2021

STATISTICAL METHODS IN GENETIC STUDIES

CHENG GAO
Michigan Technological University, chenggao@mtu.edu

STATISTICAL METHODS IN GENETIC STUDIES

By

Cheng Gao

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Statistics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2021

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Statistics.

Department of Mathematical Sciences

Dissertation Advisor:	*Dr. Kui Zhang*
Committee Member:	*Dr. Qiuying Sha*
Committee Member:	*Dr. Hairong Wei*
Committee Member:	*Dr. Shuanglin Zhang*
Committee Member:	*Dr. Xiao Zhang*
Department Chair:	*Dr. Jiguang Sun*

Contents

List of Figures

List of Tables

Preface

This dissertation is submitted for the degree of Doctor of Philosophy at Michigan Technological University. The research achievements described herein were conducted under the supervision of Prof. Kui Zhang in the Department of Mathematical Sciences, Michigan Technological University, between September 2016 and May 2021.

The research work presented in Chapter 1 is a collaborative work with Prof. Shuanglin Zhang and Prof. Qiuying Sha. Prof. Kui Zhang conceived the study, Cheng Gao developed the method, Cheng Gao implemented the R program, Cheng Gao performed the simulation studies and real data analysis, Cheng Gao drafted the manuscript, Prof. Shuanglin Zhang and Prof. Qiuying Sha proofread the manuscript, Prof. Kui Zhang and Cheng Gao revised the manuscript. Prof. Kui Zhang supervised the project. The research achievements have been published in a paper *MF-TOWmuT: Testing an optimally weighted combination of common and rare variants with multiple traits using family data*, *Genetic Epidemiology*, 2021 (`https://onlinelibrary.wiley.com/doi/10.1002/gepi.22355`).

The research work presented in Chapter 2 is a collaborative work with Prof. Hairong Wei. Prof. Kui Zhang conceived the study, Cheng Gao developed the method, Cheng Gao developed the R program, Cheng Gao performed simulation studies and real data analysis, Cheng Gao drafted the manuscript, Cheng Gao, Prof. Hairong Wei and Prof. Kui Zhang revised the manuscript. Prof. Kui Zhang supervised the project. The research achievements have been summarized in a manuscript, we have submitted the manuscript to *Frontiers in Genetics*, it is under review now.

The research work presented in Chapter 3 is from an ongoing project. Prof. Kui Zhang conceived the study, Cheng Gao developed the method, implemented the R program, and performed simulation studies. Cheng Gao drafted the manuscript, Cheng Gao and Prof. Kui Zhang revised the manuscript. Prof. Kui Zhang supervised the project. We will prepare a manuscript for publication in near future.

Acknowledgments

In the past few years at Michigan Technological University, I have obtained valuable research experience in statistical genetics and applied statistics. Through the coursework, I have laid a solid foundation in probability, statistics, and programming. I would like to express my deep and sincere gratitude to all the people who have helped me, including but not limited to my academic advisor and course instructors.

First of all, I sincerely dedicate my deepest gratitude to my advisor, Prof. Kui Zhang. Prof. Zhang provides me with the financial support for the last five academic years. As a graduate research assistant, I am fortunate to be able to concentrate all of my time and efforts on my research projects. Prof. Zhang is a very nice advisor, he always offers me timely support and help when I need. His rich academic experience and insights make great influence on my growing up to be a researcher. Without Prof. Zhang's guidence, support, and help, I cannot imagine I can make these achievements presented in this dissertation.

Secondly, I would like to express my gratitude to all the course instructors. From their instructions, I have learned new knowledge and built a solid foundation in statistics. In particular, I would like to thank Professor Iosif Pinelis. I took his four courses and audited one course. His love for mathematics, broad knowledge, and exquisite course-teaching impressed me a lot.

Thirdly, I would like to thank Prof. Shuanglin Zhang and Prof. Qiuying Sha for organizing the weekly journal club. From the journal club, The research papers presented in the journal club and the discussions have provided me with a great opportunity to learn the cutting-edge research in statistical genetics and have greatly helped my research projects presented in this dissertation.

Fourthly, I would like to express sincere appreciation to the committee members: Prof. Kui Zhang, Prof. Qiuying Sha, Prof. Hairong Wei, Prof. Shuanglin Zhang,

and Prof. Xiao Zhang. I thank them for spending their precious time on carefully proofreading my dissertation and providing me with valuable suggestions. It is my great honor to have them on my committee.

Finally, I would like to give my thanks to all the people who collaborated with me, helped me, and instructed me. I cannot list their name here, I will forever be grateful.

Abstract

This dissertation includes three Chapters. A brief description of each chapter is organized as follows.

In Chapter 1, we proposed a new method, called MF-TOWmuT, for genome-wide association studies with multiple genetic variants and multiple phenotypes using family samples. MF-TOWmuT uses kinship matrix to account for sample relatedness. It is worth mentioning that in simulations, we considered hidden polygenic effects and varied the proportion of variance contributed by it to generate phenotypes. Simulation studies show that MF-TOWmuT can preserve the type I error rates and is more powerful than several existing methods in different simulation scenarios, MF-TOWmuT is also quite robust to the proportion of variance explained by invisible polygenic effects and to the direction of effects of genetic variants.

In Chapter 2, we proposed a fast and efficient low rank penalized regression with the Elastic Net penalty for the eQTL mapping, called LORSEN. By considering the Elastic Net penalty instead of the L_1 penalty, our method can overcome two crucial drawbacks of the L_1 penalty, and outperforms two commonly used methods for the eQTL mapping, LORS and FastLORS, in many simulation scenarios in terms of average Area Under the Curve (AUC).

In Chapter 3, we proposed a bipartite network-based penalized regression model for the eQTL mapping, called BiNetPeR. This method takes into account the SNP-gene marginal association evidence to construct the SNP-gene bipartite network, then uses such a bipartite network to obtain the projected SNP network. Based on the normalized Laplacian matrix of the projected SNP network, we then formulate the eQTL mapping into a penalized regression model. Our simulation results show that our proposed method can maintain the appropriate false positive rate and outperforms two competing methods for the eQTL mapping, FastLORS and mtLasso2G.

Chapter 1

MF-TOWmuT: Testing An Optimally Weighted Combination Of Common and Rare Variants With Multiple Traits Using Family Data

Abstract

With rapid advancements of sequencing technologies and accumulations of electronic health records, a large number of genetic variants and multiple correlated human complex traits have become available in many genetic association studies. Thus, it becomes necessary and important to develop new methods that can jointly analyze the association between multiple genetic variants and multiple traits. Compared with methods that only use a single marker or trait, the joint analysis of multiple genetic variants and multiple traits is more powerful since such an analysis can fully incorporate the correlation structure of genetic variants and/or traits and their mutual dependence patterns. However, most of existing methods that simultaneously analyze multiple genetic variants and multiple traits are only applicable to unrelated samples. We develop a new method called MF-TOWmuT to detect association of multiple phenotypes and multiple genetic variants in a genomic region with family samples. MF-TOWmuT is based on an optimally weighted combination of variants. Our method can be applied to both rare and common variants and both qualitative and quantitative traits. Our simulation results show that (1) the type I error of MF-TOWmuT is preserved; (2) MF-TOWmuT outperforms two existing methods such as Multiple Family-based Quasi-Likelihood Score Test (MFQLS) and Multivariate Family-based Rare Variant Association Test (mFARVAT) in terms of power. We also illustrate the usefulness of MF-TOWmuT by analyzing genotypic and phenotipic data from the Genetics of Kidneys in Diabetes (GoKinD) study. R program is available at `https://github.com/gaochengPRC/MF-TOWmuT`.

1.1 Introduction

Genome-wide association studies (GWAS) and sequencing based association studies play an important role in revealing relationships between genetic variants and human complex traits. An important feature of many such large studies is that they generally collect a large number of correlated traits and genotypes at millions of genetic markers for thousands of samples. Therefore, such studies potentially have greater power to decipher the complicated relationship between genetic variations and human complex traits. For example, UK Biobank (`https://www.ukbiobank.ac.uk`) recruited 500,000 people aged between 40-69 years and collected correlated traits that are related to cancer, heart diseases, stroke, diabetes, etc. for these 500,000 people. Genome-wide genetic data at 805,426 markers are also available for 488,000 UK Biobank participants. At the same time, there are great challenges to developing more powerful statistical methods that can fully take advantage of such large scale studies and more efficiently analyze the huge volume of data generated.

To date, a variety of multi-marker based statistical methods (e.g., the Combined Multivariate and Collapsing (CMC) method (Li and Leal, 2008), Generalized T^2 (Zhu and Xiong, 2012), SNP-set Sequence Kernel Association Test (SKAT) (Wu et al., 2011), Sum of Squared Score U Statistic (SSU) (Pan, 2009), etc.) have been developed for detecting association between multiple genetic variants and a single trait (dichotomous or continuous). Such multiple markers based tests can combine information within all genetic variants available in a gene or a genomic region. It has been demonstrated that such methods are more powerful to detect the association between genetic variants and human complex traits than the methods based on single marker (e.g., (Li and Leal, 2008; Wu et al., 2010)). This is partially due to the fact that human complex traits are generally controlled by multiple genetic variants. In addition, with the advancements of next sequencing technologies, rare variant association studies such as Data-Adaptive Sum Test (aSum) (Han and Pan, 2010), Optimal Unified Test (SKAT-O) (Lee et al., 2012), CMC, Weighted Sum Statistic (WSS) (Madsen and Browning, 2009), SKAT, etc. (see (Lee et al., 2014) for an extensive review) have become readily available. Due to the extremely low allele frequencies of rare variants, single-marker based methods have lower power while multiple markers based tests are preferred in this situation.

