHI), have collected and measured health conditions and phenotypic traits on study participants over the years. Such studies provide rich resources for the investigation of genetic architecture and biological variations of complex disorders. In recent years, more and more genetic studies have exploited genetic data from human biobanks and extracted health information from electronic health record (EHR) data to gain better understanding of the genetic mechanism over the course of complex diseases.

Traditional genetic association analyses on single time point measure fail to capture the phenotypic variation over time and may lose statistical power to identify disease-related variants. Thus, genetic studies with longitudinal phenotypes require powerful analytical tools for optimal analysis. Statistical methods that account for dependence structure among observations from the same subject have been developed in GWAS to make full use of longitudinal data, such as mixed effects models [14, 15, 16], generalized estimating equations (GEEs) [17, 16], growth mixture models [18, 19], and empirical Bayes models [20]. Most of these methods assume that genetic contribution is constant over time. However, disease development and progression is a complicated process that changes over time. Windows of susceptibility and critical periods across the lifespan exist in disease onset and development. Studies have shown that genetic influence on the trait variations fluctuates

with the passage of time [21, 22, 23, 24, 25]. Modeling time-varying genetic effects is essential to identify and validate causal genetic loci that are associated with time-dependent variation of disease progression.

Varying coefficient models are a class of generalized regression models in which the coefficients are allowed to vary smoothly with the value of other variables [26]. They are semi-parametric models that explore dynamic pattern in the data to improve model fitting [27], reduce model bias by specifying the coefficients as smooth nonparametric functions [28], and overcome the "curse of dimensionality" in the nonparametric estimation of multiple regression problems [29]. There are several approaches to estimate time-varying coefficients in varying coefficient models, including kernel-local polynomial smoothing [30, 31, 32, 33], polynomial spline [34, 35, 36], and smoothing spline [37, 26, 38]. Other statistical methods have been developed to model dependency in longitudinal data, such as GEE [39] and Gaussian copula [40]. Models allowing for time-varying covariate effects include semiparametric regression with GEE [41], and time-varying copula models [42, 43].

Varying coefficient models have been used for two types of applications in longitudinal GWAS. The first application focuses on feature selection for longitudinal outcomes with ultra high-dimensional predictors such as SNPs. As computational burden is a major concern when handling millions of SNPs simultaneously in a model, feature screening becomes an efficient solution to filter out unimportant SNPs. Feature screening in varying coefficient models has been developed based on conditional Pearson correlation [24], extended B-splines [44], modified weighted least squares estimation [22] and functional regression with group penalty [45] to retain important SNPs associated with continuous and binary traits [46, 47]. The second application focuses on the detection of time-varying effect of quantitative trait nucleotide, such as functional GWAS [18, 48, 49]. These methods fit the model at each SNP separately and use likelihood ratio tests to determine statistical significance, thus can be computationally intensive to analyze genome-wide SNPs, especially for

binary outcomes. For large GWAS, score tests are popular and computational efficient because they only fit the null model once for all SNPs. However, current score test-based methods commonly treat the effects of genetic variants as constant over time and are not able to capture the dynamic contribution to disease progression.

To handle rare variants for longitudinal GWAS, researchers have extended the burden and kernel tests to repeatedly measured outcomes through mixed effects models or generalized estimating equations [50, 51, 52, 53]. Generalized score type tests based on a longitudinal genetic random field model were proposed to test the association between a trait measured over time and a set of genetic variants [50]. Later, they used a GEE based inference with a perturbation method to obtain tests that are theoretically robust to misspecification of within-subject correlation [51]. For a better control of type I error rates, a practical strategy was proposed by combining multiple working correlation structure in their longitudinal SNP set/sequence kernel association test (LSKAT) [52]. To overcome the ascertainment bias and model misspecification, a group of retrospective association tests for longitudinal binary traits was proposed based on generalized linear mixed-effect model (GLMM) and GEE [53]. However, all these longitudinal set-based association tests assume time-invariant genetic effects and are incapable of capturing the contribution from the genetic variants to the dynamic pattern of disease progression.

1.3 Outline of the dissertation

To fill the gap of analyzing time-varying genetic effects in longitudinal GWAS for binary traits, we develop a retrospective varying coefficient mixed model association test (RVMMAT) for common SNPs and a group of retrospective variant set varying coefficient mixed model association tests (RSVMMATs) for rare SNPs. The dissertation is organized as follows:

In Chapter 1, we review genome-wide association studies as well as its development. Then we

review popular statistical models for longitudinal data and their application to longitudinal GWAS. We focus on the varying-coefficient models which is the model we apply to longitudinal GWAS in the dissertation.

In Chapter 2, we apply the generalized linear mixed effect model with varying coefficient to longitudinal binary traits and common SNPs. Then we develop a retrospective varying coefficient mixed model association test, RVMMAT, to detect time-varying genetic effect for common SNPs. For comparison, we also develop a prospective varying coefficient mixed model association test, VMMAT.

In chapter 3, we extend the RVMMAT to a group of retrospective variant set varying coefficient mixed model association tests, RSVMMATs, to detect time-varying effects of a set of genetic variants on a longitudinal binary trait for rare SNPs. As a comparison, we also develop a group of prospective variant set varying coefficient mixed model association tests, SVMMATs.

In Chapter 4, we evaluate our proposed methods and compare them with the existing association methods assuming constant genetic effect. We apply those tests to a genome-wide association analysis of longitudinal measure of hypertension in the Multi-Ethnic Study of Atherosclerosis (MESA) and identified hypertension-related genes and pathways.

In Chapter 5, we summarize our proposed methods in Chapters 2 and 3 and discuss future directions.

CHAPTER 2
RETROSPECTIVE VARYING COEFFICIENT MIXED MODEL ASSOCIATION TEST FOR COMMON SNPS

2.1 Introduction

Motivated by a genome-wide association analysis of longitudinal measure of hypertension in the Multi-Ethnic Study of Atherosclerosis (MESA), we develop a retrospective varying coefficient mixed model association test, RVMMAT, to detect time-varying genetic effect on longitudinal binary traits. We model dynamic genetic effect using smoothing splines, estimate model parameters by maximizing a double penalized quasi-likelihood function, design a joint test using a Cauchy combination method, and evaluate significance of the test via a retrospective approach in which genotypes are treated as random conditional on the phenotype and covariates. Retrospective association tests have been shown to be robust to the trait model misspecification and improve statistical power [54, 55, 16, 56]. In RVMMAT, flexible assumptions on the effect function increased power to detect genetic variants associated with dynamic traits. The validity of RVMMAT does not depend on the variance estimation in the trait model due to the tuning parameters in penalty terms. For comparison, we also develop VMMAT, a prospective varying coefficient mixed model association test.

2.2 GLMM with varying coefficients

Suppose a binary phenotype is measured repeatedly on a sample of n subjects. A set of covariates and genemo-wide genetic variants are also measured. Both static factors like gender and dynamic variables such as body weight are allowed as covariates. Let $\boldsymbol{X}_{ij}$ and Y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, denote the p-dimensional covariate vector and the binary trait measured on subject i at time t_{ij}.

Here, the measurement time and length are allowed to be different for different subjects. We let $\boldsymbol{X} = (\boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{1,m_1}, \ldots, \boldsymbol{X}_{n,1}, \ldots, \boldsymbol{X}_{n,m_n})^T$ denote the $N \times p$ covariate matrix and $\boldsymbol{Y} = (Y_{1,1}, \ldots, Y_{1,m_1}, \ldots, Y_{n,1}, \ldots, Y_{n,m_n})^T$ denote the outcome vector of length N, where $N = \sum_{i=1}^{n} m_i$ is the total number of observations. Let $\boldsymbol{G} = (G_1, \ldots, G_n)^T$ denote the genotype vector of the n subjects, where $G_i = 0$, 1 or 2, representing number of copies of minor allele of ith individual at the tested variant. We are interested in the problem of testing time-varying genetic effect between a genetic variant and the longitudinal binary trait adjusted with the effects of covariates.

We consider a generalized linear mixed model (GLMM) with varying coefficients, specified as

$$g(\mu_{ij}) = \gamma_0(t_{ij}) + G_i\gamma_1(t_{ij}) + \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + a_i + r_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, m_i, \tag{2.1}$$

where $\mu_{ij} = E(Y_{ij} \mid G_i, \boldsymbol{X}_{ij}, a_i, r_{ij})$, Y_{ij} is the phenotype measured at time t_{ij} for individual i, G_i is the genotype, $\boldsymbol{X}_{ij}$ is a vector of covariates, $\gamma_0(t)$ and $\gamma_1(t)$ are smooth nonparametric functions of time t representing a time-varying intercept and a time-varying genetic effect of the tested variant, $\boldsymbol{\beta}$ is a vector of effects of the covariates, and $g(\cdot)$ is the link function. For binary traits, we use the logit link function. The correlations among repeated measurements are captured by two random effects: a_i and r_{ij} which are the subject randon effect and the subject-specific time-dependent random effect [52, 16]. We assume that the a_i are independent and $a_i \sim N(0, \sigma_a^2)$. The vector $\boldsymbol{r}_i = (r_{i,1}, \ldots, r_{i,m_i})^T$ is assumed to follow a multivariate normal distribution, $\boldsymbol{r}_i \sim MVN(\boldsymbol{0}, \sigma_r^2\boldsymbol{R}_i)$, where the correlation matrix $\boldsymbol{R}_i$ is modeled by an AR(1) structure in which τ is the unknown parameter. Given the random effects a_i and r_{ij}, the response Y_{ij} are assumed to be independent. When both functions $\gamma_0(t)$ and $\gamma_1(t)$ are constants, model (2.1) reduces to a standard GLMM in [16].

Following [57, 38], we estimate $\gamma_0(t)$ and $\gamma_1(t)$ by maximizing the following double penalized

quasi-likelihood (DPQL) function:

$$\begin{aligned} l_{dp}\{\gamma_0(\cdot),\gamma_1(\cdot),\boldsymbol{\beta},\sigma_a^2,\sigma_r^2,\tau\} = &-\frac{1}{2}\sum_{i,j} D_{ij}\left(Y_{ij},\mu_{ij}\right) - \frac{1}{2\sigma_a^2}\boldsymbol{a}^T\boldsymbol{a} - \frac{1}{2\sigma_r^2}\boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r} \\ &- \frac{\lambda_0}{2}\int\left\{\gamma_0^{(h_0)}(t)\right\}^2 dt - \frac{\lambda_1}{2}\int\left\{\gamma_1^{(h_1)}(t)\right\}^2 dt, \end{aligned} \tag{2.2}$$

where $\boldsymbol{a} = (a_1,\ldots,a_n)^T$, $\boldsymbol{r} = (\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n)^T$ are the two vectors of random effects, $\boldsymbol{R} = \text{diag}\{\boldsymbol{R}_1,\ldots,\boldsymbol{R}_n\}$ is a block diagonal matrix, $D_{ij}(Y_{ij},\mu_{ij}) = -2\int_{Y_{ij}}^{\mu_{ij}} \frac{Y_{ij}-u}{u(1-u)}du$ is the conditional deviance function of binary outcome Y_{ij} given random effects a_i and r_{ij}, λ_k $(k=0,1)$ are tuning parameters that control the smoothness of $\gamma_k(t)$, and h_k are positive integers for the derivative order of $\gamma_k(t)$.

The maximizers for the nonparametric functions $\gamma_k(t)$ in the DPQL function of Eq. (2.2) are smoothing splines of order $2h_k$ [58]. Let $0 < t_1^0 < \cdots < t_m^0 < 1$ be the m distinct knots of t_{ij}. Then, the smoothing splines can be expressed as

$$\gamma_k(t) = \sum_{s=1}^{h_k} c_{ks}F_{ks}(t) + \sum_{l=1}^{m} d_{kl}V_k(t,t_l^0), \quad k=0,1,$$

where $F_{ks}(t)$ is a polynomial of order $s-1$ (e.g., $F_{ks}(t) = t^{(s-1)}/(s-1)!$, $s=1,\ldots,h_k$), and $V_k(t_1,t_2) = \frac{1}{[(h_k-1)!]^2}\int_0^1 (t_1-u)_+^{h_k-1}(t_2-u)_+^{h_k-1}du$ with $u_+ = \max\{u,0\}$. We denote $\boldsymbol{c}_k = (c_{k,1},\ldots,c_{k,h_k})^T$, $\boldsymbol{d}_k = (d_{k1},\ldots,d_{km})^T$, and $\boldsymbol{\gamma}_k = (\gamma_k(t_1^0),\ldots,\gamma_k(t_m^0))^T$ for $k=0,1$. Then $\boldsymbol{\gamma}_k$ can be expressed as

$$\boldsymbol{\gamma}_k = \boldsymbol{F}_k\boldsymbol{c}_k + \boldsymbol{V}_k\boldsymbol{d}_k,$$

where $\boldsymbol{F}_k$ is an $m \times h_k$ matrix with its (l,s)th entry equal to $F_{ks}(t_l^0)$, and $\boldsymbol{V}_k$ is a positive definite matrix with the (l,s)th entry equal to $V_k(t_l^0,t_s^0)$. Similar to Eq. (5) in [38], the DPQL function of Eq. (2.2) becomes

$$\begin{aligned} l_{dp}\{\gamma_0(\cdot),\gamma_1(\cdot),\boldsymbol{\beta},\sigma_a^2,\sigma_r^2,\tau\} = &-\frac{1}{2}\sum_{i,j} D_{ij}(Y_{ij},\mu_{ij}) - \frac{1}{2\sigma_a^2}\boldsymbol{a}^T\boldsymbol{a} - \frac{1}{2\sigma_r^2}\boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r} \\ &- \frac{\lambda_0}{2}\boldsymbol{d}_0^T\boldsymbol{V}_0\boldsymbol{d}_0 - \frac{\lambda_1}{2}\boldsymbol{d}_1^T\boldsymbol{V}_1\boldsymbol{d}_1. \end{aligned} \tag{2.3}$$

If we treat $\boldsymbol{d}_k$ in $\boldsymbol{\gamma}_k$ as random effects distributed as $\boldsymbol{d}_k \sim N(\boldsymbol{0},\theta_k\boldsymbol{V}_k^{-1})$ for $k=0,1$, where $\theta_k = \lambda_k^{-1}$, it follows that the maximizers of Eq. (2.3) can be obtained by fitting the GLMM

representation of model (2.1), expressed in a matrix form as

$$g(\boldsymbol{\mu}) = \boldsymbol{M}\boldsymbol{F}_0\boldsymbol{c}_0 + \boldsymbol{\Delta}_G\boldsymbol{M}\boldsymbol{F}_1\boldsymbol{c}_1 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{V}_0\boldsymbol{d}_0 + \boldsymbol{\Delta}_G\boldsymbol{M}\boldsymbol{V}_1\boldsymbol{d}_1 + \boldsymbol{B}\boldsymbol{a} + \boldsymbol{r}, \tag{2.4}$$

where $\boldsymbol{M}$ is an $N \times m$ incidence matrix mapping t_{ij} to the m distinct knots $t_1^0, \ldots, t_m^0$, $\boldsymbol{B}$ is an $N \times n$ design matrix mapping the subject-level genotype vector $\boldsymbol{G}$ to a measurement-level genotype vector $\boldsymbol{BG}$, with its (l, i)th entry $B_{li} = 1$ if the lth entry of $\boldsymbol{Y}$ is a measurement on subject i and 0 otherwise, and $\boldsymbol{\Delta}_G = \text{diag}\{\boldsymbol{BG}\} = \text{diag}\{G_1, \ldots, G_1, \ldots, G_n, \ldots, G_n\}$ is an N-dimensional diagonal matrix of the genotypes for the n subjects. Model (2.4) is a specific implementation of the GLMM with varying coefficients in [38]. Here, the tuning parameters λ_0 and λ_1 in the DPQL function of Eq. (2.2) are re-parameterized as θ_0 and θ_1, and treated as the unknown variance component parameters in model (2.4).

2.3 Varying coefficient mixed model association test

To test time-varying genetic effect between the variant and the trait, we test $H_0: \ \gamma_1(t) = 0$ in Model (2.1), which is equivalent to test $H_0: \ \boldsymbol{c}_1 = \boldsymbol{0}$ and $\theta_1 = 0$ in Model (2.4). The reduced GLMM under the null hypothesis specifies that

$$g(\boldsymbol{\mu}_0) = \boldsymbol{M}\boldsymbol{F}_0\boldsymbol{c}_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{V}_0\boldsymbol{d}_0 + \boldsymbol{B}\boldsymbol{a} + \boldsymbol{r}, \tag{2.5}$$

where $\boldsymbol{\mu}_0 = E(\boldsymbol{Y} \mid \boldsymbol{F}_0, \boldsymbol{X}, \boldsymbol{d}_0, \boldsymbol{a}, \boldsymbol{r})$.