In addition to polygenic effects, pleiotropic effects are important for describing the relationship between genetic variants and human complex traits. Pleiotropy refers to when one gene has effects on multiple phenotypes simultaneously. Some methods (Trait-based Association Test that uses Extended Simes procedure (TATES) (van der Sluis et al., 2013), MultiPhen that tests the linear combination of phenotypes most associated with the genotypes at each SNP (O'Reilly et al., 2012), etc.) have been proposed to detect the association between a single genetic variant and multiple traits. Such methods are desirable and have more power because many large studies have collected multiple correlated traits. In addition, human complex diseases are better characterized by multiple correlated traits. For example, hypertension can be characterized by systolic and diastolic blood pressure (Wang et al., 2005). As another example, diabetes is closely related with high-density lipoprotein (HDL), systolic blood pressure (SBP), diastolic blood pressure (DBP), and body mass index (BMI) (Bays et al., 2007). However, neither methods based on single genetic variant and multiple traits nor methods based on multiple genetic variants and single trait can take into account polygenic effects and pleiotropic effects simultaneously, which can lead to the loss of power. Therefore, it is both essential and beneficial to develop methods that can test the association between multiple genetic variants and multiple traits. A number of methods (MFQLS (Won et al., 2015), MF-KM (approach for multivariate family data using kernel machine regression) (Yan et al., 2015), multi-trait variant-set association test (MSKAT) (Wu and Pankow, 2016), Gene Association with Multiple Traits (GAMuT) (Broadaway et al., 2016), etc.) based on multiple genetic variants and multiple traits have been proposed recently. Additionally, family samples instead of unrelated samples are often collected. Family samples have greater power than unrelated samples to detect the association between rare variants and traits due to the enriched rare variants in family samples. A number of methods have been developed to detect association between multiple genetic variants and multiple traits using family as well as unrelated samples (Won et al., 2015; Yan et al., 2015; Fischer et al., 2018; Jiang and McPeek, 2014; Chen et al., 2013, 2009; Lasky-Su et al., 2010; Jiang et al., 2014; Schifano et al., 2012; Zhu and Xiong, 2012; Wang et al., 2016; Feng et al., 2011). These include methods based on linear and generalized linear mixed models that can incorporate the relatedness of family samples (Wu et al., 2011; Wu and Pankow, 2016; Lee et al., 2012; Yan et al., 2015; Jiang et al., 2014; Wu et al., 2010; Schifano et al., 2012; Lee et al., 2017b) and methods based on quasi-likelihood (Wang et al., 2016; Won et al., 2015).

In all aforementioned methods for detecting association between multiple markers and multiple traits, there are some constraints or drawbacks which restrict their applicability to association studies. First, genetic variants can be rare or common or a mixture thereof. As of now, a large number of methods have been developed only for rare variants association studies (Yan et al., 2015; Wu et al., 2011; Broadaway et al., 2016; Wang et al., 2016; Lee et al., 2012; Jiang and McPeek, 2014; Madsen and Browning, 2009; Li and Leal, 2008; Wu and Pankow, 2016; Lee et al., 2014). Methods that can combine common variants and rare variants in association studies have also been developed (Ionita-Laza et al., 2013; Maity et al., 2012; He et al., 2013; Feng et al., 2011; Zhu and Xiong, 2012; Lasky-Su et al., 2010; Kim et al., 2016; Fischer et al., 2018). However, due to the proportion and composition of causal variants, those methods are not uniformly the most powerful. For example, a burden test is for rare variants and powerful with a large proportion of causal variants while a variance-component test is for rare variants and powerful with a small proportion of causal variants. Second, traits can be binary or continuous or a mixture thereof. Some methods are applicable when all traits are quantitative (MF-KM, MSKAT, MONSTER (MinimumP-value Optimized Nuisance parameter Score Test Extended to Relatives) (Jiang and McPeek, 2014), etc.) or all traits are qualitative (Generalized Disequilibrium Test (GDT) (Chen et al., 2009), Generalized T^2, etc.) but not a mixture of them. Third, covariates (e.g., age, gender) can affect traits, so it is essential to incorporate covariates in the analysis. Some methods such as (Generalized T^2 (Zhu and Xiong, 2012), etc.) cannot consider covariates in the analysis. Fourth, in the genomic region of interest, risk variants and protective variants usually coexist, so it is important for a method to be robust to the proportion of risk or protective variants. Lee et al. (Lee et al., 2017b) discussed approaches (MAAUSS) and found some methods that work well only for genetic variants with the same direction of effects (e.g. burden tests) (Lee et al., 2014). Fifth, as we have mentioned, it is important to handle samples from arbitrary family structure. Although methods proposed in (Zhu and Xiong, 2012) do not require assumptions on relationships among individuals and can allow for unknown or partially known pedigree structures, they are mainly extended for case-control study. That is, they are only applicable to a single qualitative trait. Additionally, as mentioned before, these methods cannot incorporate covariates in their analysis, which will be undoubtedly less powerful. The method proposed in (Fischer et al., 2018) is actually a two-step scheme and only applicable to trios, and their method cannot allow for arbitrary pedigree structures. Similarly, the method proposed in (Feng et al., 2011) is only applicable to sib-pairs and cannot be applied to family data with arbitrary pedigree structures. Their method is developed for case-control study and does not offer a strategy in the presence of multiple quantitative/qualitative traits. Therefore, it is necessary to develop novel statistical methods for association studies with multiple genotypes and multiple traits using family as well as unrelated samples.

In this paper, we develop a new method called MF-TOWmuT to detect association between multiple genotypes and multiple traits using family as well as unrelated samples. This method is an extension of a method developed by us, TOWmuT - testing an optimally weighted combination of common and/or rare variants with multiple traits. MF-TWOmuT can accommodate covariates and relatedness among family samples. The method can be applied to multiple rare and common variants and their mixture and to multiple quantitative and qualitative traits and their mixture. We conduct extensive simulations to evaluate and compare MF-TOWmuT with several existing methods including MFQLS and mFARVAT(Burden, SKAT-O) (Wang et al., 2016). We find that MF-TOWmuT is robust to the proportion of dichotomous or continuous traits and to the proportion of risk or protective variants and outperforms MFQLS and mFARVAT in terms of power in different scenarios. We also apply MF-TOWmuT to genetic data and multiple traits from GoKinD Study (Mueller et al., 2006; Pezzolesi et al., 2009) and identify a genome-wide significant gene showing cross-phenotype effects.

1.2 Materials and Methods

We consider a study consisting of n family and/or unrelated samples. Each sample has K potentially correlated quantitative or qualitative traits and genotypes at M bi-allelic marker loci (SNPs). Let x^*_{im} denote the genotype score of the i-th individual at the m-th marker, coded in an additive manner. Let y^*_{ik} denote the phenotype of the i-th individual for the k-th trait. We first centralize x^*_{im} and y^*_{ik}: $x_{im} = x^*_{im} - \bar{x}_m$ and $y_{ik} = y^*_{ik} - \bar{y}_k$ where $\bar{x}_m = \frac{1}{n}\sum_{i=1}^{n} x^*_{im}$ and $\bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y^*_{ik}$. Let $Y = (y_1^T, y_2^T, ..., y_n^T)^T$ be an $n \times K$ matrix of phenotypes of all n individuals for all traits, $X_i = (x_{i1}, x_{i2}, ..., x_{iM})^T, i = 1, 2, ..., n$, $X = (X_1^T, X_2^T, ..., X_n^T)^T$ be an $n \times M$ matrix of genotypes of all n individuals at all M marker loci. To take into account genotypes at all M markers, we consider the weighted combination of genotypes as the new genotype at a "super marker" for the i-th individual, $x_i = w^T X_i$, $x = (x_1, x_2, ..., x_n)^T$, in which the optimal weight vector w will be determined later.

1.2.1 Without Covariates

We consider the following linear model without covariates to explore the relationship between multiple genetic variants and multiple traits:

$$x = Y\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2\Phi), \tag{1.2.1}$$

where $\beta = (\beta_1, \beta_2, \cdots, \beta_K)^T$, $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$, and Φ is the kinship matrix and is considered as known. To test the null hypothesis of $\beta_i = 0, i = 1, 2, ..., K$, the corresponding score test statistic is:

$$T_{score} = \frac{w^T X^T \Phi^{-1} Y (Y^T \Phi^{-1} Y)^{-1} Y^T \Phi^{-1} X w}{\hat{\sigma}^2}, \tag{1.2.2}$$

The maximum likelihood estimate (MLE) of σ^2 is $\hat{\sigma}^2 = \frac{1}{n} x^T \Phi^{-1} x = \frac{1}{n} w^T X^T \Phi^{-1} X w$. We use D to represent $\frac{1}{n} X^T \Phi^{-1} X$ in the score test statistic. So, we have

$$T^0_{score} = \frac{w^T X^T \Phi^{-1} Y (Y^T \Phi^{-1} Y)^{-1} Y^T \Phi^{-1} X w}{w^T D w}. \tag{1.2.3}$$

The final test statistic is defined as:

$$T_{MF-TOWmuT} = \max_{w} T^0_{score} = \lambda_{max}((Y^T\Phi^{-1}Y)^{-1}Y^T\Phi^{-1}B\Phi^{-1}Y), \quad B = XD^{-1}X^T. \tag{1.2.4}$$

For simplicity, we use $\lambda_{max}(A)$ to denote the largest eigenvalue of the matrix A. Detailed derivations of the test statistic are shown in the Appendix. Note that for a given weight vector w, the score test statistic T^0_{score} is a function of w, X, and Y. Its power depends on w, X, and Y. However, the score test statistic T^0_{score} has an approximate χ^2 distribution under the null hypothesis. The null distributions of the score test statistics are the same for different w. Therefore, we can maximize the power by maximizing the test statistic T^0_{score} (or equivalently, minimizing the p-value).

We use permutations to derive p-value of proposed test statistic, $T_{MF-TOWmuT}$. Specifically, we permute phenotypes B times, calculate a test statistic $T^{(b)}_{MF-TOWmuT}$ for each permutation, $b = 1, 2, ..., B$, and then use the formula below to calculate the corresponding p-value:

$$P - Value = \frac{\#\{b : T^{(b)}_{MF-TOWmuT} \geq T^{(o)}_{MF-TOWmuT}, b = 1, 2, ..., B\}}{B}, \tag{1.2.5}$$

where $T^{(o)}_{MF-TOWmuT}$ is calculated based on the original data.

1.2.2 With Covariates

When covariates are present in the model, we regress phenotypes and genotypes on the covariates, respectively. Then we use corresponding residuals to replace them in the formulas above. Detailed derivations of test statistic are shown in the Appendix.

1.2.3 Methods Compared

As we have argued in introduction, most methods for family data analysis are not suitable for comparison with our proposed method. Our method can be used to conduct association studies with multiple (rare and/or common) genetic variants and multiple (qualitative and/or quantitative) traits for unrelated as well as family samples with arbitrary pedigree structures. MF-KM is computationally intensive and their software cannot be easily adapted to an arbitrary number of traits, so we excluded it from our comparison. Fischer et al. (2018) proposed a two-stage method for gene association with multiple traits from case-parent trios. Specifically, in the first stage, GAMuT (gene association with multiple traits) is performed for each gene using the phenotypes and genotypes of the parents. In the second stage, GAMuT is used again for a subset of top genes selected from the first stage using the robust within-family information from offspring. Since this method is only applicable to the case-parent trios design, it is excluded from our comparison in simulation studies. For the purpose of comparison, we choose MFQLS and mFARVAT as two wrestlers in the context of common and rare genetic variants, respectively. MFQLS is a quasi-likelihood based score test and developed specifically for common variants. It can be applied for both quantitative and dichotomous phenotypes and is robust against population substructures as long as large-scale genomic data is available. Similarly, mFARVAT is also a quasi-likelihood based score test, but developed specifically for rare variants with multiple phenotypes, and tests both homogeneous and heterogeneous effects of each variant on multiple phenotypes. mFARVAT actually generalizes SKAT, burden, and SKAT-O tests.