If we test $H_0: \ \boldsymbol{c}_1 = \boldsymbol{0}$ under the assumption that $\theta_1 = 0$, a score test can be constructed as

$$T_f = (\boldsymbol{U}_0(\boldsymbol{c}_1))^T [\text{Var}(\boldsymbol{U}_0(\boldsymbol{c}_1))]^{-1} \boldsymbol{U}_0(\boldsymbol{c}_1), \tag{2.6}$$

where $\boldsymbol{U}_0(\boldsymbol{c}_1) = (\boldsymbol{\Delta}_G\boldsymbol{M}\boldsymbol{F}_1)^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)$ is the score function for $\boldsymbol{c}_1$ and $\widehat{\boldsymbol{\mu}}_0 = g^{-1}(\boldsymbol{M}\boldsymbol{F}_0\widehat{\boldsymbol{c}}_0 + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{M}\boldsymbol{V}_0\widehat{\boldsymbol{d}}_0 + \boldsymbol{B}\widehat{\boldsymbol{a}} + \widehat{\boldsymbol{r}})$ is a vector of fitted values under Model (2.5), which can be obtained using the penalized quasi-likelihood method [59]. For binary traits, we used a bias correction procedure

[60, 57, 38] to produce less biased variance component estimates. Given the genotype and covariates, the variance of the score $\boldsymbol{U}_0(\boldsymbol{c}_1)$ under H_0 is

$$\mathrm{Var}(\boldsymbol{U}_0(\boldsymbol{c}_1)) = (\boldsymbol{\Delta}_G \boldsymbol{M} \boldsymbol{F}_1)^T \boldsymbol{P} \boldsymbol{\Delta}_G \boldsymbol{M} \boldsymbol{F}_1,$$

where $\boldsymbol{P} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}_F(\boldsymbol{X}_F^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}_F)^{-1}\boldsymbol{X}_F^T\boldsymbol{\Psi}^{-1}$, $\boldsymbol{X}_F = (\boldsymbol{M}\boldsymbol{F}_0, \boldsymbol{X})$ and $\boldsymbol{\Psi} = \widehat{\boldsymbol{\Gamma}}_0^{-1} + \widehat{\theta}_0 \boldsymbol{M}\boldsymbol{V}_0\boldsymbol{M}^T + \widehat{\sigma}_a^2 \boldsymbol{B}\boldsymbol{B}^T + \widehat{\sigma}_r^2 \widehat{\boldsymbol{R}}$. Here, $\boldsymbol{\Gamma} = \mathrm{diag}\{\mu_{1,1}(1-\mu_{1,1}), \ldots, \mu_{1,m_1}(1-\mu_{1,m_1}), \ldots, \mu_{n,1}(1-\mu_{n,1}), \ldots, \mu_{n,m_n}(1-\mu_{n,m_n})\}$ is a diagonal N-dimensional matrix, and $\widehat{\boldsymbol{\Gamma}}_0$ and $\widehat{\boldsymbol{R}}$ are $\boldsymbol{\Gamma}$ and $\boldsymbol{R}$ evaluated under model (2.5). Under the null hypothesis, the T_f test statistic has an asymptotic χ^2 distribution with h_1 degrees of freedom.

Under the assumption that $\boldsymbol{c}_1 = \boldsymbol{0}$, if we test $H_0 : \ \theta_1 = 0$, we can construct a variance component score test as

$$T_{vc} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{\Delta}_G \boldsymbol{M} \boldsymbol{V}_1 \boldsymbol{M}^T \boldsymbol{\Delta}_G (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0). \tag{2.7}$$

Under the null hypothesis, T_{vc} asymptotically follows a mixture of χ^2 distribution, $T_{vc} \sim \sum_{l=1}^{m} \xi_l \chi^2_{1,l}$, where $(\xi_1, \ldots, \xi_m)$ are the eigenvalues of the matrix $\boldsymbol{V}_1^{1/2}\boldsymbol{M}^T\boldsymbol{\Delta}_G\boldsymbol{P}\boldsymbol{\Delta}_G\boldsymbol{M}\boldsymbol{V}_1^{1/2}$ and $\chi^2_{1,l}$ are independent χ^2_1 variables. The p-value of T_{vc} can be evaluated by a moment-matching method [61].

We propose a joint test for testing $H_0 : \ \boldsymbol{c}_1 = \boldsymbol{0}$ and $\theta_1 = 0$ using a Cauchy combination test [62] that combines the test of fixed effect, T_f, and the test of variance component, T_{vc}, which we named as Varying-coefficient Mixed Model Association Test (VMMAT). Specifically, the VMMAT test statistic is

$$T_{\mathrm{VMMAT}} = \frac{1}{2}[\tan\{(0.5 - p_f)\pi\} + \tan\{(0.5 - p_{vc})\pi\}], \tag{2.8}$$

where p_f and p_{vc} are p-values of T_f and T_{vc}. Under the null hypothesis, T_{VMMAT} asymptotically follows a Cauchy distribution. Its p-value can be approximated by $p_{\mathrm{VMMAT}} = 0.5 - \arctan(T_{\mathrm{VMMAT}})/\pi$.

The asymptotic null distributions of T_f and T_{vc} are based on the GLMM of Eq. (2.4) which is an equivalent representation of the GLMM with varying coefficients using smoothing splines.

Because parameters are estimated from the DPQL function of Eq. (2.2) with two penalty terms, the estimated variance can be larger than that from the model without penalties. Therefore, the null distributions of the score tests, T_f and T_{vc}, as well as the combined test, T_{VMMAT}, depend on the tuning parameter values. We further assessed the null distribution of VMMAT through type I error experiments in simulation studies.

2.4 Retrospective varying coefficient mixed model association test

Retrospective association tests have been shown to be robust to the trait model misspecification and improve statistical power [54, 55, 16, 56]. In what follows, we introduce a new varying-coefficient test, RVMMAT (Retrospective Varying-coefficient Mixed Model Association Test), for testing time-varying genetic effect between the variant and the trait. RVMMAT also uses a Cauchy combination test to combine two tests: a test for $H_0: \boldsymbol{c}_1 = \mathbf{0}$ under the constraint $\theta_1 = 0$ and a test for $H_0: \theta_1 = 0$ under the constraint $\boldsymbol{c}_1 = \mathbf{0}$. In contrast to the two prospective tests, T_f and T_{vc}, in VMMAT, the two tests for testing fixed effect and variance component in RVMMAT are based on a retrospective model of the genotype given the trait and covariates, such that RVMMAT is less dependent on the correct specification of the trait model. The quasi-likelihood model of the genotype $\boldsymbol{G}$ conditional on the phenotype $\boldsymbol{Y}$ and covariates $\boldsymbol{X}$ under null hypothesis of no time-varying genetic effect is assumed to be

$$E_0(\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}) = 2p\mathbf{1}_n, \qquad \text{Var}_0(\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}) = \sigma_g^2 \boldsymbol{\Phi}, \tag{2.9}$$

where p is a nuisance parameter representing the tested variant's MAF, $\mathbf{1}_n = (1, 1, \ldots, 1)^T$ is a length n vector, σ_g^2 is an unknown variance parameter, and $\boldsymbol{\Phi}$ is an $n \times n$ genetic relationship matrix (GRM) measuring the similarity among individuals due to genetic variation, which can be estimated using genome-wide data.

When we test $H_0: \boldsymbol{c}_1 = \mathbf{0}$ under the assumption that $\theta_1 = 0$, the same score function $\boldsymbol{U}_0(\boldsymbol{c}_1)$

is considered. Because the column space of $\boldsymbol{X}_F = (\boldsymbol{M}\boldsymbol{F}_0, \boldsymbol{X})$ is orthogonal to the vector of null phenotypic residuals $\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0$ obtained by fitting model (2.5), the mean model of $\boldsymbol{G}$ under the null in Eq. (2.9) satisfies

$$E_0(\boldsymbol{U}_0(\boldsymbol{c}_1) \mid \boldsymbol{Y}, \boldsymbol{X}) = E_0[(\boldsymbol{\Delta}_G \boldsymbol{M}\boldsymbol{F}_1)^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0) \mid \boldsymbol{Y}, \boldsymbol{X}] = 2p(\boldsymbol{M}\boldsymbol{F}_1)^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0) = \boldsymbol{0},$$

if $h_1 \leq h_0$. In practice, we commonly use smoothing splines of the same order for $\gamma_0(t)$ and $\gamma_1(t)$ in Model (2.1). Thus, we consider the score function $\boldsymbol{U}_0(\boldsymbol{c}_1)$ and construct a score test under the retrospective model of Eq. (2.9), given by

$$T_f^R = (\boldsymbol{U}_0(\boldsymbol{c}_1))^T[\mathrm{Var}_0(\boldsymbol{U}_0(\boldsymbol{c}_1) \mid \boldsymbol{Y}, \boldsymbol{X})]^{-1}\boldsymbol{U}_0(\boldsymbol{c}_1). \tag{2.10}$$

Here, the variance of $\boldsymbol{U}_0(\boldsymbol{c}_1)$ is evaluated by

$$\begin{aligned}
\mathrm{Var}_0(\boldsymbol{U}_0(\boldsymbol{c}_1) \mid \boldsymbol{Y}, \boldsymbol{X}) &= \mathrm{Var}_0((\boldsymbol{\Delta}_G \boldsymbol{M}\boldsymbol{F}_1)^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0) \mid \boldsymbol{Y}, \boldsymbol{X}) \\
&= \mathrm{Var}_0((\boldsymbol{\Delta}_R \boldsymbol{M}\boldsymbol{F}_1)^T \boldsymbol{B}\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}) \\
&= \widehat{\sigma}_g^2(\boldsymbol{\Delta}_R \boldsymbol{M}\boldsymbol{F}_1)^T \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}^T\boldsymbol{\Delta}_R \boldsymbol{M}\boldsymbol{F}_1,
\end{aligned}$$

where $\boldsymbol{\Delta}_R = \mathrm{diag}\{\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0\} = \mathrm{diag}\{(Y_{1,1} - \widehat{\mu}_{0;1,1}), \ldots, (Y_{1,m_1} - \widehat{\mu}_{0;1,m_1}), \ldots, (Y_{n,1} - \widehat{\mu}_{0;n,1}), \ldots, (Y_{n,m_n} - \widehat{\mu}_{0;n,m_n})\}$ is an N-dimensional diagonal matrix of the phenotypic residuals. Under Hardy-Weinberg equilibrium, the variance of the genotype is estimated by $\widehat{\sigma}_g^2 = 2\hat{p}(1 - \hat{p})$, where $\hat{p}$ is the sample MAF of the tested variant. Under the null hypothesis, the T_f^R test statistic has an asymptotic χ^2 distribution with h_1 degrees of freedom.

If we test $H_0 : \ \theta_1 = 0$ under the assumption that $\boldsymbol{c}_1 = \boldsymbol{0}$, a retrospective variance component score test under Model (2.9) can be constructed as

$$T_{vc}^R = (\boldsymbol{B}\boldsymbol{G})^T \boldsymbol{\Delta}_R \boldsymbol{M}\boldsymbol{V}_1\boldsymbol{M}^T\boldsymbol{\Delta}_R \boldsymbol{B}\boldsymbol{G} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{\Delta}_G \boldsymbol{M}\boldsymbol{V}_1\boldsymbol{M}^T\boldsymbol{\Delta}_G(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0), \tag{2.11}$$

which has the same form as the prospective variance component test T_{vc}. However, under the null hypothesis, given the trait and covariates, T_{vc}^R asymptotically follows a mixture of χ^2 distribution, $T_{vc}^R \sim \sum_{l=1}^m \zeta_l \chi_{1,l}^2$, where $(\zeta_1, \ldots, \zeta_m)$ are the eigenvalues of the matrix $\widehat{\sigma}_g^2 \boldsymbol{V}_1^{1/2}\boldsymbol{M}^T\boldsymbol{\Delta}_R \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}^T\boldsymbol{\Delta}_R \boldsymbol{M}\boldsymbol{V}_1^{1/2}$.

The RVMMAT test statistic is defined by combining the two retrospective tests, T_f^R and T_{vc}^R, expressed as

$$T_{\text{RVMMAT}} = \frac{1}{2}[\tan\{(0.5 - p_f^R)\pi\} + \tan\{(0.5 - p_{vc}^R)\pi\}], \tag{2.12}$$

where p_f^R and p_{vc}^R are p-values of T_f^R and T_{vc}^R respectively. Under the null hypothesis, T_{RVMMAT} asymptotically follows a Cauchy distribution.

The asymptotic null distributions of T_f^R and T_{vc}^R are based on the retrospective model of Eq. (2.9) in which genotypes are treated as random conditional on the phenotype and covariates. Therefore, the estimated variance in the trait model due to the tuning parameters in penalty terms does not impact the null distributions of the retrospective score tests, T_f^R and T_{vc}^R, as well as the combined test, T_{RVMMAT}. We have also assessed the null distribution of RVMMAT through type I error experiments in simulation studies.

CHAPTER 3
RETROSPECTIVE VARIANT-SET VARYING COEFFICIENT MIXED MODEL ASSOCIATION TESTS FOR RARE SNPS

3.1 Introduction

In this chapter, we propose a group of retrospective variant set varying coefficient mixed model association tests, RSVMMATs, to detect time-varying effects of a set of genetic variants on a binary trait measured repeatedly over time. Our proposed tests include a burden test, a SKAT (sequence kernel association test) -type test, and an omnibus test that combines those two using the Cauchy combination test [62]. In a burden test, RSVMMAT-B, time-varying genetic effects in a region are factorized as pre-specified weights multiplied by a smooth function of time. In a SKAT-type test, RSVMMAT-S, time-varying genetic effects are treated as separated as fixed effects plus random effects and all variance components of the random effect can be factorized as pre-specified weights multiplying a common variance component. Finally, an omnibus test, RSVMMAT-A, combines RSVMMAT-B and RSVMMAT-S by the Cauchy combination test [62]. As a comparison, we have also developed SVMMAT-B, SVMMAT-S, and SVMMAT-A, a group of prospective variant set varying coefficient mixed model association tests.

3.2 GLMM with varying coefficients

Consider a group of n subjects and the ith subject has m_i repeated observations on a binary trait. Genome-wide measures of genetic variation and measurements of p non-genetic covariates are available for each subject. Let Y_{ij} and $\boldsymbol{X}_{ij}$ be the value of the binary trait and the $p-$dimensional covariate vector on subject i measured at time t_{ij}; $\boldsymbol{Y} = (Y_{1,1}, \ldots, Y_{1,m_1}, \ldots, Y_{n,1}, \ldots, Y_{n,m_n})^T$ be the outcome vector of length N and $\boldsymbol{X} = \left(\boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{1,m_1}, \ldots, \boldsymbol{X}_{n,1}, \ldots, \boldsymbol{X}_{n,m_n}\right)^T$ be the $N \times p$

covariate matrix, where $N = \sum_{i=1}^{n} m_i$ is the total number of observations. Both static factors like gender and dynamic variables such as body weight are allowed as covariates. For a genetic region with q variants, let $\boldsymbol{G} = (\boldsymbol{G}_1, \ldots, \boldsymbol{G}_n)^T$ be the $n \times q$ genotype matrix of n subjects at the q variants, where $\boldsymbol{G}_i = (G_{i1}, \ldots, G_{iq})^T$ with $G_{ik} = 0$, 1, or 2, representing number of copies of minor allele at the kth variant.

We consider the following generalized linear mixed model (GLMM) with varying coefficients for a longitudinal binary trait

$$\begin{aligned} logit\left(\mu_{ij}\right) = \gamma_0\left(t_{ij}\right) + G_{i1}\gamma_1\left(t_{ij}\right) + \cdots + G_{iq}\gamma_q\left(t_{ij}\right) + \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + a_i + r_{ij}, \\ i = 1, \ldots, n; j = 1, \ldots, m_i, \end{aligned} \tag{3.1}$$

where $\mu_{ij} = E\left(Y_{ij} \mid \boldsymbol{G}_i, \boldsymbol{X}_{ij}, a_i, r_{ij}\right)$, Y_{ij} is the phenotype measured at time t_{ij} for individual i, $\boldsymbol{G}_i$ is a vector of q genetic variants, $\boldsymbol{X}_{ij}$ is a vector of covariates,. Time-varying effects of the q genetic variants $\left(\gamma_1\left(t\right), ..., \gamma_q\left(t\right)\right)$ and a time-varying intercept $\gamma_0\left(t\right)$ are modeled as smooth nonparametric functions of time t, and $\boldsymbol{\beta}$ is a vector of the effects of the p covariates. The correlations among repeated measurements are captured by two random effects: the subject random effect a_i and the subject-specific time-dependent random effect r_{ij} [52, 16]. We assume that $a_i \sim N\left(0, \sigma_a^2\right)$, are independent. The vector $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{im_i})^T$ is assumed to follow a multivariate normal distribution, $\boldsymbol{r}_i \sim \text{MVN}\left(\boldsymbol{0}, \sigma_r^2\boldsymbol{R}_i\right)$, where the correlation matrix $\boldsymbol{R}_i$ is modeled by an AR(1) structure in which τ is the unknown parameter. Given the random effects a_i and r_{ij}, the response Y_{ij} are assumed to be independent.

We maximize the following double penalized quasi-likelihood (DPQL) function to estimate $\gamma_0\left(t\right)$, $\gamma_1\left(t\right), ..., \ \gamma_q\left(t\right)$ [57, 38]:

$$\begin{aligned} l_{dp}\{\gamma_0\left(\cdot\right), \gamma_1\left(\cdot\right), ..., \ \gamma_q\left(\cdot\right), \boldsymbol{\beta}, \sigma_a^2, \sigma_r^2, \ \tau\} = -\frac{1}{2}\sum_{i,j} D_{ij}\left(Y_{ij}, \mu_{ij}\right) - \frac{1}{2\sigma_a^2}\boldsymbol{a}^T\boldsymbol{a} - \frac{1}{2\sigma_r^2}\boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r} \\ - \frac{\lambda_0}{2}\int\left\{\gamma_0^{(h_0)}\left(t\right)\right\}^2 dt - \frac{\lambda_1}{2}\int\left\{\gamma_1^{(h_1)}\left(t\right)\right\}^2 dt - ... - \frac{\lambda_q}{2}\int\left\{\gamma_q^{(h_q)}\left(t\right)\right\}^2 dt, \end{aligned} \tag{3.2}$$

where $\boldsymbol{a} = (a_1, \ldots, a_n)^T$ and $\boldsymbol{r} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)^T$ are the two vectors of random effects, $\boldsymbol{R} = \text{diag}\{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n\}$ is a block diagonal matrix, $D_{ij}(Y_{ij}, \mu_{ij}) = -2\int_{Y_{ij}}^{\mu_{ij}} \frac{Y_{ij}-u}{u(1-u)}du$ is the conditional deviance function of a binary outcome Y_{ij} given random effects a_i and r_{ij}, λ_k $(k = 0,\ 1, ...,\ q)$ are smoothing parameters that control the smoothness of $\gamma_k(t)$, and h_k are positive integers for the derivative order of $\gamma_k(t)$.

The maximizers for the nonparametric functions $\gamma_k(t)$ in the DPQL function in Eq. (3.2) are natural smoothing splines of order $2h_k$, which can be expressed as [58]

$$\gamma_k(t) = \sum_{s=1}^{h_k} c_{ks}F_{ks}(t) + \sum_{l=1}^{m} d_{kl}V_k\left(t, t_l^0\right), \qquad k = 0,\ 1,\ ...,\ q,$$

where $0 < t_1^0 < \ldots < t_m^0 < 1$ are m distinct knots of t_{ij}, $F_{ks}(t)$ is a polynomial of order $s-1$ (e.g., $F_{ks}(t) = t^{(s-1)}/(s-1)!, \quad s = 1, \ldots, h_k$), and $V_k(t_1, t_2) = \frac{1}{[(h_k-1)!]^2}\int_0^1 (t_1-u)_+^{h_k-1}(t_2-u)_+^{h_k-1}du$ with $u_+ = \max\{u, 0\}$. Let $\boldsymbol{c}_k = (c_{k,1}, \ldots,\ c_{k,h_k})^T$, $\boldsymbol{d}_k = (d_{k1}, \ldots,\ d_{km})^T$, and $\boldsymbol{\gamma}_k = \left(\gamma_k(t_1^0), \ldots, \gamma_k(t_m^0)\right)^T$ for $k = 0,\ 1,\ ...,\ q$. Then $\boldsymbol{\gamma}_k$ can be expressed as

$$\boldsymbol{\gamma}_k = \boldsymbol{F}_k\boldsymbol{c}_k + \boldsymbol{V}_k\boldsymbol{d}_k,$$

where $\boldsymbol{F}_k$ is an $m \times h_k$ matrix with its (l, s)th entry equal to $F_{ks}\left(t_l^0\right)$, and $\boldsymbol{V}_k$ is a positive definite matrix with the (l, s)th entry equal to $V_k\left(t_l^0, t_s^0\right)$. Therefore, the DPQL function in Eq. (3.2) becomes

$$\begin{aligned} l_{dp}\{\gamma_0(\cdot), \gamma_1(\cdot), ..., \gamma_q(\cdot), \boldsymbol{\beta}, \sigma_a^2, \sigma_r^2,\ \tau\} = &-\frac{1}{2}\sum_{i,j} D_{ij}(Y_{ij}, \mu_{ij}) - \frac{1}{2\sigma_a^2}\boldsymbol{a}^T\boldsymbol{a} - \frac{1}{2\sigma_r^2}\boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r} \\ &- \frac{\lambda_0}{2}\boldsymbol{d}_0^T\boldsymbol{V}_0\boldsymbol{d}_0 - \frac{\lambda_1}{2}\boldsymbol{d}_1^T\boldsymbol{V}_1\boldsymbol{d}_1 - ... - \frac{\lambda_q}{2}\boldsymbol{d}_q^T\boldsymbol{V}_q\boldsymbol{d}_q. \end{aligned} \tag{3.3}$$

We can treat $\boldsymbol{d}_k$ in the $\boldsymbol{\gamma}_k$ as random effects distributed as $\boldsymbol{d}_k \sim N(\boldsymbol{0}, \theta_k\boldsymbol{V}_k^{-1})$, with $\theta_k = \lambda_k^{-1}$, $k = 0,\ 1,\ ...,\ q$. Then, it follows that the maximizers in Eq. (3.3) are obtained by fitting the following GLMM representation of model (3.1), expressed in a matrix form as

$$\begin{aligned} logit(\boldsymbol{\mu}) = &\boldsymbol{M}\boldsymbol{F}_0\boldsymbol{c}_0 + \boldsymbol{\Delta}_{G_1}\boldsymbol{M}\boldsymbol{F}_1\boldsymbol{c}_1 + ... + \boldsymbol{\Delta}_{G_q}\boldsymbol{M}\boldsymbol{F}_q\boldsymbol{c}_q + \boldsymbol{M}\boldsymbol{V}_0\boldsymbol{d}_0 \\ &+ \boldsymbol{\Delta}_{G_1}\boldsymbol{M}\boldsymbol{V}_1\boldsymbol{d}_1 + ... + \boldsymbol{\Delta}_{G_q}\boldsymbol{M}\boldsymbol{V}_q\boldsymbol{d}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{a} + \boldsymbol{r}. \end{aligned} \tag{3.4}$$

Here we used an $N \times m$ incidence matrix $\boldsymbol{M}$ to map t_{ij} to the m distinct knots $t_1^0, \ldots, t_m^0$, an $N \times n$ design matrix $\boldsymbol{B}$ to map the subject-level genotype vector $\boldsymbol{G}$ to a measurement-level genotype vector $\boldsymbol{BG}$, with its (l,i)th entry $B_{li} = 1$ if the lth entry of $\boldsymbol{Y}$ is a measurement on subject i and 0 otherwise, and an $N \times N$ matrix $\boldsymbol{\Delta}_{G_k} = \text{diag}\{G_{1k}, \ldots G_{1k}, \ldots, G_{nk}, \ldots G_{nk}\}$ of genotypes of the n subjects at the kth variant.

3.3 Variant set varying coefficient mixed model association tests

We are interested in investigating time-varying effects of the genetic variants $\boldsymbol{G}$ on trait $\boldsymbol{Y}$, adjusted for $\boldsymbol{X}$. We test $H_0 : \ \gamma_1(t) = \cdots = \gamma_q(t) = 0$ in model (2.1) which is equivalent to testing $H_0 : \ \boldsymbol{c}_1 = \cdots = \ \boldsymbol{c}_q = \boldsymbol{0}$ and $\theta_1 = \cdots = \ \theta_q = 0$ in model (2.4). The reduced GLMM under the null hypothesis becomes

$$\text{logit}(\boldsymbol{\mu}_0) = \boldsymbol{MF}_0\boldsymbol{c}_0 + \boldsymbol{MV}_0\boldsymbol{d}_0 + \boldsymbol{X\beta} + \boldsymbol{Ba} + \boldsymbol{r}, \tag{3.5}$$

with $\boldsymbol{\mu}_0 = E(\boldsymbol{Y} \ \mid \ \boldsymbol{F}_0, \boldsymbol{X}, \boldsymbol{d}_0, \boldsymbol{a}, \boldsymbol{r})$. We use smooth splines of same order for $\gamma_k(t)$ and set $h_k = h$, for $1 \leq k \leq q$, then $\boldsymbol{F}_k = \boldsymbol{F}$, and $\boldsymbol{V}_k = \boldsymbol{V}$, for $1 \leq k \leq q$.

We first propose a burden-type test by assuming that $\boldsymbol{c}_k = w_k\boldsymbol{c}$ and $\boldsymbol{d}_k = w_k\boldsymbol{d}$ with a prespecified weight w_k for kth variant and $\boldsymbol{d} \sim \boldsymbol{N}(\boldsymbol{0}, \theta\boldsymbol{V}^{-1})$. If we test $H_0 : \ \boldsymbol{c} = \boldsymbol{0}$ under the assumption that $\theta = 0$, a score test can be constructed as:

$$T_f = [\boldsymbol{U}_0(\boldsymbol{c})]^T \{Var[\boldsymbol{U}_0(\boldsymbol{c})]\}^{-1} U_0(\boldsymbol{c}), \tag{3.6}$$

where $\boldsymbol{U}_0(\boldsymbol{c}) = (\boldsymbol{\Delta_G MF})^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)$ is the score function for $\boldsymbol{c}$, $\boldsymbol{\Delta_G} = \text{diag}\{\boldsymbol{G}_1^T\boldsymbol{w}, ..., \boldsymbol{G}_1^T\boldsymbol{w}, ..., \boldsymbol{G}_n^T\boldsymbol{w}, ..., \boldsymbol{G}_n^T\boldsymbol{w}\}$ with $\boldsymbol{w} = (w_1, \ldots, w_q)^T$, and $\widehat{\boldsymbol{\mu}}_0 = logit^{-1}(\boldsymbol{MF}_0\widehat{\boldsymbol{c}}_0 + \boldsymbol{MV}_0\widehat{\boldsymbol{d}}_0 + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{B}\widehat{\boldsymbol{a}} + \widehat{\boldsymbol{r}})$ is a vector of fitted values under model (2.5), which can be obtained using the penalized quasi-likelihood methods [59]. Conditioning on the genotype $\boldsymbol{G}$ and covariate $\boldsymbol{X}$, the variance of the score function $\boldsymbol{U}_0(\boldsymbol{c})$ under H_0 is

$$\text{Var}[\boldsymbol{U}_0(\boldsymbol{c})] = (\boldsymbol{\Delta}_G\boldsymbol{MF})^T\boldsymbol{P}\boldsymbol{\Delta}_G\boldsymbol{MF}_{\boldsymbol{c}},$$

where $\boldsymbol{P} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}_F(\boldsymbol{X}_F^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}_F)^{-1}\boldsymbol{X}_F^T\boldsymbol{\Psi}^{-1}$, $\boldsymbol{X}_F = (\boldsymbol{MF}_0, \boldsymbol{X})$ and $\boldsymbol{\Psi} = \widehat{\boldsymbol{\Gamma}}_0^{-1} + \widehat{\theta}_0\boldsymbol{MV}_0\boldsymbol{M^T} + \widehat{\sigma}_a^2\boldsymbol{BB}^T + \widehat{\sigma}_r^2\widehat{\boldsymbol{R}}$. Here, $\boldsymbol{\Gamma} = \text{diag}\{\mu_{1,1}(1-\mu_{1,1}), \ldots, \mu_{1,m_1}(1-\mu_{1,m_1}), \ldots, \mu_{n,1}(1-\mu_{n,1}), \ldots, \mu_{n,m_n}(1-\mu_{n,m_n})\}$ is an N-dimensional diagonal matrix, and $\widehat{\boldsymbol{\Gamma}}_0$ and $\widehat{\boldsymbol{R}}$ are $\boldsymbol{\Gamma}$ and $\boldsymbol{R}$ evaluated under Model (3.5). Under the null hypothesis, the T_f test statistic has an asymptotic χ^2 distribution with h degrees of freedom.

On the other hand, if we assume that $\boldsymbol{c} = \boldsymbol{0}$ and test for $H_0 : \theta = 0$, a variance component score test takes the form

$$T_{vc-B} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{\Delta}_G \boldsymbol{MVM}^T \boldsymbol{\Delta}_G (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0). \tag{3.7}$$

Under the null hypothesis, T_{vc-B} asymptotically follows a mixture of χ^2 distribution, $T_{vc-B} \sim \sum_{l=1}^{m} \xi_l \chi_{1,l}^2$, where $\chi_{1,l}^2$ are independent χ_1^2 variables and $(\xi_1, \ldots, \xi_m)$ are eigenvalues of $\boldsymbol{V}^{1/2}\boldsymbol{M^T\Delta_G P\Delta_G MV}^{1/2}$. A moment-matching method [61] can be used to calculate the P-value of T_{vc-B}.

To test $H_0 : \boldsymbol{c} = 0$ and $\theta = 0$ jointly, we combine the test of fixed effect, T_f, and the test of variance component, T_{vc-B}, using a Cauchy combination test [62] and propose variant-Set Varying-coefficient Mixed Model Association Test - Burden (SVMMAT-B). Specifically, the SVMMAT-B test statistic is,

$$T_{SVMMAT-B} = \frac{1}{2}\left[\tan\left\{(0.5 - p_f)\,\pi\right\} + \tan\left\{(0.5 - p_{vc-B})\,\pi\right\}\right], \tag{3.8}$$

where p_f and p_{vc-B} are p-values of T_f and T_{vc-B} respectively. Under the null hypothesis, $T_{\text{SVMMAT}-\text{B}}$ asymptotically follows a Cauchy distribution. Its p-value can be approximated by $p_{\text{SVMMAT}-\text{B}} = 0.5 - \arctan\left(T_{\text{SVMMAT}-\text{B}}\right)/\pi$.

Next, we propose a SKAT-type test by assuming that $\boldsymbol{c}_k = w_k\boldsymbol{c}$ and $\boldsymbol{d}_k \sim \boldsymbol{N}(\boldsymbol{0}, w_k^2\theta\boldsymbol{V}^{-1})$ for prespecified weights w_k, $k = 0,\ 1, ..., q$. We use the same T_f as in SVMMAT-B to test $H_0 :\ \boldsymbol{c} = \boldsymbol{0}$ under the assumption of $\theta = 0$. To test the variance component $H_0 : \theta = 0$ under the assumption

of $\boldsymbol{c} = \mathbf{0}$, we use the following score statistic:

$$T_{vc-S} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \left(\boldsymbol{\Delta}_{G_1}\boldsymbol{M}, \ldots, \boldsymbol{\Delta}_{G_q}\boldsymbol{M}\right) \left(\boldsymbol{W}^2 \otimes \boldsymbol{V}\right) \left(\boldsymbol{\Delta}_{G_1}\boldsymbol{M}, \ldots, \boldsymbol{\Delta}_{G_q}\boldsymbol{M}\right)^T (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0), \quad (3.9)$$

where $\otimes$ is the Kronecker product and $\boldsymbol{W} = \text{diag}(w_1, ..., w_q)$. Under the null hypothesis, T_{vc-S} asymptotically follows a mixture of χ^2 distribution, $T_{vc-S} \sim \sum_{l=1}^{m} \zeta_m \chi^2_{1,l}$ where $(\zeta_1, \ldots, \zeta_m)$ are eigenvalues of $(\boldsymbol{W} \otimes \boldsymbol{V}^{1/2})(\boldsymbol{\Delta}_{G_1}\boldsymbol{M}, \ldots, \boldsymbol{\Delta}_{G_q}\boldsymbol{M})^T \boldsymbol{P} (\boldsymbol{\Delta}_{G_1}\boldsymbol{M}, \ldots, \boldsymbol{\Delta}_{G_q}\boldsymbol{M})(\boldsymbol{W} \otimes \boldsymbol{V}^{1/2})$. The p-value of T_{vc-S} can be calculated by the moment-matching method of [61]. Similarly, we propose a joint test for testing $H_0 : \boldsymbol{c} = 0$ and $\theta = 0$ using the Cauchy combination test [62] that combines the test of fixed effect, T_f, and the test of variance component, T_{vc-S}, which we named as variant-Set Varying-coefficient Mixed Model Association Test - SKAT (SVMMAT-S). Specifically, the SVMMAT-S test statistic is,

$$T_{SVMMAT-S} = \frac{1}{2} \left[\tan\left\{(0.5 - p_f)\,\pi\right\} \; + \tan\left\{(0.5 - p_{vc-S})\,\pi\right\} \right], \quad (3.10)$$

where p_f and p_{vc-S} are the p-values of T_f and T_{vc-S} respectively. Its p-value can be approximated by $p_{\text{SVMMAT-S}} = 0.5 - \arctan\left(T_{\text{SVMMAT-S}}\right)/\pi$.