1.3 Simulations

The samples from the parents-offspring trios or a three-generation pedigree (Figure 1.1) are used. Coalescent simulator $\mathcal{C}$ (Schaffner et al., 2005) is first used to simulate 20,000 haplotypes over a 250 kb chromosome region mimicking European populations. $\mathcal{C}$ is a software to simulate haplotypes using a coalescent model based on empirical genetic patterns observed from different populations. For each founder, two haplotypes among 20,000 haplotypes are randomly selected with replacement to form his/her genotypes. For each nonfounder, one haplotype from his/her father and one hapolotype from his/her mother are randomly chosen to pair them together and form his/her genotypes. For the parents-offspring trios, genotypes for 1,500 samples of 500 trios are generated. For the three-generation pedigree, genotypes of 1,000 samples from 100 families are generated. To select genetic markers used in simulations, 60 ($M = 60$) genetic variants are randomly selecetd. Among 60 genetic variants, we assume n_c variants are causal of which n_p variants are protective, $n_r (= n_c - n_p)$ variants are risk variants. Rare variants are defined as those variants with minor allele frequency (MAF) in (0.25%, 3%), and common variants defined as those variants with MAF $\geq$ 5%. Note that all selected 60 genetic variants have MAF in the range (0.25%, 99.75%).

In simulations, quantitative traits are generated based on six distinct factor models (Aschard et al., 2014; Wang et al., 2018b). Qualitative traits are generated by setting a threshold (e.g. quantiles of trait values) (Won et al., 2015; Fischer et al., 2018) for quantitative phenotypes. Here, the 75% percentile of quantitative phenotypes is set as the threshold. In other words, trait values falling below the 75 % percentile are set to be 0 (unaffected), and 1 (affected) otherwise. Furthermore, we consider presence or absence of two covariates Z_1 and Z_2, where Z_1 is a continuous covariate generated from standard normal distribution, and Z_2 is a binary covariate generated from Bernoulli distribution with success probability 0.5.

The similar procedures are used to generate polygenic effects of other genes on phenotypes. Specifically, (1) Cosi is used to simulate an independent set of haplotypes; (2) genotypes of samples are generated according to the procedures described above; (3) a set of genetic variants are randomly selected as causal variants; and (4) the traits value based on the selected variants are generated and added to traits values generated from the first set of genetic variants. The effect sizes are determined by **P**roportion of **V**ariance explained by invisible **P**olygenic effects (PVP). Note the genotypes from this set of genetic variants are not used in MF-TOWmuT to detect the association between genetic variants and traits. This workflow is illustrated in

Figure 1.2.

The following factor models (Wang et al., 2018b; van der Sluis et al., 2013; Aschard et al., 2014) are used to generate K(=10) correlated trait values of an individual based on his/her genotypes:

$$y = (0.5Z_1 + 0.5Z_2)e + \beta x + c\gamma f + \sqrt{1-c^2} \times \epsilon \tag{1.3.6}$$

In the above fomula, $y = (y_1, ..., y_K)^T$ are K trait values. $x = (x_1, ..., x_{n_c})^T$ are the genotypes at n_c causal markers. $\beta = (\beta_1^T, \beta_1^T, ..., \beta_K^T)^T$ is a $K \times n_c$ matrix of coefficients of genotypes of causal markers. $e = (1, ..., 1)^T$ is a vector of 1. $f = (f_1, ..., f_R)^T$ has multivariate normal distribution with mean 0 and covariance matrix Σ, where $\Sigma = (1-\rho)I + \rho A$ and A is a matrix with elements of 1, I is the identity matrix, and ρ is the correlation between f_i and f_j. γ is a $K \times R$ matrix, R is the number of factors. c^2 is within-factor correlation where c is a constant. Different ρ, A, γ, R, c can be used to generate different degrees of relatedness among traits. $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_K)^T$: a vector of random noise, and $\epsilon_k \overset{i.i.d}{\sim} N(0,1)$, for $k = 1, 2, ..., K$. We also use $x_i^r, i = 1, 2, ..., n_r$ to represent genotype at the i-th risk marker and $x_j^p, j = 1, 2, ..., n_p$ to represent genotype at the j-th protective marker, respectively. Z_1 and Z_2 are two covariates as described above.

In this paper, the following six models are considered.

- Model 1: There is only one factor and genotypes impact 6 traits, $R = 1$ and $\gamma = (1, ..., 1)^T$.

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{i=1}^{n_r} \beta_{ki}^r x_i^r - \sum_{j=1}^{n_p} \beta_{kj}^p x_j^p + cf_1 + \sqrt{1-c^2} \times \epsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_1 + \sqrt{1-c^2} \times \epsilon_k, & k > 6 \end{cases}$$

- Model 2: There are five factors and genotypes impact 6 traits, $R = 5$ and $\gamma = diag(D_1, D_2, ..., D_5)$, where

$$D_i = \left(\underbrace{1, \cdots, 1}_{\text{K/5}} \right) \text{ for } i = 1, \cdots, 5.$$

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{i=1}^{n_r} \beta_{ki}^r x_i^r - \sum_{j=1}^{n_p} \beta_{kj}^p x_j^p + cf_{\lfloor (k-1)/2 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_{\lfloor (k-1)/2 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & k > 6 \end{cases}$$

- Model 3: There are two factors and genotypes impact 6 traits, $R = 2$ and $\gamma = diag(D_1, D_2)$, where $D_i = \left(\underbrace{1, \cdots, 1}_{\text{K/2}}\right)$ for $i = 1, \cdots, 2$.

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{i=1}^{n_r} \beta_{ki}^r x_i^r - \sum_{j=1}^{n_p} \beta_{kj}^p x_j^p + cf_{\lfloor (k-1)/5 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_{\lfloor (k-1)/5 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & k > 6 \end{cases}$$

- Model 4: There are five factors and genotypes impact one trait, $R = 5$ and $\gamma = diag(D_1, D_2, ..., D_5)$, where $D_i = \left(\underbrace{1, \cdots, 1}_{\text{K/5}}\right)$ for $i = 1, \cdots, 5$.

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{i=1}^{n_r} \beta_{ki}^r x_i^r - \sum_{j=1}^{n_p} \beta_{kj}^p x_j^p + cf_{\lfloor (k-1)/2 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & k = 1 \\ 0.5Z_1 + 0.5Z_2 + cf_{\lfloor (k-1)/2 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & k > 1 \end{cases}$$

- Model 5: There are two factors and genotypes impact one trait, $R = 2$ and $\gamma = diag(D_1, D_2)$, where $D_i = \left(\underbrace{1, \cdots, 1}_{\text{K/2}}\right)$ for $i = 1, \cdots, 2$.

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{i=1}^{n_r} \beta_{ki}^r x_i^r - \sum_{j=1}^{n_p} \beta_{kj}^p x_j^p + cf_{\lfloor (k-1)/5 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & k = 1 \\ 0.5Z_1 + 0.5Z_2 + cf_{\lfloor (k-1)/5 \rfloor + 1} + \sqrt{1-c^2} \times \epsilon_k, & k > 1 \end{cases}$$

- Model 6: There are K factors and genotypes impact 6 traits, $R = K$, $\gamma = I$, and $c = 1$.

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{i=1}^{n_r} \beta_{ki}^r x_i^r - \sum_{j=1}^{n_p} \beta_{kj}^p x_j^p + cf_1 + \sqrt{1-c^2} \times \epsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_1 + \sqrt{1-c^2} \times \epsilon_k, & k > 6 \end{cases}$$

In summary, the following scenarios are considered to evaluate the type I error and power of MF-TOWmuT and methods compared: (1) six factor models for generating traits; (2) two family pedigree structures (parents-offspring trio and a three-generation pedigree); (3) presence or absence of covariates; (4) different proportions of variance explained by invisible polygenic effects; (5) three types of phenotypes (qualitative phenotypes, quantitative phenotypes and their mixture).

To evaluate the type I error rate, β, the matrix of coefficients of genotypes of causal markers is set as 0. To evaluate the power, β is determined by the following ways. Let h_{all} and h_k be the heritability of all causal variants for all K traits and the

k-th trait, respectively. First, the value of h_{all} is specified. Second, K random numbers $r_1, r_2, \cdots r_K$ from uniform distribution of (0, 1) are generated and used to define the heritability of the k-th trait: $h_k = h_{all} * r_k / \sum_{k=1}^{K} r_k$. Third, given h_k, n_c random numbers $t_1, t_2, \cdots t_{n_c}$ from uniform distribution of (0, 1) are generated and used to determine the heritability of the m-th variant for the k-th trait: $h_k^m = h_k t_m / \sum_{j=1}^{n_c} t_j$. Different proportions of protective variants $(0\%, 20\%, 40\%, 60\%, 80\%, 100\%)$, different PVPs $(0\%, 5\%, 10\%, 15\%, 20\%, 25\%, 30\%)$, and different numbers of quantitative traits (0, 6, and 10) are considered.

1.4 Results

1.4.1 Type I Error Rate

In our simulation, 1,000 permutations are used to estimate p-values, and 500 replicates are used to estimate type I error rates and corresponding 95% Wald confidence intervals. If the 95% confidence interval doesn't contain the significance level, e.g., 0.01 or 0.05, then the type I error rate is inflated or conservative. Table 1.1 shows type I error rates in the following scenarios considered: mixture of four qualitative and six quantitative traits, two covariates, three-generation pedigree, seven distinct PVPs using Model 2 at significance level 0.01 and 0.05. Figure 1.3 shows corresponding Q-Q plots for PVP = 0.1 and PVP = 0.25, respectively. Table 1.2 shows type I error rates at significance level 0.01 and 0.05 with fixed PVP (= 0.5), mixture of four qualitative and six quantitative traits, two covariates using six models for three-generation case. Figure 1.3 shows corresponding Q-Q plots for each model. From Table 1.1 and Table 1.2 and Figure 1.3 and 1.4, we can see that: First, our newly developed method, MF-TOWmuT has the appropriate type I error in all situations. Second, TOWmuT has inflated type I error rates in most situations, and its type I error rates increase with the increased proportion of variance explained by invisible polygenic effects. This is expected since TOWmuT is developed for unrelated samples. Third, the type I error rates from MFQLS are not stable. They are either too conservative in some situations or inflated in some other situations. Fourth, mFARVAT (Burden, SKAT-O) has correct type I error rates in the absence of invisible polygenic effects. However, as the proportion of variance explained by invisible polygenic effects increases, its type I error rate inflates consistently.