Third, we combine $T_{SVMMAT-B}$ and $T_{SVMMAT-S}$ using the Cauchy combination test [62] and propose the following variant-Set Varying-coefficient Mixed Model Association Test – A (SVMMAT-A)

$$T_{SVMMAT-A} = \frac{1}{2} \left[\tan\left\{(0.5 - p_{SVMMAT-B})\,\pi\right\} \; + \tan\left\{(0.5 - p_{SVMMAT-S})\,\pi\right\} \right],$$

where $p_{SVMMAT-B}$ and $p_{SVMMAT-S}$ are the p-values of $T_{SVMMAT-B}$ and $T_{SVMMAT-S}$ respectively and p-value of $T_{SVMMAT-A}$ can be approximated by $p_{SVMMAT-A} = 0.5 - \arctan(T_{SVMMAT-A})/\pi$.

3.4 Retrospective variant set varying coefficient mixed model association test

It has been shown that retrospective association tests are robust to the trait model misspecification and have improved statistical power [54, 55, 16, 56]. In the ascertained sample, it preserves

additional information on genotype-phenotype association carried by the joint distribution of the genotype and covariates [54, 56]. In what follows, we propose a group of retrospective variant set varying coefficient mixed model association tests, RSVMMATs, to detect time-varying effects of a set of genetic variants on a binary trait measured repeatedly over time. Similar to RVMMAT, RSVMMATs are based on a model in which the genotype is assumed to be random given the phenotype and covariate. The quasi-likelihood model of the genotype $\widetilde{\boldsymbol{G}}$ conditional on the phenotype $\boldsymbol{Y}$ and covariates $\boldsymbol{X}$ under the null hypothesis of no time-varying genetic effects is assumed to be

$$E_0\left(\widetilde{\boldsymbol{G}} \mid \boldsymbol{Y}, \boldsymbol{X}\right) = 2\boldsymbol{p} \otimes \mathbf{1}_n, \qquad \mathrm{Var}_0\left(\widetilde{\boldsymbol{G}} \mid \boldsymbol{Y}, \boldsymbol{X}\right) = \boldsymbol{\Sigma}_G \otimes \boldsymbol{\Phi}, \tag{3.11}$$

where $\widetilde{\boldsymbol{G}} = \mathrm{vec}(\boldsymbol{G}) = (G_{11}, \ldots, G_{n1}, \ldots, G_{1q}, \ldots, G_{nq})^T$ is an nq-dimensional vector denoting the vectorization of the genotype matrix $\boldsymbol{G}$, $\boldsymbol{p} = (p_1, \ldots, \ p_q)^T$ is a vector of the MAFs of the q genetic variants which are treated as nuisance parameters, $\boldsymbol{\Sigma}_G = \boldsymbol{D}^{1/2}\boldsymbol{R}\boldsymbol{D}^{1/2}$ is an $q \times q$ covariance matrix with $\boldsymbol{D} = diag(\sigma_1^2, \ldots, \sigma_q^2)$ consisting of q unknown variance parameters of the genetic variants and a correlation matrix $\boldsymbol{R}$ capturing the linkage disequilibrium (LD) structure of the variants, and $\boldsymbol{\Phi}$ is an $n \times n$ GRM. The matrices $\boldsymbol{R}$ and $\boldsymbol{\Phi}$ can be estimated respectively using the genotype matrix $\boldsymbol{G}$ and the genome-wide data.

When we test $H_0 : \boldsymbol{c} = \mathbf{0}$ under the assumption that $\theta = 0$, the same score function $\boldsymbol{U}_0(\boldsymbol{c})$ is considered. The vector of null phenotypic residuals $\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0$ is obtained by fitting model (3.5) and orthogonal to the column space of $\boldsymbol{X}_F = (\boldsymbol{M}\boldsymbol{F}_0, \boldsymbol{X})$. Thus, the null mean model of $\boldsymbol{G}$ in Eq. (3.9) ensures that

$$E_0\left(\boldsymbol{U}_0\left(\boldsymbol{c}\right) \mid \boldsymbol{Y}, \boldsymbol{X}\right) = E_0\left[\left(\boldsymbol{\Delta}_G\boldsymbol{M}\boldsymbol{F}\right)^T\left(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0\right) \mid \boldsymbol{Y}, \boldsymbol{X}\right] = 2\boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{w}\left(\boldsymbol{M}\boldsymbol{F}\right)^T\left(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0\right) = \mathbf{0},$$

if $h \leq h_0$. In practice, we commonly use natural smoothing splines of the same order for $\gamma_0(t)$ and $\gamma_1(t), ..., \ \gamma_q(t)$ in model (3.1). Therefore, we construct a score test under the retrospective model

of Eq. (3.11) using the score function $\boldsymbol{U}_0(\boldsymbol{c})$ as follows

$$T_f^R = (\boldsymbol{U}_0(\boldsymbol{c}))^T [\mathrm{Var}_0(\boldsymbol{U}_0(\boldsymbol{c}) \mid \boldsymbol{Y}, \boldsymbol{X})]^{-1} \boldsymbol{U}_0(c). \tag{3.12}$$

Here, the variance of $\boldsymbol{U}_0(\boldsymbol{c})$ is evaluated by

$$\begin{aligned}
\mathrm{Var}_0(\boldsymbol{U}_0(\boldsymbol{c}) \mid \boldsymbol{Y}, \boldsymbol{X}) &= \mathrm{Var}_0\left((\boldsymbol{\Delta}_G \boldsymbol{M}\boldsymbol{F})^T (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0) \mid \boldsymbol{Y}, \boldsymbol{X}\right) \\
&= \mathrm{Var}_0\left((\boldsymbol{\Delta}_R \boldsymbol{M}\boldsymbol{F})^T \boldsymbol{B}\boldsymbol{G}\boldsymbol{w} \mid \boldsymbol{Y}, \boldsymbol{X}\right) \\
&= (\boldsymbol{w}^{\boldsymbol{T}} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{G}} \boldsymbol{w})(\boldsymbol{\Delta}_{\mathrm{R}} \boldsymbol{M}\boldsymbol{F})^{\boldsymbol{T}} \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}^T \boldsymbol{\Delta}_{\mathrm{R}} \boldsymbol{M}\boldsymbol{F},
\end{aligned}$$

where $\boldsymbol{\Delta}_R = \mathrm{diag}\{\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0\} = \mathrm{diag}(Y_{1,1} - \widehat{\mu}_{0;1,1}), \ldots, (Y_{1,m_1} - \widehat{\mu}_{0;1,m_1}), \ldots, (Y_{n,1} - \widehat{\mu}_{0;n,1}), \ldots, (Y_{n,m_n} - \widehat{\mu}_{0;n,m_n})$is an N-dimensional diagonal matrix of the phenotypic residuals and $\widehat{\boldsymbol{\Sigma}}_G = \widehat{\boldsymbol{D}}^{1/2} \boldsymbol{R} \widehat{\boldsymbol{D}}^{1/2}$ with $\widehat{\boldsymbol{D}} = \mathrm{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_q^2)$. Under Hardy-Weinberg equilibrium, we can estimate the variance parameters as $\widehat{\sigma}_k^2 = 2\hat{p}_k(1 - \hat{p}_k)$, $1 \leq k \leq q$, using sample MAFs of the testing variants. The null distribution of the test statistic T_f^R is an asymptotic χ^2 distribution with h degrees of freedom.

If we test $H_0 : \theta = 0$ under the assumption that $\boldsymbol{c} = \boldsymbol{0}$, a retrospective variance component score test under model (3.9) takes the same form as the prospective variance component test T_{vc-B}, that is

$$T_{vc-B}^R = (\boldsymbol{B}\boldsymbol{G}\boldsymbol{w})^T \boldsymbol{\Delta}_R \boldsymbol{M}\boldsymbol{V}\boldsymbol{M}^T \boldsymbol{\Delta}_R \boldsymbol{B}\boldsymbol{G}\boldsymbol{w} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{\Delta}_G \boldsymbol{M}\boldsymbol{V}\boldsymbol{M}^T \boldsymbol{\Delta}_G (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0), \tag{3.13}$$

However, the null distribution of T_{vc-B}^R, given the trait and covariates, is an asymptotic mixture of χ^2 distributions $T_{vc-B}^R \sim \sum_{l=1}^m \xi_{rl} \chi_{1,l}^2$, where $(\xi_{r1}, \ldots, \xi_{rm})$ are eigenvalues of the matrix $(\boldsymbol{w}^T \boldsymbol{\Sigma}_G \boldsymbol{w}) \boldsymbol{V}^{1/2} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{\Delta}_{\boldsymbol{R}} \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}^T \boldsymbol{\Delta}_{\mathrm{R}} \boldsymbol{M}\boldsymbol{V}^{1/2}$.

The RSVMMAT-B test statistic is obtained by combining the two retrospective tests, T_f^R and T_{vc-B}^R, and expressed as

$$T_{\mathrm{RSVMMAT-B}} = \frac{1}{2}\left[\tan\left\{\left(0.5 - p_f^R\right)\pi\right\} + \tan\left\{\left(0.5 - p_{vc-B}^R\right)\pi\right\}\right], \tag{3.14}$$

where p_f^R and p_{vc-B}^R are p-values of T_f^R and T_{vc-B}^R. Under the null hypothesis, $T_{\mathrm{RSVMMAT-B}}$ asymptotically follows a Cauchy distribution.

A SKAT-type test statistic under retrospective model takes the same form as the prospective SKAT-type test T_{vc-S} but with a different null distribution, that is

$$\begin{aligned} T^R_{vc-S} &= (\boldsymbol{BGW})^T \boldsymbol{\Delta}_R \boldsymbol{MVM}^T \boldsymbol{\Delta}_R \boldsymbol{BGW} \\ &= (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \left(\boldsymbol{\Delta}_{G_1}\boldsymbol{M}, \ldots, \boldsymbol{\Delta}_{G_q}\boldsymbol{M}\right) \left(\boldsymbol{W^2} \otimes \boldsymbol{V}\right) \left(\boldsymbol{\Delta}_{G_1}\boldsymbol{M}, \ldots, \boldsymbol{\Delta}_{G_q}\boldsymbol{M}\right)^{\boldsymbol{T}} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0). \end{aligned}$$

Under H_0, T^R_{vc-S} asymptotically follows a mixture of χ^2 distributions, $T^R_{vc-S} \sim \sum_{l=1}^{m} \zeta_{rl}\chi^2_{1,l}$ with $(\zeta_{r1}, \ldots, \zeta_{rm})$ being eigenvalues of $(\boldsymbol{W\Sigma}_G\boldsymbol{W})\otimes(\boldsymbol{V}^{1/2}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{\Delta_R}\boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}^T\boldsymbol{\Delta}_{\mathrm{R}}\boldsymbol{MV}^{1/2})$. The RSVMMAT-S test statistic is defined by combining the two retrospective tests, T^R_f and T^R_{vc-S}, expressed as

$$T_{\mathrm{RSVMMAT-S}} = \frac{1}{2}\left[\tan\left\{\left(0.5 - p^R_f\right)\pi\right\} + \tan\left\{\left(0.5 - p^R_{vc-S}\right)\pi\right\}\right], \tag{3.15}$$

where p^R_f and p^R_{vc-S} are p-values of T^R_f and T^R_{vc-S}. Under the null hypothesis, $T_{\mathrm{RSVMMAT-S}}$ asymptotically follows a Cauchy distribution.

Finally, we combine $T_{\mathrm{RSVMMAT-B}}$ and $T_{\mathrm{RSVMMAT-S}}$ to define RSVMMAT-A test statistic as follows

$$T_{\mathrm{RSVMMAT-A}} = 0.5\tan\left\{(0.5 - p_{\mathrm{RSVMMAT-B}})\right\}\pi + 0.5\tan\left\{(0.5 - p_{\mathrm{RSVMMAT-S}})\right\}\pi,$$

where $p_{\mathrm{RSVMMAT-B}}$ and $p_{\mathrm{RSVMMAT-S}}$ are the p-values of $T_{\mathrm{RSVMMAT-B}}$ and $T_{\mathrm{RSVMMAT-S}}$ respectively. Under the null hypothesis, $T_{\mathrm{RSVMMAT-A}}$ asymptotically follows a Cauchy distribution.

CHAPTER 4
SIMULATION STUDY AND APPLICATION

4.1 Introduction

In this chapter, we evaluate our proposed single variant association tests, RVMMAT and VMMAT, and compare them with the association tests assuming constant genetic effect and a Gaussian copula method that allows for heterogenous genetic effect. The type I error results demonstrate that the retrospective varying-coefficient test had better control of type I error when the trait model was misspecified, and was robust to various ascertainment schemes. Moreover, the retrospective varying-coefficient test was more powerful than the prospective test. We applied RVMMAT and VMMAT to the genome-wide association analysis of longitudinal measure of hypertension in MESA and identified hypertension-related genetic loci and pathways.

Furthermore, we also conduct the simulation studies for our proposed variant set association tests, RSVMMATs and SVMMATs, and compare them with existing variant set association tests assuming constant genetic effect. Our results demonstrate that all retrospective tests are robust to the ascertainment bias and model misspecification and are more powerful than prospective tests. The usage of the method is illustrated by applying RSVMMATs and SVMMATs to a genome-wide association analysis of longitudinal measure of hypertension in the MESA and identification of some hypertension-related genes and pathways.

4.2 Simulation studies for common SNPs

We conducte simulation studies to assess the type I error and power of VMMAT and RVMMAT, and compare them to a Gaussian copula method with weighted scores that allows for heterogenous

genetic effect [63] and the two association tests that assume constant genetic effect, GMMAT [64] and RGMMAT [16]. In all simulations, VMMAT and RVMMAT were implemented with cubic smoothing splines.The Gaussian copula method was implemented with a binomial marginal model with the logit link function and an AR(1) structure in the Gaussian copula correlation matrix [63]. VMMAT and RVMMAT are designed to detect time-varying genetic effect between a genetic variant and the longitudinal binary trait. Because we test one variant at a time, these methods tend to have limited power for rare variants and are more appropriate for common variants. The performance of all methods was evaluated on common variants in simulation studies. We considered two trait models and three ascertainment schemes to evaluate the robustness of VMMAT and RVMMAT in the presence of model misspecification and ascertainment.

4.2.1 Simulation settings

First, we simulated 10,000 chromosomes over a 1 Mb region using a coalescent model to mimic the recombination rates and linkage disequilibrium (LD) pattern of the European population [65, 66]. Then we randomly selected 1,000 non-causal SNPs with MAF > 0.05. In addition, we simulated two causal SNPs that were assumed to influence the trait value with epistasis. In each simulation setting, we generated binary phenotypes at five time points for a given sample size with 1,000 replicates. In the type I error experiments, for each phenotype dataset, we tested the time-varying genetic effect at the 1,000 non-causal SNPs. In total, 10^6 test results were used for the type I error assessment. We tested time-varying genetic effect at the first of the two causal SNPs in the power simulations while the untested SNPs were not included as covariates in the model. We evaluated power using 1,000 simulated phenotype datasets.

We simulated binary phenotypes under two types of trait models at five time points. We further assumed that the phenotype was influenced by an epistatic interaction of two unlinked causal SNPs.