1.4.2 Power

To evaluate the power of MF-TOWmuT and show comparison with MFQLS and mFARVAT, seven specific scenarios were considered and summarized in Table 1.3. For notational simplicity, we use PPV to represent **P**roportion of **P**rotective **V**ariants in Table 1.3. Notice that all power is evaluated at significance level 0.05.

Figure 1.5 shows the power comparison between MF-TOWmuT and MFQLS for common variants in the first scenario. We can see that MF-TOWmuT achieves higher power than MFQLS consistently for six models for distinct proportion of protective variants.

Figure 1.6 shows the power comparison between MF-TOWmuT and mFARVAT for rare variants in the second scenario. We can see that MF-TOWmuT achieves higher power than both mFARVAT-Burden and mFARVAT-SKAT-O for six models and distinct proportion of protective variants, and mFARVAT-Burden is sensitive to the proportion of protective variants, especially in models 2, 3 and 6. mFARVAT-SKAT-O has the comparable power with MF-TOWmuT only in model 2 and 6.

Figure 1.7 shows the power comparison between MF-TOWmuT and MFQLS for common variants in the third scenario. We can see that MF-TOWmuT achieves higher power than MFQLS consistently for six models and different PVP.

Figure 1.8 shows the power comparison between MF-TOWmuT and mFARVAT for rare variants in the fourth scenario. MF-TOWmuT has relatively high power for six models and distinct proportion of protective variants. mFARVAT-Burden has the lowest power among these three methods, especially in model 4 and 5 where only one phenotype is affected by genetic variants. This indicates that mFARVAT is not optimal at detecting association between multiple markers and multiple traits when genetic variants are only associated with a small number of traits. Although mFARVAT-SKAT-O achieves high power, especially in models 2, 3, and 6, this may be just due to the fact that mFARVAT-SKAT-O has inflated type I error rates when PVP is high. Moreover, from models 1, 4, and 5, we can see as PVP increases, power of mFARVAT-SKAT-O increases consistently.

Figure 1.9 shows the power comparison among MF-TOWmuT, MFQLS and mFARVAT in the last three scenarios. We can see that for a mixture of rare variants and common variants, both MF-TOWmuT and MFQLS can achieve high power no matter in trio case or three-generation case, though MFQLS is designed for common variants. However, mFARVAT has very low power in these three scenarios.

1.5 Application To Real Data

To demonstrate performance of our proposed method, MF-TOWmuT was applied to genotypic and phenotypic data from Genetics of Kidneys in Diabetes (GoKinD) study (dbGaP accession numbers phs000018.v2.p1 and phs000088.v1.p1). Quality control was performed with Plink (Purcell et al., 2007): SNPs with missing rate greater than 10% were removed, then individuals with missing rate greater than 10% were removed. Hardy-Weinberg equilibrium (HWE) exact test was applied and SNPs with p-value less than 1×10^{-6} were removed. Missing genotypes were replaced by the average of genotypes at the marker. Missing phenotypes were imputed with the median value of that phenotype. After quality control, 1,792 individuals containing 542 trios were remained in the analysis. A lift over tool from UCSC Genome Browser (`https://genome.ucsc.edu/`) was used to change the coordinates of SNPs from GRCh36 to GRCh 38, then the main annotation file in GENCODE (Frankish et al., 2019) was used to locate SNPs and assign a SNP to a gene if that SNP lay within a 1-kb flank region of the gene on either side. Only genes with at least 16 SNPs were included in the final analysis. Specifically, 4,006 genes, 730 rare SNPs ($0 < \text{MAF} < 0.03$) and 169,567 common SNPs ($0.03 \leq \text{MAF} < 1$) (a total of 170,297 SNPs) were used. Therefore, the Bonferroni-corrected genome wide significance level is $0.05/4006 \approx 1.248 \times 10^{-5}$. It is well known that Bonferroni correction is quite conservative, so we suggest using 1×10^{-4} as significance level as in (Fischer et al., 2018).

MF-TOWmuT was used to test the association between 4,006 genes and four correlated phenotypes (SBP, DBP, HDL, BMI). Following (Fischer et al., 2018), 16 covariates (age, gender, renal function status (proteinuric, dialysis, renal transplant, or other), smoking status, insulin intake (yes or no), antihypertension drug intake (yes or no), and lipid-lowering medication intake (yes or no)) were included in analysis, and ten principal components from genotype data were used as covariates to account for potential population stratification. In order to save computational time, we took a hierarchical exclusion strategy to derive p-values for significant genes. We first selected genes that showed evidence of association based on a small number of permutations (e.g. 5,000), and then used a larger number of permutations (e.g. 100,000) to test the selected genes. We repeated this process with increasing number of permutations. In the final stage, 1,000,000 permutations were used to derive p-value of significant genes. MF-TOWmuT is able to identify one novel gene Long Intergenic Non-Protein Coding RNA 535(LINC00535, containing 69 common variants, chr8:93213302 - 93700433) based on suggested significance level, the derived p-value is 2.2×10^{-5} using MF-TOWmuT. As a comparison, the derived p-value for LINC00535 is 1.02854×10^{-2} using MFQLS. Additionally, we applied MFQLS to discover another novel gene named LINC00393 to be associated with SBP, DBP, HDL and BMI with p-value 9.25345×10^{-8}, however, MF-TOWmuT had higher p-value 0.6098 for LINC00393, failed to discover it. In (Fischer et al., 2018), their found gene was VPS41 with p-value less than 5×10^{-6}, however, neither MF-TOWmuT nor MFQLS dug out this gene with p-values 0.0404 and 0.136333, respectively. We summarize these results in Table 1.4. From Genome Catalog (https://www.ebi.ac.uk/gwas/), we know that gene LINC00535 has been revealed to be associated with IgG glycosylation (Lauc et al., 2013), temperament (Service et al., 2012), facial morphology (factor 16) (Lee et al., 2017a), diisocyanate-induced asthma and lack of perseverance, but has not been found to be associated with diabetes-related traits. VPS41 (VPS41 subunit of HOPS complex, chr7:38722974 - 38932394), as a member of Vesicle medicated protein sorting family, plays an important role in segregation of intracellular molecules into distinct organelles. Expression studies indicate that VPS41 may be involved in the formation and fusion of transport vesicles from the Golgi (https://www.ncbi.nlm.nih.gov/gene/27072). LINC00393 (Long Intergenic Non-protein Coding RNA 393, chr: 13:73413473 - 73661891) has been discovered to be associated with eczema, respiratory diseases (Kichaev et al., 2019), colorectal cancer (Huyghe et al., 2019) and other complex diseases, but has not been found to be associated with diabetes-related traits.

1.6 Discussion

With advancements in high-throughput sequencing technology and availability of large scale genetic association studies, genotypes at millions of genetic markers and a large number of correlated human complex traits for thousands of samples have been collected. Advanced statistical methods are needed to fully take advantage of such data to investigate the relationship between genetic variants and human complex traits. However, most of available methods based on multiple genetic variants and multiple traits are for unrelated samples. Family samples are routinely collected. It has been shown that rare variants play an important role in human complex traits. Family samples can be more powerful for rare variant association studies due to the enriched rare variants. Therefore, it is necessary to develop such flexible method called MF-TOWmuT to carry out association studies with multiple genetic variants and multiple traits using family as well as unrelated samples.

MF-TOWmuT can be applied to rare and/or common variants association studies with qualitative and/or quantitative traits. Our simulation studies show that MF-TOWmuT preserve the desired type I error rates, and achieve higher power than MFQLS and mFARVAT in different scenarios. MF-TOWmuT provides a novel approach to genetic analysis of multivariate data for family-based studies. The computational time for MF-TOWmuT depends on a numer of factors, including the total number of genes, the number of traits, the sample size, the family pedigree structure, etc. Since MF-TOWmuT uses permutations to evaluate p-values, it can be quite computationally intensive. The computational time of MF-TOWmuT with 1,000 permutations on a data set with 1,000 individuals from a three generation pedigree, ten traits, 60 genetic variants in a genomic region on a laptop with 4 Intel(R) Core(TM) i7-7500U CPUs @ 2.70GHz and 8 GB RAM is about 3.7 minutes. To carry out such an analysis at about 25,000 genes from genome-wide association studies, a much larger number of permutations are needed to achieve the genome wide significance level. A hierarchical exclusion strategy is taken here as described in last section. We are pursuing a better strategy to improve computational efficiency in our program. Further theoretical approximation is desirable to reduce computational burden.

To estimate σ^2 in (2), we used $D = \frac{1}{n}X^T\Phi^{-1}X$ instead of the diagonal of D. When only rare genetic variants are involved, the diagonal of D may be used as an approximation of D to reduce the computational cost due to the weak correlation between rare genetic variants. The diagonal of D was used by (Wang et al., 2018b) and (Pan, 2009). However, when common variants are involved in the analysis, the correlation among genetic variants may not be simply ignored. We have performed extensive simulations to evaluate the type I error rate and power using the approximation matrix (the diagonal of D) and the original matrix (D) (data not shown). Our results show that both methods preserve the type I error rates and have the similar power in most situations. In a number of simulations, the method using D

does have the significantly higher power than the method using the diagonal of D.

In this paper, we use the linear sum of centered genotype scores as the response. The response of the i-th individual, $\sum_{m=1}^{M} w_i(x_{im}^* - \bar{x}_m)$, is a complicated mixture of dependent binomial random variables and approximate normal random variables. It is difficult to analytically derive the exact distribution and certainly it is not approximately normally distributed. However, we do not think such violation of normality for linear regression will drastically affect our method and conclusions. First, although the normality assumption is required for appropriate statistical inference in linear regression, such assumption plays a less important role. In genetic association studies, as pointed out by Bůžková (Bůžková, 2013), "We conclude that it is a combination of heteroscedasticity, minor allele frequency, sample size, and to a much lesser extent the error distribution, that matter for proper statistical inference". Second, the use of binary variable as the response in linear regression has been explored (Hellevik, 2009; Gomila, 2019). Hellevik found that the use of binary variable as the response in linear regression was acceptable and resulted in nearly identical significance tests as those obtained from logistic regression. Gomila demonstrated that linear regression was better than logistic regression for estimating causal effects of treatments on binary responses. The linear models with a binary outcome variable, called linear probability model (LPM), have also been extensively examined in the paper of Chatla and Shmueli (Chatla and Shmueli, 2017). Chatla and Shmueli and other researchers have found that LPM has several attractive advantages over logistic and probit models. First, LPM has computational advantages, especially with a large number of samples, because least square estimation is computationally cheaper than the maximum likelihood method used in logistic or probit models. Second, researchers found that LPM based coefficient directions, statistical significance, and marginal effects yielded results similar to logistic and probit models. LPM estimators are consistent with the true parameters up to a multiplicative scalar. Third, the normal linear models with genotypes being dependent variable, called reverse regression models, have been used in genetic association studies (Mägi et al., 2017; Zhang and Sun, 2018). Fourth, based on our extensive simulations, our method has the correct type I error rates. For these reasons, we do not think the normality assumption will drastically affect our method and conclusions.

In this work, we only consider individual level data from a single study. Developing a meta-analysis approach is important and necessary on the basis of MF-TOWmuT with individual level data from multiple studies. Correspondingly, challenges arising from meta-analysis need to be considered such as how to handle sample overlapping between studies and account for complex population structures in generalizing MF-TOWmuT. Additionally, in this work, we use a theoretical kinship coefficient matrix with known pedigree information. When pedigree structure is unknown, we can instead use an empirical kinship coefficient matrix which is expected to show similar results.

1.7 Tables and Figures

Table 1.1: The estimated type I error rates of MF-TOWmuT with covariates, mixture of four qualitative phenotypes and six quantitative phenotypes using Model 2 in three-generation case.

Significance level	**Method**	**PVP**						
		0	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**
0.01	TOWmuT	0.016	0.020	**0.066**	**0.088**	**0.132**	**0.188**	**0.214**
	MF-TOWmuT	0.008	0.016	0.018	**0.002**	0.006	0.014	0.010
	MFQLS	0.012	0.012	**0.004**	**0.002**	**0.002**	**0.004**	0.006
	mFARVAT-Burden	0.012	0.014	0.018	**0.032**	0.022	**0.040**	**0.036**
	mFARVAT-SKAT-O	0.010	**0.040**	**0.062**	**0.140**	**0.148**	**0.174**	**0.224**
0.05	TOWmuT	0.062	**0.088**	**0.180**	**0.208**	**0.320**	**0.352**	**0.430**
	MF-TOWmuT	0.048	0.05	0.058	0.040	0.054	0.054	0.054
	MFQLS	0.060	0.068	**0.018**	**0.022**	**0.026**	**0.016**	**0.034**
	mFARVAT-Burden	0.054	0.060	**0.084**	**0.096**	**0.098**	**0.112**	**0.114**
	mFARVAT-SKAT-O	0.050	**0.102**	**0.174**	**0.280**	**0.334**	**0.368**	**0.396**

Note: The conservative or inflated type I error rates are indicated by bold fonts.

Abbreviation: PVP, Proportion of Variance explained by invisible Polygenic effects.

Table 1.2: The estimated type I error rates of MF-TOWmuT with covariates, mixture of four qualitative phenotypes and six quantitative phenotypes, and fixed PVP(= 0.5) using six models in three-generation case.

Significance level	**Method**	**Factor Models Simulating Phenotypes**					
		1	**2**	**3**	**4**	**5**	**6**
0.01	TOWmuT	**0.311**	**0.375**	**0.355**	**0.379**	**0.340**	**0.364**
	MF-TOWmuT	0.012	0.006	**0.002**	0.008	0.008	0.012
	MFQLS	0.020	0.008	0.006	0.008	0.008	**0.004**
	mFARVAT-Burden	**0.052**	**0.066**	**0.052**	**0.056**	**0.06**	**0.062**
	mFARVAT-SKAT-O	**0.273**	**0.317**	**0.250**	**0.277**	**0.292**	**0.290**
0.05	TOWmuT	**0.509**	**0.557**	**0.561**	**0.581**	**0.498**	**0.582**
	MF-TOWmuT	0.058	0.062	0.050	0.060	0.042	0.064
	MFQLS	**0.074**	**0.020**	**0.034**	**0.034**	**0.028**	0.044
	mFARVAT-Burden	**0.128**	**0.168**	**0.158**	**0.142**	**0.142**	**0.168**
	mFARVAT-SKAT-O	**0.457**	**0.505**	**0.487**	**0.483**	**0.53**	**0.538**

Note: The conservative or inflated type I error rates are indicated by bold fonts.

Table 1.3: Seven specific scenarios considered in power comparison.

Scenario	**Phenotypes**	**Pedigree Type**	**Causal SNPs**	**PPV**	**PVP**
One	ten quantitative	trio	only common	varying	zero
Two	ten quantitative	three-generation	only rare	varying	zero
Three	ten qualitative	trio	only common	fixed	varying
Four	ten qualitative	three generation	only rare	fixed	varying
Five	ten quantitative	three-generation	common(20%)+rare(80%)	fixed	zero
Six	ten quantitative	three-generation	common(80%)+rare(20%)	fixed	zero
Seven	ten quantitative	trio	common(50%)+rare(50%)	fixed	zero

Abbreviations: PPV, proportion of protective variants; PVP, proportion of variance explained by invisible polygenic effects.

Table 1.4: Real data analysis results.

Gene	**P-Value**		
	MF-TOWmuT	**MFQLS**	**Fisher's method**
LINC00535	$\mathbf{2.20 \times 10^{-5}}$	1.03×10^{-2}	N/A
LINC00393	0.610	$\mathbf{9.25 \times 10^{-8}}$	N/A
VPS41	0.040	0.136	$\mathbf{< 5.0 \times 10^{-6}}$

Note: The newly discovered genes associated with SBP, DBP, HDL, and BMI are indicated by bold fonts.

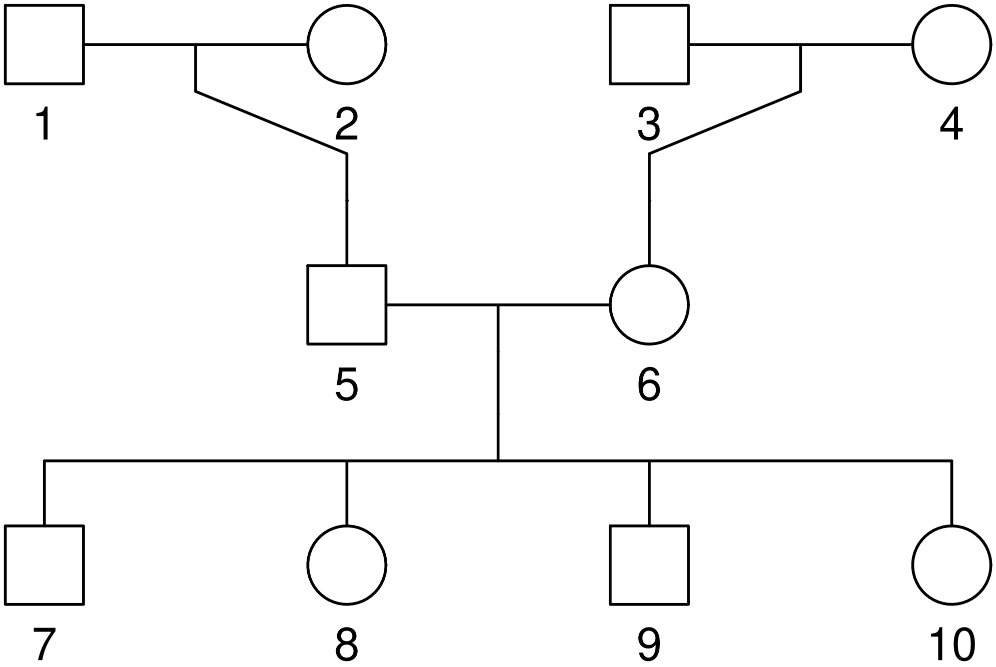


Figure 1.1: Three-generation family pedigree.

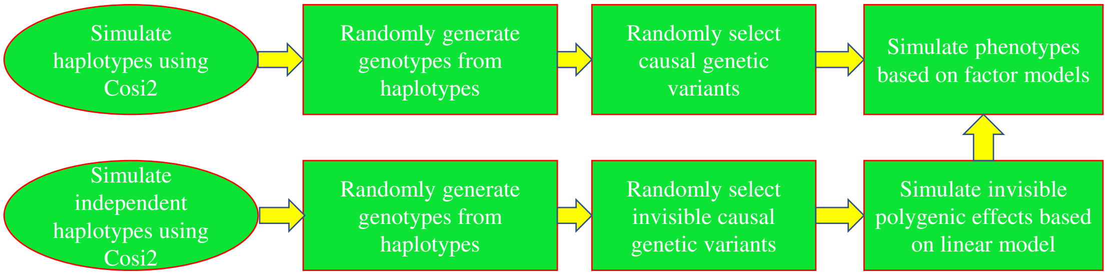

Simulate haplotypes using Cosi2
Randomly generate genotypes from haplotypes
Randomly select causal genetic variants
Simulate phenotypes based on factor models
Simulate independent haplotypes using Cosi2
Randomly generate genotypes from haplotypes
Randomly select invisible causal genetic variants
Simulate invisible polygenic effects based on linear model

Figure 1.2: Workflow in simulation of phenotypes.

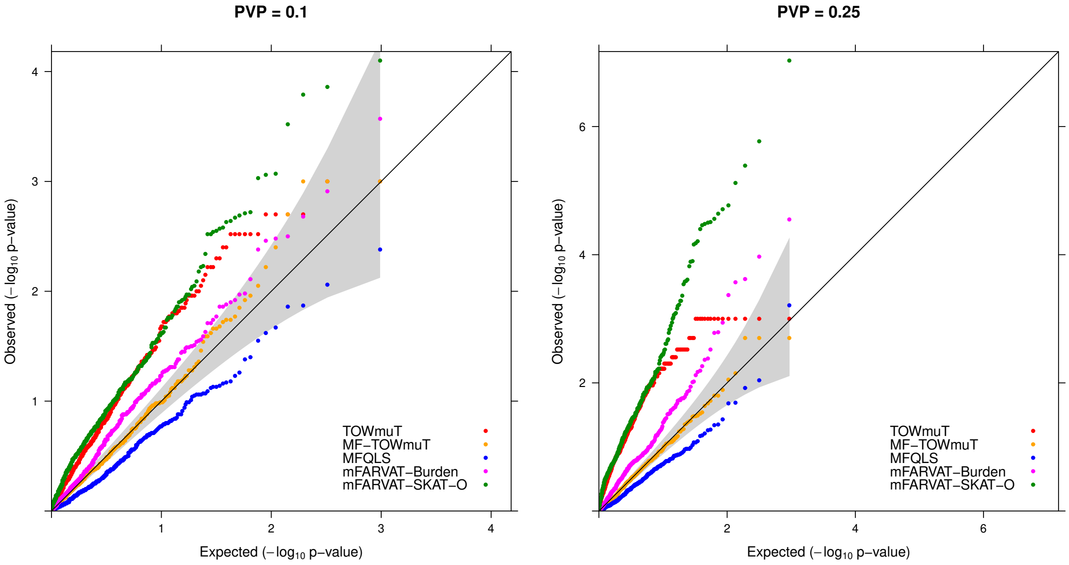


Figure 1.3: QQ plots of P-Values corresponding to Table 1 with 95% confidence band(gray area) based on Model 2. PVP = 0.1 is for left plot, PVP = 0.25 is for right plot. In these two scenarios, we consider mixture of four qualitative and six quantitative phenotypes, two covariates, three-generation pedigree.