The first type is a logistic mixed model, specified by:

$$Y_{ij}|\boldsymbol{X}_{ij}, G_{i(1)}, G_{i(2)}, a_i, r_{ij} \sim \text{Bernoulli}(\mu_{ij}),$$

$$\text{logit}(\mu_{ij}) = -1.9 + 0.2j + \gamma_1(t_{ij})\mathbb{I}_{\{G_{i(1)}>0, G_{i(2)}>0\}} + 0.5X_{ij(1)} + 0.5X_{i(2)} + a_i + r_{ij},$$

$$i = 1, \ldots, n; \quad j = 1, \ldots, 5,$$

where $\gamma_1(t_{ij})$ is a function encoding the effect of the causal SNPs, $G_{i(1)}$ and $G_{i(2)}$ are the genotypes of subject i at the two causal SNPs, $\mathbb{I}_{\{G_{i(1)}>0, G_{i(2)}>0\}}$ is an indicator function that equals to 1 when $G_{i(1)} > 0$ or $G_{i(2)} > 0$ and 0 otherwise, $X_{ij(1)}$ is generated from a multivariate normal distribution with a compound symmetry correlation matrix where the correlation is 0.5, $X_{i(2)}$ follow the Bernoulli distribution taking the value 1 with a probability of 0.5, a_i and r_{ij} are the subject-level time-independent and time-dependent random effects, respectively. We assumed $a_i \sim N(0,\ \sigma_a^2)$ and $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{i5})^T \sim MVN(\boldsymbol{0}, \sigma_r^2\boldsymbol{R})$, where $\boldsymbol{R}$ is specified by a 5×5 AR(1) correlation matrix. The two causal SNPs were assumed to be unlinked with MAFs 0.1 and 0.5, respectively. The variance components were set to $\sigma_a^2 = \sigma_r^2 = 0.64$ and $\tau = 0.7$.

The second type of trait model is a liability threshold model in which an underlying continuous liability determines the binary outcome value based on a threshold. Specifically, the phenotype Y_{ij} is determined by

$$Y_{ij} = 1 \quad \text{if } L_{ij} > 0,$$

$$\text{with } L_{ij} = -1.8 + 0.2j + \gamma_1(t_{ij})\mathbb{I}_{\{G_{i(1)}>0, G_{i(2)}>0\}} + 0.5X_{ij(1)} + 0.5X_{i(2)} + a_i + r_{ij} + e_{ij},$$

where L_{ij} is the underlying liability for subject i at time t_{ij}, and $e_{ij} \sim N(0, \sigma_e^2)$ represents independent noise, with $\sigma_e^2 = 1.96$. All other parameters are the same as those in the logistic mixed model.

In both trait models, we specified the intercept as a linear function of time $t_{ij} = j$, and the genetic effect as a logistic function $\gamma_1(t_{ij}) = \frac{\gamma}{\{1+\gamma\exp(8-2.4t_{ij})\}}$ [23]. For the type I error assessment,

the effect of the causal SNPs was set to $\gamma = 0.6$ in $\gamma_1(t_{ij})$. For the power evaluation, we considered a range of values for γ, where $\gamma = 0.6, 0.63, 0.66$, and 0.69. At the given parameter values, the prevalence of the event of interest ranges from 23.68% to 40.56% over time. The proportion of the phenotypic variance explained by the two causal SNPs ranges from 0.01% to 2.99% in the logistic mixed model and from 0.01% to 1.36% in the liability threshold model.

We considered three sampling designs as in [16]. In the "random" sampling, samples contain 2,000 subjects randomly selected from the population regardless of their phenotypes. In the "baseline" sampling, we sampled 1,000 subjects whose phenotype equal to 1 at baseline and 1,000 subjects whose phenotype equal to 0 at baseline. In the "sum" sampling, we stratified subjects into three strata based on the sum of events over time for each subject, where subjects in stratum 1 never experienced the event of interest, i.e., $\sum_j Y_{ij} = 0$, subjects in stratum 2 sometimes experienced the event, i.e., $0 < \sum_j Y_{ij} < n_i$, and subjects in stratum 3 always experienced the event, i.e., $\sum_j Y_{ij} = n_i$. We oversampled subjects with response variation over the course of the study and selected 100, 1,800, and 100 subjects from the three strata [67].

4.2.2 Simulation results

To assess type I error, we tested time-varying genetic effect at unlinked and unassociated SNPs. Empirical type I error was calculated as the proportion of simulations in which the p-value of the SNP is less than the nominal level α, for $\alpha = 0.01$, 0.001, and 0.0001. Table 4.1 gives the empirical type I error rates of RVMMAT and VMMAT, based on 10^6 replicates, under two trait models and three sampling designs. In most simulations, the type I error of RVMMAT was within the 95% confidence interval of the nominal levels. In contrast, the type I error of VMMAT in all simulation settings was much lower than the nominal level when $\alpha = 0.01$, 0.001, and 0.0001. It is well recognized that the DPQL approach underestimates variance components when data are sparse such as binary data [57, 38]. Even with bias correction, parameters estimated from the

Table 4.1: Empirical type I error of RVMMAT and VMMAT, based on 10^6 replicates

		Logistic Mixed Model			Liability Threshold Model		
Test	**Level**	**Random**	**Baseline**	**Sum**	**Random**	**Baseline**	**Sum**
RVMMAT	0.01	9.90×10^{-3}	1.01×10^{-2}	9.97×10^{-3}	1.02×10^{-2}	$\mathbf{9.70 \times 10^{-3}}$	1.00×10^{-2}
	0.001	9.52×10^{-4}	9.92×10^{-4}	$\mathbf{9.17 \times 10^{-4}}$	1.04×10^{-3}	1.00×10^{-3}	1.03×10^{-3}
	0.0001	9.80×10^{-5}	1.08×10^{-4}	9.40×10^{-5}	1.00×10^{-4}	1.01×10^{-4}	1.13×10^{-4}
VMMAT	0.01	$\mathbf{5.77 \times 10^{-3}}$	$\mathbf{6.45 \times 10^{-3}}$	$\mathbf{6.91 \times 10^{-3}}$	$\mathbf{5.73 \times 10^{-3}}$	$\mathbf{6.26 \times 10^{-3}}$	$\mathbf{7.20 \times 10^{-3}}$
	0.001	$\mathbf{4.34 \times 10^{-4}}$	$\mathbf{5.14 \times 10^{-4}}$	$\mathbf{5.33 \times 10^{-4}}$	$\mathbf{4.73 \times 10^{-4}}$	$\mathbf{5.10 \times 10^{-4}}$	$\mathbf{6.68 \times 10^{-4}}$
	0.0001	$\mathbf{3.70 \times 10^{-5}}$	$\mathbf{4.40 \times 10^{-5}}$	$\mathbf{5.70 \times 10^{-5}}$	$\mathbf{2.90 \times 10^{-5}}$	$\mathbf{5.00 \times 10^{-5}}$	$\mathbf{4.60 \times 10^{-5}}$

Rates outside of the 95% confidence interval are in bold.

DPQL function with penalty terms depend on the tuning parameter values. Thus, the prospective variance of the score $\boldsymbol{U}_0(\boldsymbol{c}_1)$ tends to be overestimated, producing a conservative test statistic. However, the retrospective variance of the score $\boldsymbol{U}_0(\boldsymbol{c}_1)$ does not depend on the estimation of variance components due to the tuning parameters in penalty terms so that the test statistic is less biased. These results suggest that the retrospective RVMMAT test had much better control of type I error and was robust to trait model misspecification and ascertainment, whereas the prospective VMMAT test was overly conservative.

To compare power, we considered four parameter values for γ to determine time-varying genetic effect at the two causal SNPs. Then we tested the association between the phenotype and the first causal SNP. Based on 1,000 replicates, we calculated the empirical power at the significance level 10^{-3}. Figure 4.1 demonstrates the power results of the five methods, RVMMAT, VMMAT, Copula, RGMMAT and GMMAT, under two trait models and three sampling designs. In all simulation settings, the two varying-coefficient tests consistently had higher power than the association tests assuming constant gene effect. The Gaussian copula method with heterogenous genetic effect had lower power than RVMMAT and VMMAT, but performed better than RGMMAT and GMMAT. Moreover, within the same type of tests, the retrospective test was more powerful than the prospective test. Both RVMMAT and VMMAT had similar power across the three sampling designs. In

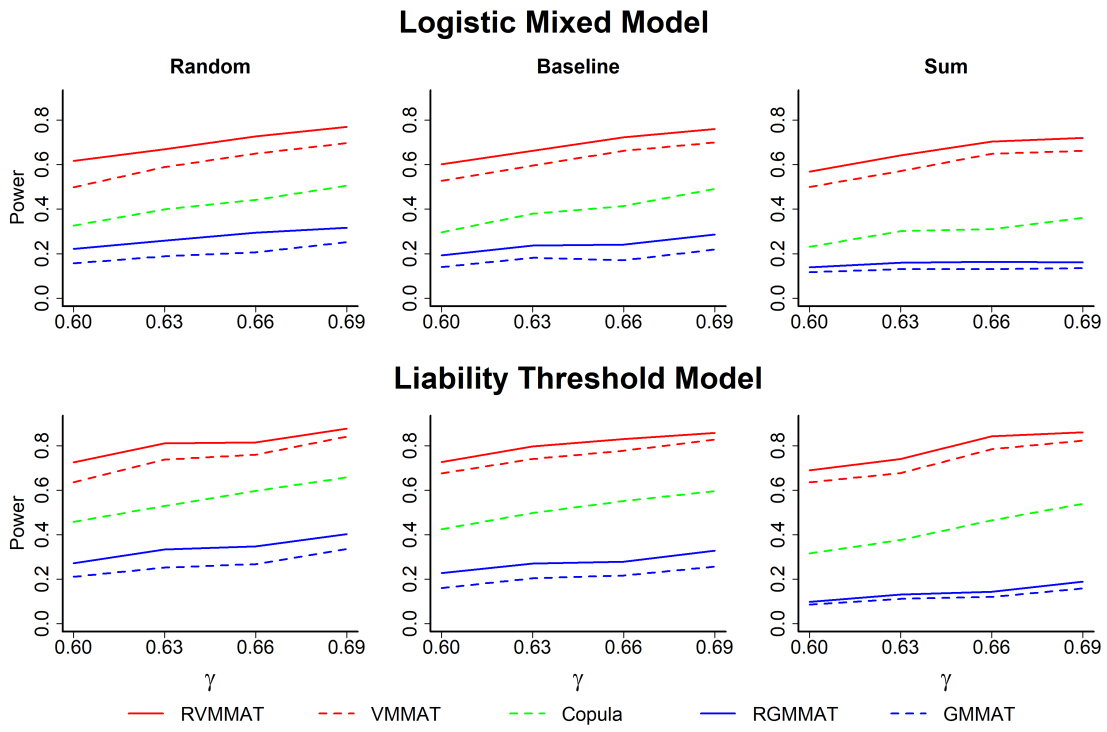

Figure 4.1: Empirical power of RVMMAT, VMMAT, Copula, RGMMAT and GMMAT are calculated from 1,000 replicates with level $\alpha = 10^{-3}$. Each replicate contains 2,000 individuals with observations at five time points. We simulated the phenotype under the logistic mixed model in the upper panel and under the liability threshold model in the lower panel. Power results are demonstrated in three ascertainment schemes: random, baseline, and sum.

contrast, Copula, RGMMAT and GMMAT had lower power under the sum sampling in both trait models. The power gain of the varying-coefficient tests was more prominent over the association tests assuming constant genetic effect in the presence of ascertainment. These results suggest that RVMMAT was the most powerful test and outperformed the association tests assuming constant genetic effect.

4.3 Application to MESA data for common SNPs

We illustrated the utility of our proposed methods by analyzing a GWAS dataset of hypertension in MESA [68]. MESA is a large longitudinal study of subclinical cardiovascular disease (CVD) whose primary objective is to understand the pathogenesis of atherosclerosis and other CVD. We analyzed

longitudinal hypertension assessed at five time points on 6,429 participants. Among them, 39.3% are White, 26.1% are African American, 22.5% are Hispanic, and 12.1% are Asian. The proportion of case subjects at each time point ranges from 44.6% ($n = 2,864$) to 59.5% ($n = 2,608$), and the missing rate at each time point ranges from 0 to 31.6%.

Samples were genotyped using the Affymetrix Human SNP Array 6.0. After data cleaning, there were 6,428 subjects available for genotype imputation. Using the 1000 Genomes Phase 3 data as a reference panel, we applied IMPUTE2 [69] for imputation. We excluded the subjects who did not meet either of the following criteria: (1) proportion of successfully imputed SNPs $> 95\%$ and (2) empirical inbreeding coefficient < 0.05. Based on the above criteria, 6,424 subjects were retained in the downstream analysis, with 3,057 males and 3,367 females, of whom 2,527 are white, 1,673 are African American, 1,449 are Hispanic, and 775 are Asian. There are 2,227 subjects who had no hypertension during the study period, 1,807 subjects who were sometimes hypertensive, i.e., exhibited response variation, and 2,390 subjects who were always hypertensive over the course of the study. We then tested Hardy-Weinberg equilibrium at each SNP within each population. We included SNPs meeting all of the following quality-control conditions: (1) missing rate $< 5\%$, (2) Hardy-Weinberg χ^2 statistic p-value $> 10^{-6}$, and (3) MAF $> 1\%$. Finally, 6,155,404 SNPs were examined in the downstream association testing.

4.3.1 Analysis of time-varying genetic effect

We performed genome-wide tests of time-varying genetic effect on hypertension using RVMMAT and VMMAT with cubic smoothing splines in the MESA sample. Age at baseline, sex, and the top ten principal components (PCs) were included as time-invariant covariates in the analysis. The top ten PCs were calculated using the LD pruned SNPs with MAF > 0.05 to control for population structure. Since hypertension was assessed in year 2000, 2002, 2004, 2005 and 2010, we coded time at the five time points as 0, 0.2, 0.4, 0.5 and 1, respectively. We also applied the Gaussian

copula method with heterogenous genetic effect, adjusting for the same covariates. To compare the performance of the varying-coefficient tests with the association tests assuming constant genetic effect, we applied RGMMAT and GMMAT to the analysis of hypertension, adjusting for age at baseline, sex, time, and the top ten PCs.

The two retrospective tests, RVMMAT and RGMMAT, showed no evidence of inflation in the quantile-quantile (Q-Q) plot. The genomic control inflation factors were 0.905 and 0.976, respectively. The prospective VMMAT test was overly conservative, with a genomic control factor of 0.774, consistent with the observed deflation in the type I error simulations. The genomic control inflation factor was 0.838 for GMMAT.