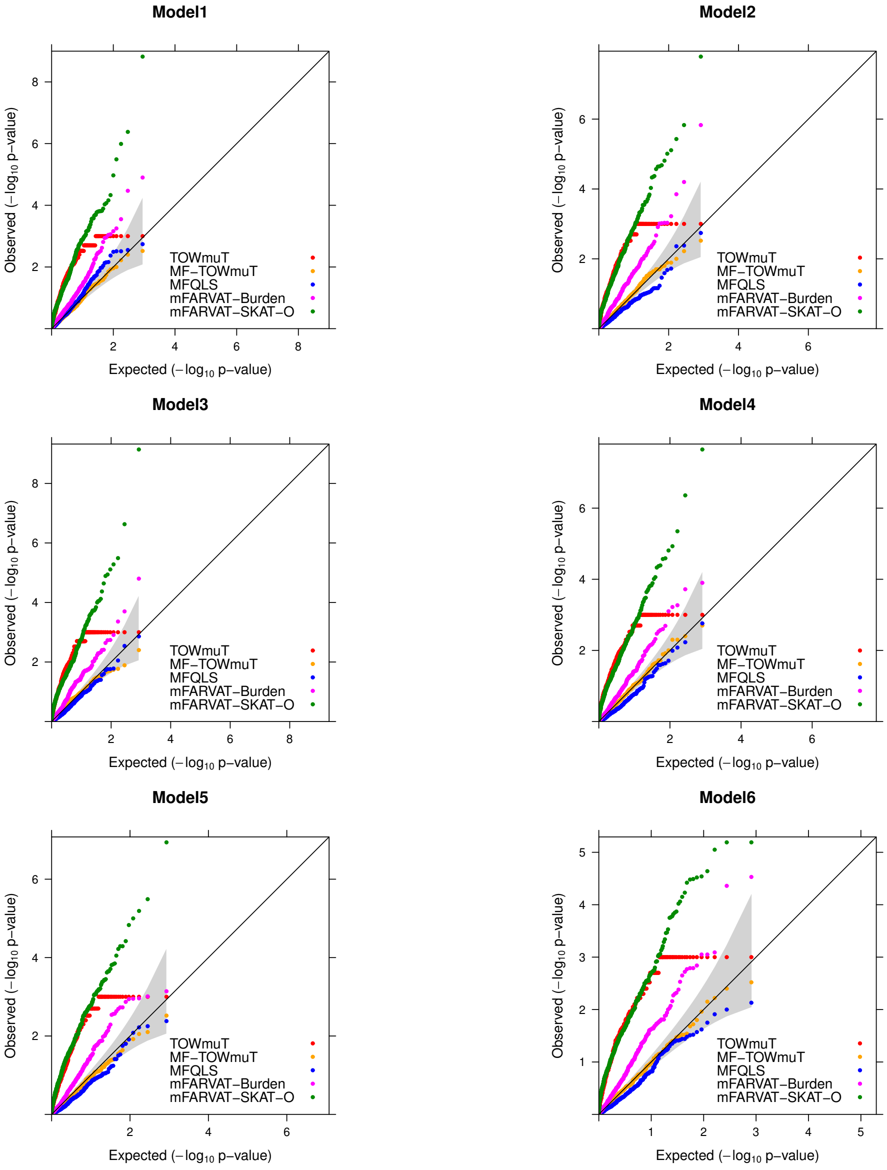


Figure 1.4: QQ plots of P-Values corresponding to Table 2 with 95% confidence band(gray area) for six models. PVP is fixed as 0.5. In these six models, we consider mixture of four qualitative and six quantitative phenotypes, two covariates, three-generation pedigree.

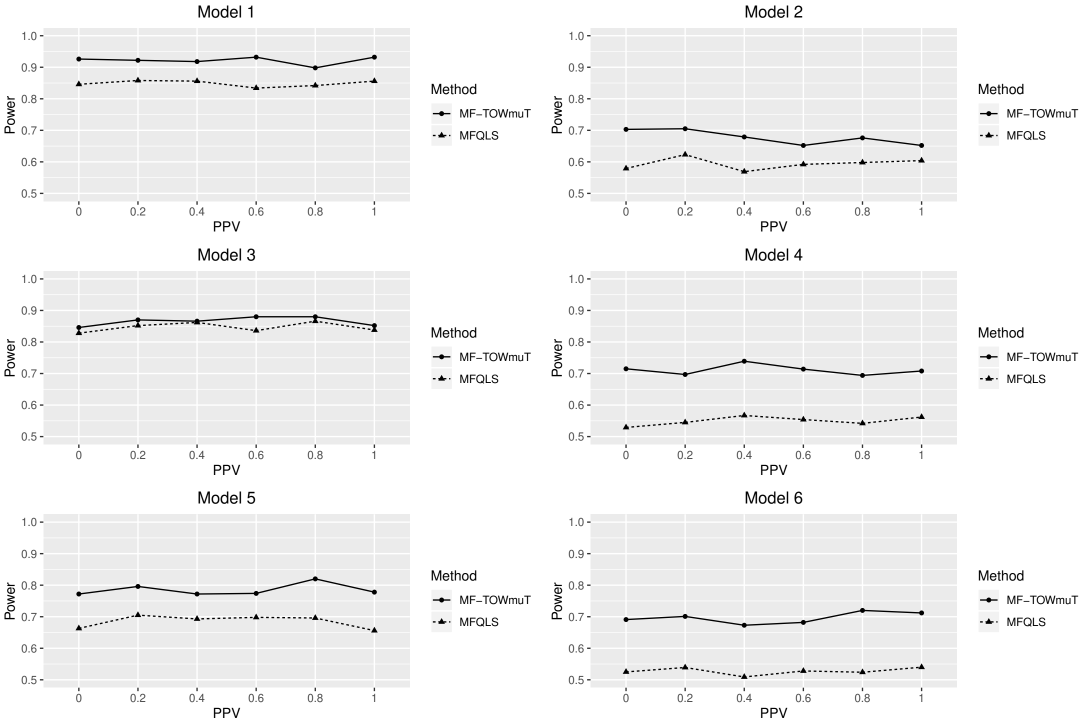


Figure 1.5: Power comparison of MF-TOWmuT and MFQLS for common variants for six models varying proportion of protective variants.

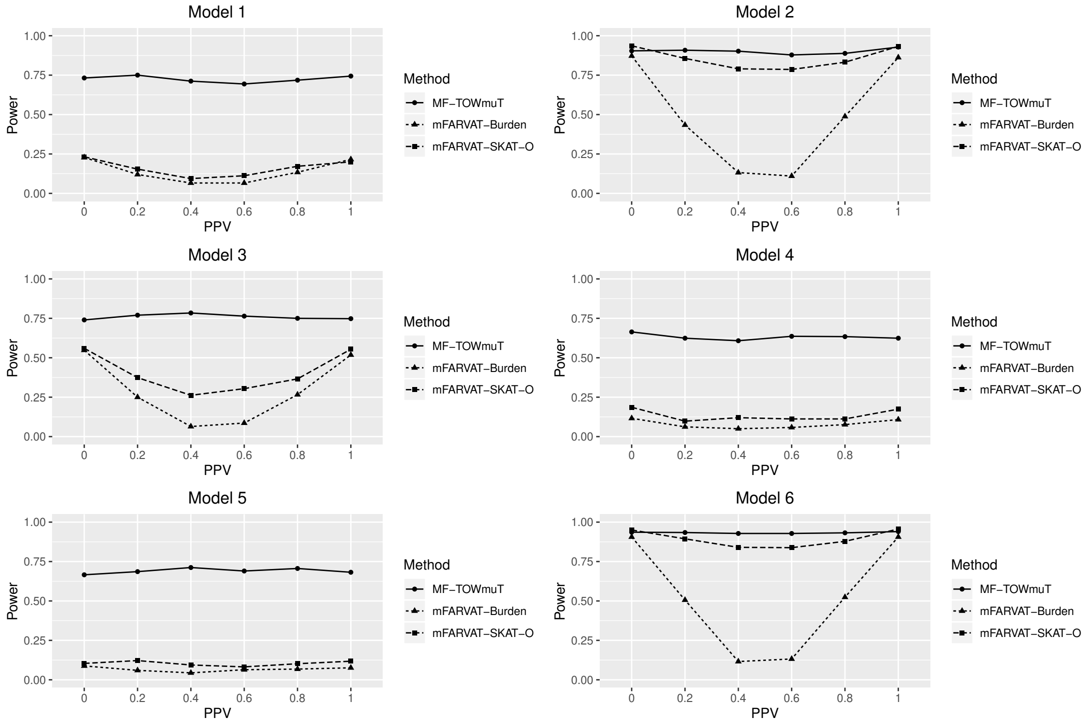


Figure 1.6: Power comparison of MF-TOWmuT and mFARVAT(Burden, SKAT-O) for rare variants for six models varying proportion of protective variants.