None of the SNPs reached genome-wide significance at the p-value threshold of 5×10^{-8} that is widely used in GWAS. Table 4.2 reports the top SNPs for which at least one of the tests gives a p-value $< 5 \times 10^{-7}$. The smallest p-values of these eight SNPs were mostly generated by RVMMAT, except at the last two SNPs. VMMAT generated much larger p-values than RVMMAT due to its conservativeness, while RGMMAT and GMMAT had comparable results. The Gaussian copula method produced p-values comparable to VMMAT, except at the last two SNPs. A cluster of six SNPs in LD ($r^2 > 0.97$), rs145659245, rs58265184, rs57719815, rs60197637, rs61327798, and rs142890225, located at 4p15, showed time-varying genetic effect on hypertension by RVMMAT (p-value= $6.78 \times 10^{-8} - 2.85 \times 10^{-7}$). Figure 4.2A demonstrated the estimated genetic effect over time at these SNPs where the estimated effect at each time point was obtained by using the observed trait values at that time point only. A straight line was used to connect the estimated values at two adjacent time points. We observed an increasing and then decreasing trend in genetic effects on hypertension across the five time points. However, RGMMAT and GMMAT lost power and generated large p-values by assuming constant genetic effect. These SNPs are in an intron of the gene *PROM1*, encoding a pentaspan transmembrane glycoprotein, which was

Table 4.2: SNPs with p-value $< 5 \times 10^{-7}$ in at least one of the tests in the MESA data

Chr	Gene Region	SNP	Position	MAF	RVMMAT	VMMAT	Copula	RGMMAT	GMMAT
4	*PROM1*	rs145659245	16,060,553	0.013	$\mathbf{6.78 \times 10^{-8}}$	4.19×10^{-6}	9.77×10^{-6}	7.67×10^{-4}	1.78×10^{-3}
		rs58265184	16,061,151	0.014	$\mathbf{2.18 \times 10^{-7}}$	7.70×10^{-5}	2.79×10^{-5}	2.51×10^{-3}	4.85×10^{-3}
		rs57719815	16,063,652	0.013	$\mathbf{2.85 \times 10^{-7}}$	8.04×10^{-5}	3.09×10^{-5}	1.65×10^{-3}	3.27×10^{-3}
		rs60197637	16,063,659	0.013	$\mathbf{2.85 \times 10^{-7}}$	8.04×10^{-5}	3.09×10^{-5}	1.65×10^{-3}	3.27×10^{-3}
		rs61327798	16,063,661	0.013	$\mathbf{2.85 \times 10^{-7}}$	8.04×10^{-5}	3.09×10^{-5}	1.65×10^{-3}	3.27×10^{-3}
		rs142890225	16,065,544	0.013	$\mathbf{2.85 \times 10^{-7}}$	8.04×10^{-5}	3.10×10^{-5}	1.65×10^{-3}	3.27×10^{-3}
17	*LRRC37B*	rs374012917	30,403,054	0.038	2.21×10^{-5}	1.95×10^{-4}	$\mathbf{1.02 \times 10^{-7}}$	2.31×10^{-6}	4.95×10^{-6}
18	*WDR7*	rs72930733	54,641,870	0.011	1.80×10^{-6}	7.06×10^{-5}	7.72×10^{-6}	$\mathbf{2.55 \times 10^{-7}}$	4.63×10^{-6}

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

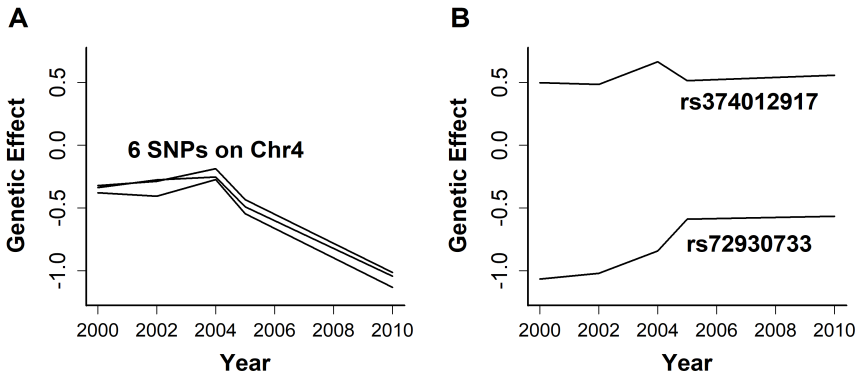

Figure 4.2: Estimated genetic effect of the top 8 SNPs on hypertension at each of the five time points. (A) six SNPs on chromosome 4; (B) two SNPs on chromosomes 17 and 18, respectively.

reported to be associated with pulse pressure [70]. The smallest p-value for rs374012917, located on chromosome 17, was generated by the Gaussian copula method (p-value $= 1.02 \times 10^{-7}$). As the estimated genetic effects at the five time points were relatively stable for this SNP (Figure 4.2B), RGMMAT and GMMAT generated slightly larger p-values. There was also evidence of association between hypertension and rs72930733 (p-value $= 2.55 \times 10^{-7}$). Although RVMMAT did not give the smallest p-value for this SNP, its p-value was slightly larger than that of RGMMAT, mostly due to the increasing trend in genetic effect (Figure 4.2B). This SNP is in an intron of the gene *WDR7*, located at 18q21. Two hypertension GWASs identified an association between *WDR7* and systolic blood pressure [70, 71].

We further assessed the model fitting of cubic smoothing splines on the top SNPs in Table 4.2 using deviance and goodness-of-fit p-value. All p-values were large, suggesting that there was no

Table 4.3: Assessment of model fitting with cubic smoothing splines at the top SNPs in the MESA data

Chr	Gene Region	SNP	Position	MAF	Deviance	Goodness-of-fit P-value
4	*PROM1*	rs145659245	16,060,553	0.013	10467.93	1
		rs58265184	16,061,151	0.014	10468.13	1
		rs57719815	16,063,652	0.013	10469.79	1
		rs60197637	16,063,659	0.013	10469.79	1
		rs61327798	16,063,661	0.013	10469.79	1
		rs142890225	16,065,544	0.013	10469.79	1
17	*LRRC37B*	rs374012917	30,403,054	0.038	10495.83	1
18	*WDR7*	rs72930733	54,641,870	0.011	10493.05	1

evidence of lack of fit (Table 4.3). We also checked the deviance residuals of the cubic smoothing splines model applied to the top SNPs. The deviance residuals range from -2.46 to 2.19, suggesting that cubic smoothing splines fit the data adequately.

4.3.2 Pathway analysis

We then performed functional pathway analysis using the MetaCore^{TM} software to identify enriched pathways related to hypertension. The top SNPs for which at least one of the tests had a p-value $< 2 \times 10^{-4}$ were included in the analysis. Fisher's exact test was used to determine whether the SNP list was enriched for a functional pathway. At the false discovery rate (FDR) < 0.05, we identified two significant pathways that were associated with G-protein signaling and DNA damage. The first one is the G-protein signaling pathway related to Rac1 activation (p-value $= 6.12 \times 10^{-5}$, FDR $= 1.65 \times 10^{-2}$). Rac1 participates in the control of blood pressure through multiple mechanisms in the arterial wall and the central nervous system [72]. Importantly, a role for Rac1 in atherosclerosis and cardiac hypertrophy has been established in response to the administration of statins in clinical trials [73]. Animal studies indicated that Rac1 is essential for endothelium-dependent vasomotor response, the redox state of blood vessels and homeostasis of blood pressure [74, 75, 76]. The second pathway is the DNA damage pathway related to the ataxia-telangiectasia

mutated (ATM) kinase activation (p-value = $4.98 \times 10^{-4}, FDR = 4.48 \times 10^{-2}$). Emerging evidence has demonstrated that accumulated DNA damage and subsequent repair pathways play a crucial role in the initiation and progression of cardiovascular disorders, such as atherosclerosis and maladaptive cardiac hypertrophy [77, 78, 79, 80]. ATM-mediated phosphorylation plays cardinal roles in response to genomic stress to preserve cellular homeostasis. DNA double-strand breaks trigger ATM activation, which mediates DNA damage response and regulate cardiac remodeling, inflammation, and systolic function, eventually promoting heart failure development [81, 82].

4.4 Stimulation studies for rare SNPs

To evaluate the performance of RSVMMAT-A, S, B and SVMMAT-A, S, B in the analyses of longitudinal binary phenotype, we conducted simulation studies to assess the empirical type I error and compare power of the tests with two association tests assuming constant genetic effects, SMMAT-E [12] and RSMMAT-E [53]. In all simulations, we used natural cubic smoothing splines for the time-varying genetic effects. Two trait models and three sampling schemes were considered to evaluate the robustness of RSVMMATs and SVMMATs to model misspecification and ascertainment.

4.4.1 Simulation settings

We simulated a set of 10^6 genotype-phenotype data at five time points under the assumption of no genetic effects on the trait values. We generated 10,000 chromosomes over a 1 Mb region using package COSI2 [66]. The package produces simulated data that closely resemble empirical data in allele frequency, linkage disequilibrium (LD), and population differentiation of the European population [65]. We generated sequence data selected from 30kb regions in each set. We further selected rare variants (MAF<0.05) in those regions and then removed variants with less than 8 copies of the minor allele.

Binary phenotype data were simulated under two traits models, a logistic mixed effects model

and a liability threshold model, and three sampling designs labeled as "random", "baseline", and "sum". We first considered a logistic mixed effects model, given by:

$$y_{ij}|X_{ij}, a_i, r_{ij} \sim \text{Bernoulli}(\mu_{ij}),$$

$$logit\left(\mu_{ij}\right) = -1.9 + 0.5X_{ij1} + 0.5X_{i2} + 0.2t_{ij} + a_i + r_{ij},$$

where $X_{ij(1)}$ is a continuous, time-varying covariate generated from a multivariate normal distribution with a compound symmetry correlation matrix where the correlation is 0.5, $X_{i(2)}$ is a binary, time-invariant covariate taking values 0 or 1 with a probability of 0.5, the subject random effect a_i and the individual-specific time-dependent random effect r_{ij} are assumed to follow $N\left(0, 0.64\right)$ and $\boldsymbol{r}_i = \left(r_{i1}, \ \ldots, r_{i5}\right)^T \sim N\left(0, 0.64\boldsymbol{R}\right)$ respectively with $\boldsymbol{R}$ being a 5×5 correlation matrix specified by the AR(1) structure with a correlation coefficient $\tau = 0.7$, and $t_{ij} = \text{j}$.

A liability threshold model that we used as a second trait model assigns the binary outcome based on a threshold that is determined by an underlying continuous liability. Specifically, the phenotype y_{ij} is given by:

$$y_{ij} = 1 \text{ if } L_{ij} > 0,$$

with

$$L_{ij} = -1.8 + 0.5X_{ij1} + 0.5X_{i2} + 0.2t_{ij} + a_i + r_{ij} + e_{ij},$$

and $e_{ij} \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 0.64$.

In each setting, we generated 1,000 sets of phenotypes at five-time points and 1,000 noncausal variant sets. In total, 10^6 replicates across 1,000 phenotype datasets were used for the type I error evaluation.

For power comparison, we considered a logistic mixed effects model, given by:

$$y_{ij}|X_{ij}, G_{i1}, .., G_{is}, a_i, r_{ij} \sim \text{Bernoulli}(\mu_{ij}),$$

where

$$logit\left(\mu_{ij}\right) = -1.9 + 0.5X_{ij1} + 0.5X_{i2} + 0.2t_{ij} + \sum_{k=1}^{s} G_{ik}\gamma_{\mathrm{k}}\left(t_{ij}\right) + a_i + r_{ij},$$

A liability threshold model that we used as a second trait model assigns the binary outcome based on a threshold that is determined by an underlying continuous liability. Specifically, the phenotype y_{ij} is given by:

$$y_{ij} = 1 \text{ if } L_{ij} > 0,$$

with

$$L_{ij} = -1.8 + 0.5X_{ij1} + 0.5X_{i2} + 0.2t_{ij} + \sum_{k=1}^{s} G_{ik}\gamma_{\mathrm{k}}\left(t_{ij}\right) + a_i + r_{ij} + e_{ij},$$

and $e_{ij} \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 0.64$.

In both trait models, a logistic function of $\gamma_{\mathrm{k}}\left(t_{ij}\right) = \gamma/\{1 + \gamma\exp\left(10 - 2.4t_{ij}\right)\}|\log_{10}MAF_k|$ is used as the form of the varying coefficient [23] and $\gamma_{\mathrm{k}}\left(t_{ij}\right) = \gamma|\log_{10}MAF_k|$ is used as the form of the time-invariant effect. We considered $\gamma = 0.08$ in the liability threshold model and $\gamma = 0.12$ in the logistic mixed effects model for time-varying effect. The parameter γ of the time-invariant effect is reduced to 1/3 of γ in time-varying effect, in other words, $\gamma = 0.0267$ in the liability threshold model and $\gamma = 0.04$ in the logistic mixed effects model for the time-invariant effect. We considered two cases for causal SNPs set. First, we selected 60% causal SNPs with time-varying effects. Second, we selected 30% causal SNPs with time-varying effects and 30% causal SNPs with time-invariant effects.

For the three different sampling schemes, the "random" sample contains 4,000 randomly picked individuals from the population regardless of their phenotypes. The "baseline" sample consists of 2,000 case subjects and 2,000 control subjects based on their phenotypic values at baseline only. In the "sum" sampling scheme, we created three groups so that subjects in group 1 never experienced the phenotypic event, *i.e.*, $\sum_j y_{ij} = 0$, subjects in group 2 experienced the event occasionally, *i.e.*, $0 < \sum_j y_{ij} < n_i$, and subjects in group 3 experienced the event all the time, *i.e.*, $\sum_j y_{ij} = n_i$. We

oversampled subjects with response variation over the course of the study by selecting 200, 3,600, and 200 individuals from the three groups according to the outcome-dependent sampling design for longitudinal binary data [67].

For power comparison, we tested the association between the trait and the causal variant set and the empirical power was calculated at the significance level of 10^{-4} based on 1,000 simulated replicates.

4.4.2 Simulation results

Empirical type I error of the RSVMMAT-A, S, B, RSMMAT-E, SVMMAT-A, S, B, and SMMAT-E tests are shown in Table 4.4. The results are based on 10^6 replicates at the nominal level of 0.01, 0.001, and 0.0001. In all simulations, type I error rates of the four retrospective tests, RSVMMAT-A, S, B, and RSMMAT-E, are well controlled at any of the nominal levels considered. In contrast, the prospective SVMMAT-A, S, B, and SMMAT-E tests had deflated type I error at all nominal levels and in all settings. These results suggest that the retrospective tests, RSVMMAT-A, S, and B, are robust to trait model misspecification as well as sample ascertainment, whereas the three prospective tests are somewhat conservative.

Figures 4.3 and 4.4 demonstrate the power results in samples of 4,000 individuals under the two trait models with different causal genetic effect proportions, different causal genetic effect directions, and three different sampling schemes at the significance level of 10^{-4}. RSVMMAT-B and RSVMMAT-A are the most powerful tests in 100% positive causal genetic effects. RSVMMAT-S and RSVMMAT-A are the most powerful tests in 80% and 50% positive causal genetic effects. The retrospective tests are more powerful than the corresponding prospective tests in all settings. RSVMMAT-S and RSVMMAT-A are more powerful than RSMMAT-E in all settings. RSVMMAT-B is more powerful than RSMMAT-E except the cases of logistic mixed model with 50% positive causal genetic effects.

Table 4.4: Empirical type I error of the longitudinal tests, based on 10^6 replicates

		Logistic Mixed Model			Liability Threshold Model		
Test	**Level**	**Random**	**Baseline**	**Sum**	**Random**	**Baseline**	**Sum**
RSVMMAT-A	0.01	1.0×10^{-2}	1.1×10^{-2}	1.1×10^{-2}	1.1×10^{-2}	1.1×10^{-2}	1.0×10^{-2}
	0.001	1.1×10^{-3}	1.1×10^{-3}	1.1×10^{-3}	1.1×10^{-3}	1.1×10^{-3}	1.2×10^{-3}
	0.0001	1.4×10^{-4}	1.3×10^{-4}	1.3×10^{-4}	1.4×10^{-4}	1.3×10^{-4}	1.5×10^{-4}
RSVMMAT-S	0.01	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.1×10^{-2}	1.0×10^{-2}
	0.001	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-3}
	0.0001	1.6×10^{-4}	1.5×10^{-4}	1.6×10^{-4}	1.5×10^{-4}	1.5×10^{-4}	1.6×10^{-4}
RSVMMAT-B	0.01	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}
	0.001	9.6×10^{-4}	9.6×10^{-4}	1.0×10^{-3}	9.8×10^{-4}	9.9×10^{-4}	1.1×10^{-3}
	0.0001	1.1×10^{-4}	9.4×10^{-5}	9.6×10^{-5}	1.1×10^{-4}	1.1×10^{-4}	1.1×10^{-4}
SVMMAT-A	0.01	5.9×10^{-3}	6.0×10^{-3}	6.0×10^{-3}	4.9×10^{-3}	5.4×10^{-3}	8.2×10^{-3}
	0.001	5.2×10^{-4}	4.9×10^{-4}	5.3×10^{-4}	3.9×10^{-4}	4.5×10^{-4}	8.4×10^{-4}
	0.0001	5.8×10^{-5}	4.0×10^{-5}	5.3×10^{-5}	3.2×10^{-5}	4.5×10^{-5}	1.1×10^{-4}
SVMMAT-S	0.01	5.2×10^{-3}	5.2×10^{-3}	5.2×10^{-3}	4.0×10^{-3}	4.6×10^{-3}	7.6×10^{-3}
	0.001	4.8×10^{-4}	5.0×10^{-4}	5.1×10^{-4}	3.6×10^{-4}	4.3×10^{-4}	8.6×10^{-4}
	0.0001	5.5×10^{-5}	4.6×10^{-5}	6.2×10^{-5}	3.5×10^{-5}	5.0×10^{-5}	1.2×10^{-4}
SVMMAT-B	0.01	6.6×10^{-3}	6.8×10^{-3}	6.7×10^{-3}	5.8×10^{-3}	6.1×10^{-3}	8.5×10^{-3}
	0.001	5.4×10^{-4}	5.2×10^{-4}	5.5×10^{-4}	4.2×10^{-4}	4.5×10^{-4}	8.1×10^{-4}
	0.0001	4.3×10^{-5}	3.3×10^{-5}	4.8×10^{-5}	4.0×10^{-5}	4.5×10^{-5}	9.1×10^{-5}

4.5 Application to MESA data for rare SNPs

We used the MESA dataset to compare the performance of the proposed three retrospective varying coefficients tests, RSVMMAT-A, S, and B, with five other methods, including four prospective tests, SVMMAT-A, S, B, and SMMAT-E, and one retrospective tests, RSMMAT-E. After data cleaning, the MESA data remains array-based genotype data on 830,535 SNPs for 6,424 individuals. We used IMPUTE2 [69] for data imputation and the 1000 Genomes Phase 3 data as a reference panel. We checked the Hardy-Weinberg equilibrium within each population and SNPs that are included in the analysis met the following quality-control (QC) conditions: (1) call rate $> 95\%$, (2) Hardy-Weinberg chi-square statistic p-value $> 10^{-6}$, and (3) alleles less than 12 copies. After imputation and QC process, we analyzed a final set of 18,670 protein coding genes on 6,424 individuals. Age at the baseline, sex, cubic polynomial of time, and top five principal components of the genetic

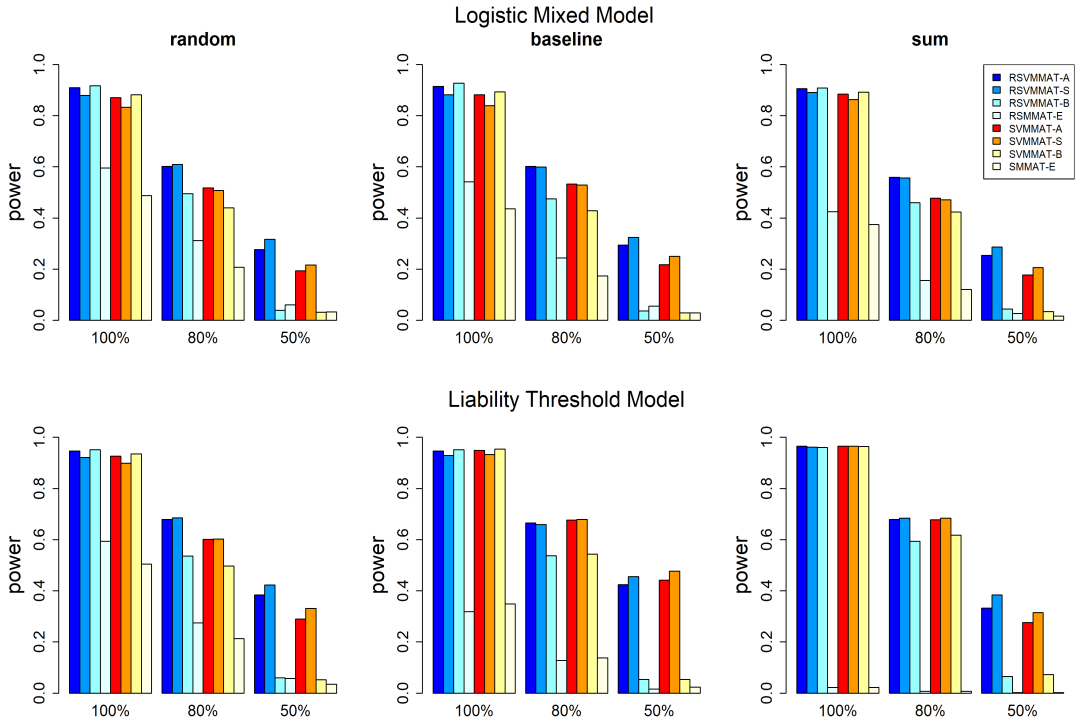

Figure 4.3: Empirical power of the longitudinal tests under two trait models and three sample ascertainment schemes with 60% time-varying causal genetic effects.

relatedness matrix, calculated using LD pruned SNPs with MAF > 0.05, were included as covariates. The MESA data were collected at the years of 2000, 2002, 2005, 2007, and 2010 which are coded as 0, 1/5, 1/2, 7/10, and 1, respectively in our study. The proportion of case subject at each visit ranges from 44.6% ($n = 2,864$) to 59.5% ($n = 2,608$), and the missing rate at each visit is less than 31.6%.

4.5.1 Results in the MESA data

Hypertension status at five time points was used as a longitudinal binary trait. For the retrospective tests, no evidence of inflation was presented with genomic control inflation factors of 0.78, 0.79, 0.87, and 0.91 for RSVMMAT-A, S, B, and RSMMAT-E, respectively. The genomic control inflation factors are 0.48, 0.43, 0.81, and 0.59 for SVMMAT-A, S, B, and SMMAT-E, respectively. The prospective tests showed some evidence of deflation and it is consistent with their deflated type I

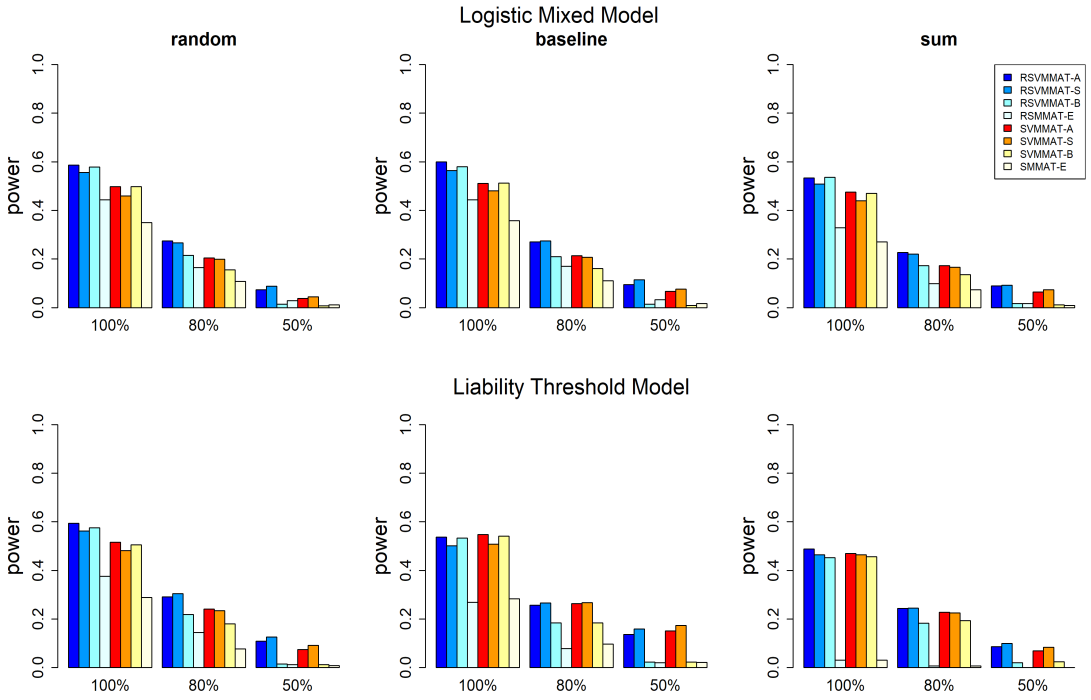

Figure 4.4: Empirical power of the longitudinal tests under two trait models and three sample ascertainment schemes with 30% time-varying causal genetic effects and 30% time-invariant causal genetic effects.

error rates in simulation studies.

We report, in Table 4.5, the genes for which at least one of the longitudinal tests gives a P-value $< 1 \times 10^{-4}$. Among them, retrospective time-varying tests, RSVMMAT-B and S produced all of the smallest P-values. The gene *TNXB* was reported to be associated with diastolic blood pressure, systolic blood pressure, and hypertension [83, 84]. What's more, an animal study has shown that TNX plays a crucial role in blood vessel formation [85]. Further, some researchers have hypothesized that TNX-deficient patients tend to protect against high blood pressure as patients with deficient Tenascin-X levels usually have cardiac arrhythmias and orthostatic hypotension *TNXB* [86].

Table 4.5: Top genes with p-value $< 10^{-4}$ in at least one of the longitudinal tests in the MESA data.

Gene	#SNPs	Chr.	RSVMMAT-A	RSVMMAT-S	RSVMMAT-B	RSMMAT-E	SVMMAT-A	SVMMAT-S	SVMMAT-B	SMMAT-E
SLFN12	88	17	4.88×10^{-5}	$\mathbf{2.44 \times 10^{-5}}$	8.33×10^{-1}	3.66×10^{-1}	3.14×10^{-3}	1.55×10^{-3}	8.79×10^{-1}	6.17×10^{-1}
GTF3C5	215	9	5.33×10^{-5}	$\mathbf{2.74 \times 10^{-5}}$	1.05×10^{-3}	5.47×10^{-1}	1.40×10^{-3}	8.53×10^{-4}	3.89×10^{-3}	6.34×10^{-1}
TNXB	393	6	9.59×10^{-5}	6.73×10^{-4}	$\mathbf{5.16 \times 10^{-5}}$	5.48×10^{-1}	1.46×10^{-3}	8.78×10^{-3}	7.96×10^{-4}	7.14×10^{-1}
CCSER2	1114	10	7.03×10^{-5}	8.11×10^{-5}	$\mathbf{6.20 \times 10^{-5}}$	2.15×10^{-1}	9.51×10^{-4}	1.38×10^{-3}	7.24×10^{-4}	4.17×10^{-1}

The smallest p-value among all tests at the given genes are in bold.

4.5.2 Pathway analysis

Using MetaCore^{TM}, we performed pathway analysis on genes for which at least one of the longitudinal tests gave a P-value $< 5 \times 10^{-3}$. We identified four significant pathways that are associated with hypertension. The first pathway is Rheumatoid arthritis (general schema) (p-value $= 5.95 \times 10^{-5}$, FDR $= 4.30 \times 10^{-3}$). A study found that cardiovascular mortality and hypertension are increased among patients with Rheumatoid arthritis[87]. The second pathway is systemic lupus erythematosus (SLE) genetic marker-specific pathways in T cells (p-value $= 6.87 \times 10^{-4}$, FDR $= 4.32 \times 10^{-3}$). The third pathway is Role of B cells in SLE (p-value $= 1.93 \times 10^{-3}$, FDR $= 4.88 \times 10^{-3}$). An animal study shows that using anti-CD20 antibody prevents the development of hypertension in a mouse model of SLE [88].

4.6 Code availability

R package implementing RVMMAT and VMMAT can be found at https://github.com/ZWang-Lab/RVMMAT. R package implementing RSVMMATs and SVMMATs can be found at https://github.com/ZWang-Lab/RSVMMAT.

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

In genome-wide association analysis of longitudinal traits, modeling time-varying genetic effect can increase power for the detection of genes underlying the development and progression of complex diseases.

In Chapter 2, we developed RVMMAT, a GLMM-based, retrospective varying-coefficient association testing method for longitudinal binary traits. RVMMAT extends the existing association methods assuming constant effect over time to testing of time-varying effect on binary traits. RVMMAT is constructed based on the trait model allowing for time-varying genetic effect. The variance of the test statistic is assessed retrospectively by considering the conditional distribution of the genotype at the variant of interest, given phenotype and covariate information, under the null hypothesis of no association. RVMMAT has the following features: (1) it is computationally feasible for genetic studies with millions of variants, (2) it has well-controlled type I error in the presence of ascertainment and trait model misspecification, and (3) it can easily be fitted as a GLMM model using popular software such as R and SAS. We also propose VMMAT, a prospective varying-coefficient association test, for performance comparison.

The RVMMAT and VMMAT methods are designed for single-variant association analysis of longitudinal binary traits. However, single-variant association tests suffer from restricted power to detect association for rare variants in whole-genome sequencing studies. As many variants influence complex traits collectively, assessing joint effects from multiple variants by aggregating weak signals at the gene or pathway level holds great promise for the identification of novel genes underlying disease risks.

In Chapter 3, we developed RSVMMATs, GLMM-based, retrospective variant set association tests detecting time-varying genetic effects for a set of variants. RSVMMATs can also be viewed as an extension of set-based association tests assuming constant effects over time to testing time-varying effects on longitudinal binary traits. RSVMMATs are constructed based on GLMM with time-varying coefficients and consist of three variant set tests: a burden type test, RSVMMAT-B, a SKAT type test, RSVMMAT-S, and a combination of RSVMMAT-B and RSVMMAT-S using Cauchy combination test, RSVMMAT-A. We retrospectively assessed the variance of RSVMMATs by modeling the genotypes in a certain genetic region given the phenotype and covariates under the null hypothesis of no association between the genotypes and the phenotype.

Our simulation results demonstrated that RVMMAT maintained correct type I error under different trait models and ascertainment schemes, whereas VMMAT was overly conservative due to the biased estimation of variance in the penalized trait model. We further demonstrated that the retrospective RVMMAT test achieved the highest power among the five tests under all the trait models and ascertainment schemes considered in the simulations. Application of RVMMAT to the MESA longitudinal hypertension data identified three novel genes that were associated with hypertension. Among them, two genes are known to be associated with systolic blood pressure and pulse pressure. Moreover, we identified two significant pathways associated with longitudinal hypertension: the G-protein signaling pathway related to Rac1 activation and the DNA damage pathway related to ATM activation. Given the established role for Rac1 and ATM in atherosclerosis and cardiac hypertrophy, our findings suggest that RVMMAT can provide enhanced statistical power in detecting biologically relevant genetic loci that are associated with trait dynamics. A better understanding of temporal variation of trait values and time-varying genetic contribution may shed light on the genetic mechanisms influencing the temporal trend of diseases and complex traits.

For rare SNPs, our simulation results suggest that the retrospective association tests are robust to model misspecifications. RSVMMAT-B is the most powerful test when the genetic effects are time-varying and in same direction. While RSVMMAT-S is the most powerful test when the genetic effects are time-varying and in different directions. We applied RSVMMATs to the whole genome genotyping data of MESA to detect the genes that are associated with longitudinal hypertension. We identified four hypertension related genes while RSVMMAT-S detects two genes and RSVMMAT-B detects two genes. Among the four genes, one gene is reported to be associated with diastolic blood pressure, systolic blood pressure, and hypertension. In pathway analysis, we identified three significant pathways that are associated with rheumatoid arthritis and systemic lupus erythematosus in the analysis. Compared to existing set-based tests for longitudinal data, our simulation studies and real data analysis suggest that RSVMMATs gains statistical power and provide a better understanding of a group of genetic variants influencing the human traits over time.

Our current model focuses on the binary phenotype and can easily be extended to analyze multi-category data. Furthermore, less than 10% of total heritability linked with cardiovascular traits are collectively explained by genetic variation in GWAS. Lacking consideration of gene-environment (G×E) interactions is a important factor that contributes to the missing heritability. Thus we plan to extend current methods to detect the association between the longitudinal phenotype and G×E interactions.

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, pp. 1–21, 2021.

[2] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[3] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009.

[4] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *The American Journal of Human Genetics*, vol. 83, no. 3, pp. 311–321, 2008.

[5] B. M. Neale, M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly, "Testing for an unusual distribution of rare variants," *PLoS Genetics*, vol. 7, no. 3, p. e1001322, 2011.

[6] 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.

[7] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, E. L. P. Team, D. C. Christiani, M. M. Wurfel, and X. Lin, "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," *The American Journal of Human Genetics*, vol. 91, no. 2, pp. 224–237, 2012.

[8] 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.

[9] D. Jiang and M. S. McPeek, "Robust rare variant association testing for quantitative traits in samples with related individuals," *Genetic Epidemiology*, vol. 38, no. 1, pp. 10–20, 2014.

[10] J. Sun, Y. Zheng, and L. Hsu, "A unified mixed-effects model for rare-variant association in sequencing studies," *Genetic Epidemiology*, vol. 37, no. 4, pp. 334–344, 2013.

[11] W. Pan, J. Kim, Y. Zhang, X. Shen, and P. Wei, "A powerful and adaptive association test for rare variants," *Genetics*, vol. 197, no. 4, pp. 1081–1095, 2014.

[12] H. Chen, J. E. Huffman, J. A. Brody, C. Wang, S. Lee, Z. Li, S. M. Gogarten, T. Sofer, L. F. Bielak, and J. C. Bis, "Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies," *The American Journal of Human Genetics*, vol. 104, no. 2, pp. 260–274, 2019.

[13] Y. Liu, S. Chen, Z. Li, A. C. Morrison, E. Boerwinkle, and X. Lin, "Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies," *The American Journal of Human Genetics*, vol. 104, no. 3, pp. 410–421, 2019.

[14] N. A. Furlotte, E. Eskin, and S. Eyheramendy, "Genome-wide association mapping with longitudinal data," *Genetic Epidemiology*, vol. 36, no. 5, pp. 463–471, 2012.

[15] K. Sikorska, F. Rivadeneira, P. J. Groenen, A. Hofman, A. G. Uitterlinden, P. H. Eilers, and E. Lesaffre, "Fast linear mixed model computations for genome-wide association studies with longitudinal data," *Statistics in Medicine*, vol. 32, no. 1, pp. 165–180, 2013.