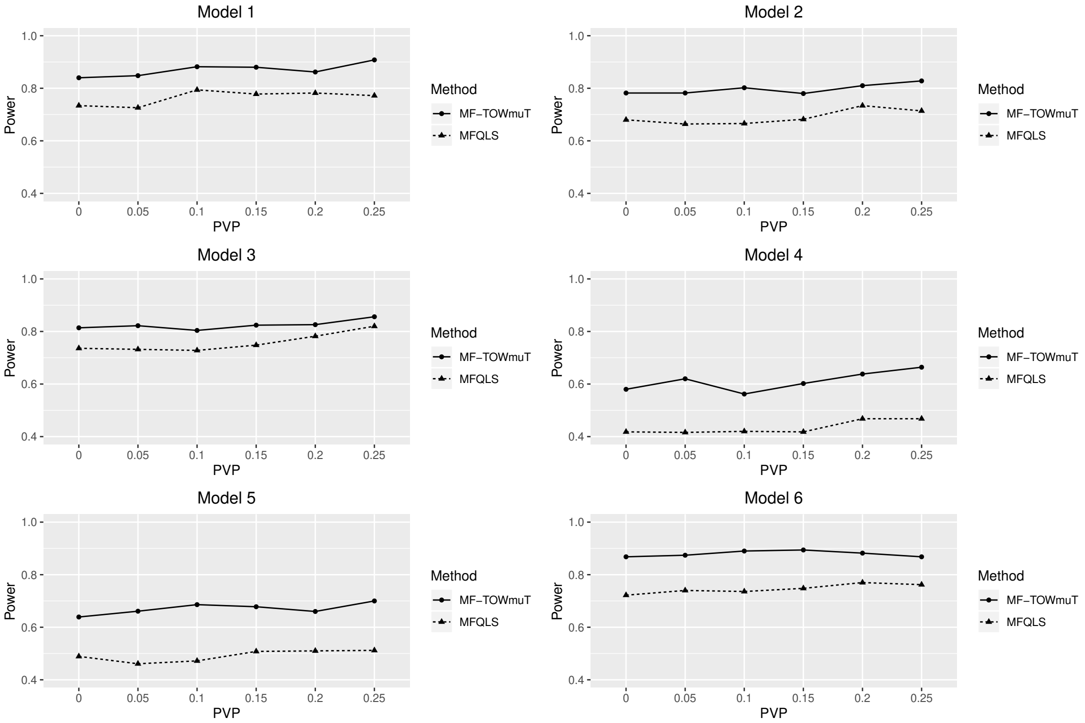


Figure 1.7: Power comparison of MF-TOWmuT and MFQLS for common variants for six models varying PVP.

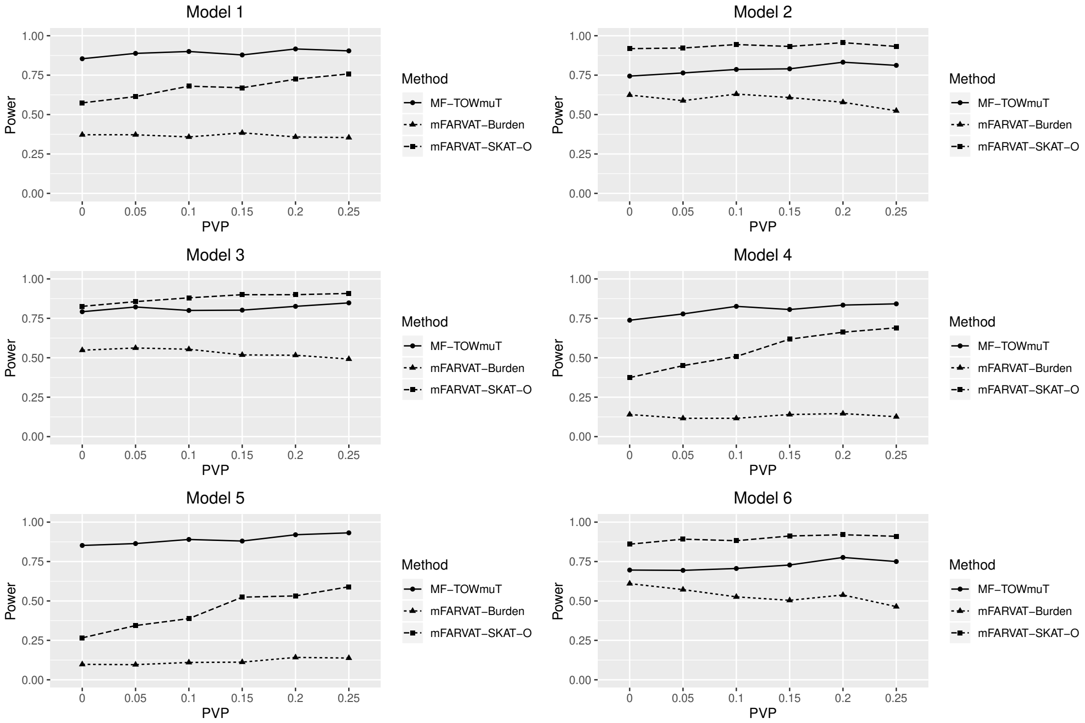


Figure 1.8: Power comparison of MF-TOWmuT and mFARVAT(Burden, SKAT-O) for rare variants for six models varying PVP.

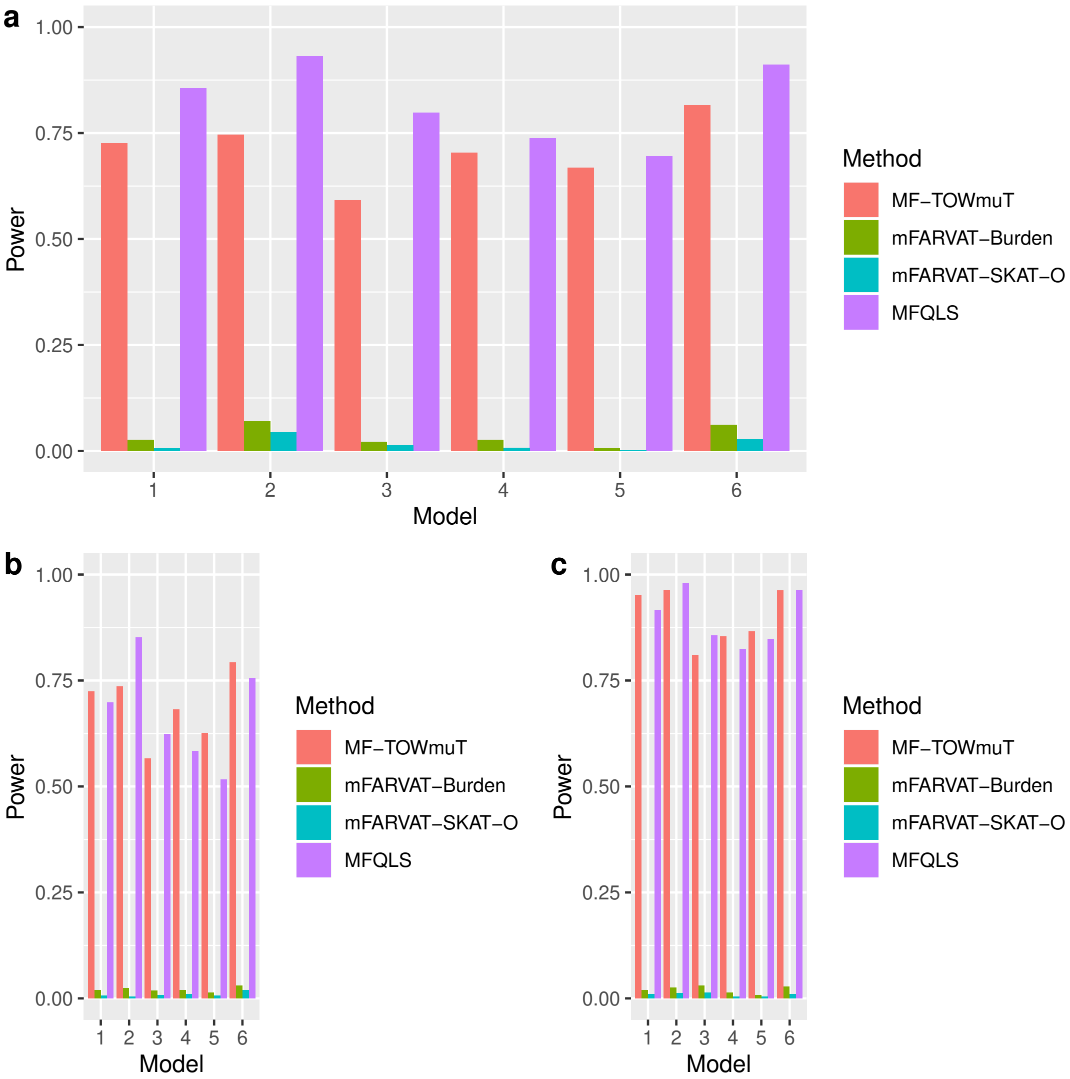


Figure 1.9: Power comparison of MF-TOWmuT, MFQLS, and mFARVAT(Burden, SKAT-O) in three special scenarios: (a) we assumed no invisible polygenic effects and fixed proportion of protective variants as 0.2 with ten quantitative phenotypes for three-generation case, common variants occupied one fifth of total genetic variants; (b) similar to (a) except that rare variants occupied one fifth of total genetic variants; (c) we assumed no invisible polygenic effects and fixed proportion of protective variants as 0.2 with ten quantitative phenotypes for trio case, rare variants occupied half of total genetic variants.

Chapter 2

LORSEN: Fast and Efficient eQTL Mapping With Low Rank Penalized Regression

Abstract

Characterization of genetic variations that are associated with gene expression levels is essential to understand cellular mechanisms that underline human complex traits. Expression quantitative trait loci (eQTL) mapping attempts to identify genetic variants, such as single nucleotide polymorphisms (SNPs), that affect the expression of one or more genes. With the availability of a large volume of large scale gene expression data, it is necessary and important to develop fast and efficient statistical and computational methods to perform eQTL mapping for such large scale data. In this paper, we propose a new method, the low rank penalized regression method, for eQTL mapping (LORSEN). We evaluate and compare the performance of LORSEN with two existing methods for eQTL mapping using extensive simulations as well as real data from the HapMap3 project. Simulation studies show that our method outperforms two commonly used methods for eQTL mapping, LORS and FastLORS, in many scenarios in terms of average Area Under the Curve (AUC). We illustrate

our method by applying it to SNP variants data and gene expression levels on four chromosomes in the HapMap3 Project.

2.1 Introduction

With rapid advancements in sequencing technologies and high-throughput technologies, a large number of single nucleotide polymorphism (SNP) data and gene expression data have become available. This allows us to investigate the associations between SNP genotypes and gene expression levels. Expression quantitative trait loci (eQTLs) are those genetic variants that can explain variation in gene expression levels and can help to elucidate the underlying genetic mechanisms of human complex traits (Albert and Kruglyak, 2015). eQTL mapping aims to identify eQTLs associated with genes of interest (Hu et al., 2015; Banerjee et al., 2020). In general, eQTLs are classified into two types: *cis*-eQTLs (or local eQTLs) and *trans*-eQTLs (or distant eQTLs) (Cookson et al., 2009). *cis*-eQTLs refer to the genetic variants that functionally act on local genes and are physically located near the target genes. *trans*-eQTLs are those genetic variants that functionally act on distant genes residing on the same chromosome or genes residing on different chromosome and are physically located far from the target genes. It is worth mentioning that *trans*-eQTLs account for a large proportion of heritability of gene expression levels, though *trans* effects are usually weaker than *cis* effects in humans (Cookson et al., 2009).