[16] W. Wu, Z. Wang, K. Xu, X. Zhang, A. Amei, J. Gelernter, H. Zhao, A. C. Justice, and Z. Wang, "Retrospective association analysis of longitudinal binary traits identifies important loci and pathways in cocaine use," *Genetics*, vol. 213, no. 4, pp. 1225–1236, 2019.

[17] C. M. Sitlani, K. M. Rice, T. Lumley, B. McKnight, L. A. Cupples, C. L. Avery, R. Noordam, B. H. Stricker, E. A. Whitsel, and B. M. Psaty, "Generalized estimating equations for genome-wide association studies using longitudinal phenotype data," *Statistics in Medicine*, vol. 34, no. 1, pp. 118–130, 2015.

[18] K. Das, J. Li, Z. Wang, C. Tong, G. Fu, Y. Li, M. Xu, K. Ahn, D. Mauger, and R. Li, "A dynamic model for genome-wide association studies," *Human Genetics*, vol. 129, no. 6, pp. 629–639, 2011.

[19] D. Londono, K.-m. Chen, A. Musolf, R. Wang, T. Shen, J. Brandon, J. A. Herring, C. A. Wise, H. Zou, and M. Jin, "A novel method for analyzing genetic association with longitudinal phenotypes," *Statistical Applications in Genetics and Molecular Biology*, vol. 12, no. 2, pp. 241–261, 2013.

[20] O. D. Meirelles, J. Ding, T. Tanaka, S. Sanna, H.-T. Yang, D. B. Dudekula, F. Cucca, L. Ferrucci, G. Abecasis, and D. Schlessinger, "SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits," *European Journal of Human Genetics*, vol. 21, no. 6, pp. 673–679, 2013.

[21] J. Bryois, A. Buil, P. G. Ferreira, N. I. Panousis, A. A. Brown, A. Viñuela, A. Planchon, D. Bielser, K. Small, and T. Spector, "Time-dependent genetic effects on gene expression implicate aging processes," *Genome Research*, vol. 27, no. 4, pp. 545–552, 2017.

[22] W. Chu, R. Li, and M. Reimherr, "Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data," *The Annals of Applied Statistics*, vol. 10, no. 2, p. 596, 2016.

[23] Y. Gong and F. Zou, "Varying coefficient models for mapping quantitative trait loci using recombinant inbred intercrosses," *Genetics*, vol. 190, no. 2, pp. 475–486, 2012.

[24] J. Liu, R. Li, and R. Wu, "Feature selection for varying coefficient models with ultrahigh-dimensional covariates," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 266–274, 2014.

[25] L. Wang, H. Li, and J. Z. Huang, "Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1556–1569, 2008.

[26] T. Hastie and R. Tibshirani, "Varying-coefficient models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 55, no. 4, pp. 757–779, 1993.

[27] J. Fan and W. Zhang, "Statistical methods with varying coefficient models," *Statistics and its Interface*, vol. 1, no. 1, p. 179, 2008.

[28] Y. Lu and R. Zhang, "Smoothing spline estimation of generalised varying-coefficient mixed model," *Journal of Nonparametric Statistics*, vol. 21, no. 7, pp. 815–825, 2009.

[29] R. Eubank, C. Huang, Y. M. Maldonado, N. Wang, S. Wang, and R. Buchanan, "Smoothing spline estimation in varying-coefficient models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 3, pp. 653–667, 2004.

[30] J. Fan and W. Zhang, "Statistical estimation in varying coefficient models," *The annals of Statistics*, vol. 27, no. 5, pp. 1491–1518, 1999.

[31] D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang, "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometrika*, vol. 85, no. 4, pp. 809–822, 1998.

[32] G. Kauermann and G. Tutz, "On model diagnostics using varying coefficient models," *Biometrika*, vol. 86, no. 1, pp. 119–128, 1999.

[33] C. O. Wu, C.-T. Chiang, and D. R. Hoover, "Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data," *Journal of the American Statistical Association*, vol. 93, no. 444, pp. 1388–1402, 1998.

[34] J. Z. Huang and H. Shen, "Functional coefficient regression models for non-linear time series: a polynomial spline approach," *Scandinavian Journal of Statistics*, vol. 31, no. 4, pp. 515–534, 2004.

[35] J. Z. Huang, C. O. Wu, and L. Zhou, "Varying-coefficient models and basis function approximations for the analysis of repeated measurements," *Biometrika*, vol. 89, no. 1, pp. 111–128, 2002.

[36] J. Z. Huang, C. O. Wu, and L. Zhou, "Polynomial spline estimation and inference for varying coefficient models with longitudinal data," *Statistica Sinica*, pp. 763–788, 2004.

[37] C.-T. Chiang, J. A. Rice, and C. O. Wu, "Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 605–619, 2001.

[38] D. Zhang, "Generalized linear mixed models with varying coefficients for longitudinal data," *Biometrics*, vol. 60, no. 1, pp. 8–15, 2004.

[39] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[40] H. Joe, *Dependence modeling with copulas*. CRC press, 2014.

[41] X. Lin and R. J. Carroll, "Nonparametric function estimation for clustered data when the predictor is measured without/with error," *Journal of the American statistical Association*, vol. 95, no. 450, pp. 520–534, 2000.

[42] E. Kürüm, J. Hughes, R. Li, and S. Shiffman, "Time-varying copula models for longitudinal data," *Statistics and its Interface*, vol. 11, no. 2, p. 203, 2018.

[43] E. Kürüm, R. Li, S. Shiffman, and W. Yao, "Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data," *Statistica Sinica*, vol. 26, no. 3, p. 979, 2016.

[44] J. Fan, Y. Ma, and W. Dai, "Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models," *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 1270–1284, 2014.

[45] M. Marchetti-Bowick, J. Yin, J. A. Howrylak, and E. P. Xing, "A time-varying group sparse additive model for genome-wide association studies of dynamic complex traits," *Bioinformatics*, vol. 32, no. 19, pp. 2903–2910, 2016.

[46] W. Chu, R. Li, J. Liu, and M. Reimherr, "Feature selection for generalized varying coefficient mixed-effect models with application to obesity GWAS," *The Annals of Applied Statistics*, vol. 14, no. 1, pp. 276–298, 2020.

[47] X. Xia, H. Yang, and J. Li, "Feature screening for generalized varying coefficient models with application to dichotomous responses," *Computational Statistics & Data Analysis*, vol. 102, pp. 85–97, 2016.

[48] J. Li, Z. Wang, R. Li, and R. Wu, "Bayesian group Lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies," *The Annals of Applied Statistics*, vol. 9, no. 2, pp. 640–664, 2015.

[49] C. Ning, H. Kang, L. Zhou, D. Wang, H. Wang, A. Wang, J. Fu, S. Zhang, and J. Liu, "Performance gains in genome-wide association studies for longitudinal traits via modeling time-varied effects," *Scientific Reports*, vol. 7, no. 1, 2017.

[50] Z. He, M. Zhang, S. Lee, J. A. Smith, X. Guo, W. Palmas, S. L. Kardia, A. V. D. Roux, and B. Mukherjee, "Set-based tests for genetic association in longitudinal studies," *Biometrics*, vol. 71, no. 3, pp. 606–615, 2015.

[51] Z. He, S. Lee, M. Zhang, J. A. Smith, X. Guo, W. Palmas, S. L. Kardia, I. Ionita-Laza, and B. Mukherjee, "Rare-variant association tests in longitudinal studies, with an application to the Multi-Ethnic Study of Atherosclerosis (MESA)," *Genetic Epidemiology*, vol. 41, no. 8, pp. 801–810, 2017.

[52] Z. Wang, K. Xu, X. Zhang, X. Wu, and Z. Wang, "Longitudinal SNP-set association analysis of quantitative phenotypes," *Genetic Epidemiology*, vol. 41, no. 1, pp. 81–93, 2017.

[53] W. Wu, *Improving Risk Factor Identification of Human Complex Traits in Omics Data.* Thesis, Yale University, 2021.

[54] T. J. Hayeck, N. A. Zaitlen, P.-R. Loh, B. Vilhjalmsson, S. Pollack, A. Gusev, J. Yang, G.-B. Chen, M. E. Goddard, and P. M. Visscher, "Mixed model with correction for case-control ascertainment increases association power," *The American Journal of Human Genetics*, vol. 96, no. 5, pp. 720–730, 2015.

[55] D. Jiang, J. Mbatchou, and M. S. McPeek, "Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model," *Human Heredity*, vol. 80, no. 4, pp. 187–195, 2015.

[56] X. Wu and M. S. McPeek, "L-gator: genetic association testing for a longitudinally measured quantitative trait in samples with related individuals," *The American Journal of Human Genetics*, vol. 102, no. 4, pp. 574–591, 2018.

[57] X. Lin and D. Zhang, "Inference in generalized additive mixed modelsby using smoothing splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 2, pp. 381–400, 1999.

[58] G. Wahba, *Spline models for observational data.* Society for Industrial and Applied Mathematics, 1990.

[59] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 9–25, 1993.

[60] X. Lin and N. E. Breslow, "Bias correction in generalized linear mixed models with multiple components of dispersion," *Journal of the American Statistical Association*, vol. 91, no. 435, pp. 1007–1016, 1996.

[61] 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.

[62] Y. Liu and J. Xie, "Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures," *Journal of the American Statistical Association*, vol. 115, no. 529, pp. 393–402, 2020.

[63] A. K. Nikoloulopoulos, H. Joe, and N. R. Chaganty, "Weighted scores method for regression models with dependent data," *Biostatistics*, vol. 12, no. 4, pp. 653–665, 2011.

[64] H. Chen, C. Wang, Matthew, Adrienne, Z. Li, T. Sofer, Adam, W. Chen, John, Juan, S. Redline, George, Timothy, Cathy, K. Rice, and X. Lin, "Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models," *The American Journal of Human Genetics*, vol. 98, no. 4, pp. 653–666, 2016.

[65] 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.

[66] 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.

[67] J. S. Schildcrout, E. F. Schisterman, N. D. Mercaldo, P. J. Rathouz, and P. J. Heagerty, "Extending the case–control design to longitudinal data," *Epidemiology*, vol. 29, no. 1, pp. 67–75, 2018.

[68] 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, and K. Liu, "Multi-ethnic study of atherosclerosis: objectives and design," *American Journal of Epidemiology*, vol. 156, no. 9, pp. 871–881, 2002.

[69] 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.

[70] E. Evangelou, H. R. Warren, D. Mosen-Ansorena, B. Mifsud, R. Pazoki, H. Gao, *et al.*, "Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits," *Nature Genetics*, vol. 50, no. 10, pp. 1412–1425, 2018.

[71] G. Kichaev, G. Bhatia, P. R. Loh, S. Gazal, K. Burch, M. K. Freund, A. Schoech, B. Pasaniuc, and A. L. Price, "Leveraging polygenic functional enrichment to improve GWAS power," *The American Journal of Human Genetics*, vol. 104, no. 1, pp. 65–75, 2019.

[72] G. Loirand and P. Pacaud, "The role of Rho protein signaling in hypertension," *Nature Reviews Cardiology*, vol. 7, no. 11, pp. 637–647, 2010.

[73] C. Maack, T. Kartes, H. Kilter, H.-J. Schäfers, G. Nickenig, M. Böhm, and U. Laufs, "Oxygen free radical release in human failing myocardium is associated with increased activity of rac1-gtpase and represents a target for statin treatment," *Circulation*, vol. 108, no. 13, pp. 1567–1574, 2003.

[74] M. Moustafa-Bayoumi, S. Wisel, P. J. Goldschmidt-Clermont, and H. H. Hassanain, "Hypertension caused by transgenic overexpression of rac1," *Medicine & Science in Sports & Exercise*, vol. 35, no. 5, p. S186, 2003.

[75] M. Satoh, H. Ogita, K. Takeshita, Y. Mukai, D. J. Kwiatkowski, and J. K. Liao, "Requirement of Rac1 in the development of cardiac hypertrophy," *Proceedings of the National Academy of Sciences*, vol. 103, no. 19, pp. 7432–7437, 2006.

[76] N. Sawada, S. Salomone, H.-H. Kim, D. J. Kwiatkowski, and J. K. Liao, "Regulation of endothelial nitric oxide synthase and postnatal angiogenesis by Rac1," *Circulation Research*, vol. 103, no. 4, pp. 360–368, 2008.

[77] A. Shah, K. Gray, N. Figg, A. Finigan, L. Starks, and M. Bennett, "Defective base excision repair of oxidative DNA damage in vascular smooth muscle cells promotes atherosclerosis," *Circulation*, vol. 138, no. 14, pp. 1446–1462, 2018.

[78] N. R. Shah and M. Mahmoudi, "The role of DNA damage and repair in atherosclerosis: A review," *Journal of Molecular and Cellular Cardiology*, vol. 86, pp. 147–157, 2015.

[79] A. Uryga, K. Gray, and M. Bennett, "DNA damage and repair in vascular disease," *Annual Review of Physiology*, vol. 78, pp. 45–66, 2016.

[80] L. Wu, J. R. Sowers, Y. Zhang, and J. Ren, "Targeting DNA damage response in cardiovascular diseases: from pathophysiology to therapeutic implications," *Cardiovascular Research*, 2022.

[81] Y. Shiloh and Y. Ziv, "The ATM protein kinase: regulating the cellular response to genotoxic stress, and more," *Nature Reviews Molecular Cell Biology*, vol. 14, no. 4, pp. 197–210, 2013.

[82] T. Uziel, Y. Lerenthal, L. Moyal, Y. Andegeko, L. Mittelman, and Y. Shiloh, "Requirement of the MRN complex for ATM activation by DNA damage," *The EMBO Journal*, vol. 22, no. 20, pp. 5612–5621, 2003.

[83] X. Lu, L. Wang, X. Lin, J. Huang, C. Charles Gu, M. He, H. Shen, J. He, J. Zhu, and H. Li, "Genome-wide association study in chinese identifies novel loci for blood pressure and hypertension," *Human Molecular Genetics*, vol. 24, no. 3, pp. 865–874, 2015.

[84] L. V. Wain, A. Vaez, R. Jansen, R. Joehanes, P. J. Van Der Most, A. M. Erzurumluoglu, P. F. O'Reilly, C. P. Cabrera, H. R. Warren, and L. M. Rose, "Novel blood pressure locus and gene discovery using genome-wide association study and expression data sets from blood and the kidney," *Hypertension*, vol. 70, no. 3, pp. e4–e19, 2017.

[85] H. Sakai, S. Yokota, N. Kajitani, T. Yoneyama, K. Kawakami, Y. Yasui, and K.-i. Matsumoto, "A potential contribution of tenascin-X to blood vessel formation in peripheral nerves," *Neuroscience Research*, vol. 124, pp. 1–7, 2017.

[86] J. W. Petersen and J. Y. Douglas, "Tenascin-X, collagen, and ehlers–danlos syndrome: Tenascin-X gene defects can protect against adverse cardiovascular events," *Medical Hypotheses*, vol. 81, no. 3, pp. 443–447, 2013.

[87] P. Anyfanti, E. Gavriilaki, S. Douma, and E. Gkaliagkousi, "Endothelial dysfunction in patients with rheumatoid arthritis: the role of hypertension," *Current Hypertension Reports*, vol. 22, no. 8, pp. 1–10, 2020.

[88] K. W. Mathis, K. Wallace, E. R. Flynn, C. Maric-Bilkan, B. LaMarca, and M. J. Ryan, "Preventing autoimmunity protects against the development of hypertension and renal injury," *Hypertension*, vol. 64, no. 4, pp. 792–800, 2014.

CURRICULUM VITAE

Gang Xu

xug517@gmail.com

Degrees:

Master of Science, Statistics, 2017

University of Science and Technology of China, Hefei, China

Bachelor of Science, Statistics, 2014

University of Science and Technology of China, Hefei, China

Honors and Awards:

- Graduate College Finishing Fellowship — Fall 2022
- Mathematical Sciences Summer Fellowship — Summer 2022
- Invited to The Honor Society of Phi Kappa Phi — Spring 2022
- Wolzinger Family Science Research Scholarship — 2021-2022
- Chris McNamee Memorial Scholarship — 2021-2022
- Summer Doctoral Research Fellowship — Summer 2020
- UNLV Access Grant — 2017-2022

Publications:

- 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 (Under review).

Dissertation Title:

Retrospective Varying Coefficient Association Analysis of Longitudinal Binary Traits

Dissertation Examination Committee:

Chairperson, Dr. Amei Amei, Ph.D.

Committee Member, Dr. Kaushik Ghosh, Ph.D.

Committee Member, Dr. Malwane Ananda, Ph.D.

Graduate Faculty Representative, Dr. Edwin Oh, Ph.D.