In fact, gene expression levels are not only regulated by genetic variants but also influenced by non-genetic factors which are known or hidden, for example, batch effects. Therefore, in eQTL mapping, how to account for confounding factors is an important issue and can influence the detection power of eQTLs. Up to now, a number of methods have been proposed to account for confounding factors in eQTL mapping, for example, PANAMA (Fusi et al., 2012), PEER (Stegle et al., 2010), LORS (Yang et al., 2013), HEFT (Gao et al., 2014), LMM-EH-PS (Listgarten et al.,

2010) and ECCO (Yue et al., 2020). Another challenge in eQTL mapping is that the number of SNPs involved is usually very large (Yang et al., 2013). This will result in heavy computational burden for estimating model parameters but also will generally result in reduced detection power if all SNPs are included in eQTL mapping. This is because the signal-to-noise ratio (SNR) is very low, meaning only a very small portion of SNPs that are actually associated with gene expression levels. To overcome this problem, a number of SNP screening procedures (Yang et al., 2013; Jeng et al., 2020; Wang et al., 2011) and variable selection techniques (Fan and Lv, 2008) that aim to reduce the number of SNPs and only keep informative SNPs in eQTL mapping have been developed. More importantly, a number of methods based on the penalized regression have been developed to model such sparsity of eQTLs (Yang et al., 2013; Jeng et al., 2020; Cheng et al., 2014; Lee and Xing, 2012).

LORS, a method based on the low rank sparse regression, was proposed for eQTL mapping in (Yang et al., 2013). LORS is based on a linear model with gene expression levels as response variables and SNP genotypes as predictors. To model the sparsity of regression coefficients, LORS poses the L_1 penalty on the regression coefficient matrix. In addition, LORS includes one unknown matrix with the nuclear norm penalty to account for variations caused by non-genetic factors. Yang et al. (2013) applied the coordinate descent algorithm to optimize the objective function and estimate the model parameters. A SNP screening method, called LORS-Screening, was also developed to reduce the number of SNPs involved in the subsequent joint modeling, thus reduce the computational burden. Similar to LORS, FastLORS (Jeng et al., 2020) employs the same low rank sparse regression model that is used in LORS. Different from LORS, FastLORS uses generic proximal gradient algorithm to optimize the objective function and estimate the model parameters. Moreover, Jeng et al. (2020) proposed a SNP screening method based on the Higher Criticism (HC) statistic, called HC-Screening.

To improve the detection power of eQTL mapping, a number of methods have been proposed to incorporate the structure information from SNP variants data and gene expression levels, for example, clustering based on gene expression levels (Chun and Keles, 2009; Kendziorski et al., 2006) and gene regulatory networks (Rakitsch and Stegle, 2016), into eQTL mapping. A number of studies have shown that such use of structure information from SNP variants data and gene expression levels can be effectively used in penalized regression to boost the detection power of eQTL mapping (Chen et al., 2012; Kim and Xing, 2012, 2009). For example, the graph-regularized dual lasso (GDL) proposed by (Cheng et al., 2014) can simultaneously integrate the correlation structures among SNPs and gene expression levels. Through extensive experimental evaluations, Cheng et al. (2014) showed that GDL significantly outperformed the existing method for eQTL mapping. Similar to GDL, the graph-guided fused lasso (GFlasso) proposed by (Lee and Xing, 2012) can also consider the structure of the genetic variants and the structure of the gene expression levels. As a penalized regression method, GFlasso also inherits the benefits from the group lasso. Lee and Xing (2012) showed that GFlasso was able to detect weak association signals between the genetic variants and the gene expression levels.

However, there are some drawbacks for most of the aforementioned methods. First, if two SNPs are highly correlated with each other, and one SNP is associated with some genes, but the other SNP is not associated with them, we should not expect that these two SNPs have similar coefficients for those genes. Similarly, if some SNPs are classified into one group, we should not expect that the SNPs within the same group have similar coefficients for common genes. Second, the group structures of SNP data and gene expression data are usually identified by performing clustering on the data, however, clustering is an unsupervised leaning approach, the number of clusters is usually artificially determined. When we use the resulting clusters of SNPs and gene expressions to design the penalty term, it may lead to loss of detection power and even spurious associations. Third, complicated design of penalty term in penalized regression modeling can result in untractable computational bottleneck, especially when dealing with large volumn of data.

To overcome some limitations of existing methods for eQTL mapping, we propose a novel method for eQTL mapping, **LO**w **R**ank **S**parse regression with **E**lastic **N**et penalty, abbreviated as LORSEN. Different from LORS (Yang et al., 2013) and FastLORS (Jeng et al., 2020), we apply the Elastic Net penalty to the association coefficients instead of the L_1 penalty in LORSEN. In addition, we use the low rank approximation to account for non-genetic factors in LORSEN (Yang et al., 2013). There are several advantages to use the Elastic Net penalty instead of the L_1 penalty (Tibshirani, 1996). First, when the number of SNPs p is much larger than the sample size n, theoretically, the methods based on the L_1 penalty can only yield at most n non-zero coefficients. This can lead to the substantial loss of detection power in eQTL mapping since the number of samples is generally much smaller than the number of eQTLs in gene expression studies. Second, when several eQTLs are in linkage disequilibrium (LD), the methods based on the L_1 penalty can only select one of them. In theory, the Elastic Net penalty can overcome these two drawbacks. For the estimation of the model parameters in LORSEN, we develop an efficient optimization algorithm based on the proximal gradient method (Parikh and Boyd, 2014). Our algorithm allows us to perform the eQTL mapping for a large number of SNPs and genes. We evaluate and compare the performance of LORSEN with LORS and FastLORS using extensive simulation studies as well as the HapMap3 data.

2.2 Material and Methods

2.2.1 Model

We assume that the collected data are the genotypes for p SNPs and the gene expression levels for q genes over n samples. Let X denote the $n \times p$ matrix of SNP genotypes coded in an additive manner, and Y denote the $n \times q$ matrix of gene expression levels. To model the association between SNPs and gene expressions, we can use the linear model as proposed in (Yang et al., 2013):

$$Y = XB + L + \mathbf{1}\mu^T + e, \tag{2.2.1}$$

where B is a $p \times q$ matrix for the regression coefficients, $\mathbf{1}$ is a n-dimensional all-ones vector, μ is a q-dimensional vector for the intercepts in the regression model, e is a $n \times q$ matrix for the error terms and each element in e has a normal distribution with zero mean and variance σ^2, all e_{ij} are independent, L is $n \times q$ matrix which is introduced to account for variations caused by non-genetic factors.

For the convenience of description, we first introduce the following notations used in this paper. For a n-dimensional vector v with the elements $v_i (i = 1, \cdots, n)$: the L_1 norm of v is defined as $\|v\|_1 = \Sigma_{i=1}^n |v_i|$; the L_2 norm of v is defined as $\|v\|_2 = \sqrt{\Sigma_{i=1}^n v_i^2}$, respectively. For a $m \times n$ matrix M with the elements $M_{ij} (i = 1, \cdots, m; j = 1, \cdots, n)$, the Frobenius norm of M is defined as $\|M\|_F = \sqrt{\Sigma_{i=1}^m \Sigma_{j=1}^n M_{ij}^2}$; the nuclear norm $\|M\|_* = \Sigma_{i=1}^r \sigma_i$, where $\sigma_1, \cdots, \sigma_r$ are the singular values of M and r is the rank of M; and the L_1 norm of M is defined as $\|M\|_1 = \Sigma_{i=1}^m \Sigma_{j=1}^n |M_{ij}|$.

In this paper, we follow the same sparsity assumptions used in (Yang et al., 2013). First, there are only a few non-genetic factors that influence the gene expression levels globally, not locally. Second, there are only a small fraction of SNPs that influence the gene expression levels, which implies that the regression coefficient matrix B should be sparse. Yang et al. (2013) proposed the following LORS procedure to estimate B, L, μ by solving the optimization problem

$$\underset{B,L,\mu}{min} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda\|B\|_1, \tag{2.2.2}$$

where ρ and λ are regularization (tuning) parameters that control the rank of L and the sparsity of B, respectively. When L and μ are fixed, the optimization problem becomes a least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) problem with respect to B. As pointed out in (Zou and Hastie, 2005), the Lasso has some limitations that affect its usefulness. First, when $n < p$ (the number of samples is less than the number of SNPs), the Lasso selects at most n SNPs. In the context of eQTL mapping, there are usually a small number of samples available. Even though the proportion of SNPs that are associated with the gene expression levels is small, it is highly likely that the number of SNPs associated with the gene expressions can still be larger than the number of samples. In this case, the L_1 penalty on B will fail to identify some SNPs that are associated with the gene expressions. Second, the Lasso tends to select only one variable among a group of highly correlated variables. This can be problematic in eQTL mapping. For example, if two SNPs are in high linkage disequilibrium and both of them are associated with gene expressions, only one SNP will be selected by the Lasso. Furthermore, if two SNPs are in high linkage disequilibrium and only one of them is associated with gene expressions, the selected SNP by the Lasso may not even be associated with gene expressions.

The use of the Elastic Net penalty (Zou and Hastie, 2005) instead of the L_1 penalty on B can overcome the limitations of the Lasso. Therefore, we propose the following optimization problem to estimate B, L, μ:

$$\underset{B,L,\mu}{min} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2, \tag{2.2.3}$$

where ρ, λ_1 and λ_2 are non-negative tuning parameters. For real data sets, it is quite possible that some entries in Y are unobserved (missing). In such scenarios, the missing data will not be used in (2.2.3). As done in (Yang et al., 2013), we use Ω to index the observed entries in Y. Specifically, Ω is a $n \times q$ matrix with the entry

$$\Omega_{ij} = \begin{cases} 0, & Y_{ij} \quad missing \\ 1, & otherwise. \end{cases} \tag{2.2.4}$$

Then we define the projection of a matrix A onto Ω as $\tilde{A} = P_\Omega(A) = \Omega \odot A$, where A has the same dimension as Ω and $\odot$ represents Hadamard product, that is, $\tilde{A}_{ij} = A_{ij} \times \Omega_{ij}$. Based on the observed data, the optimization problem becomes

$$\underset{B,L,\mu}{min} \quad \frac{1}{2}\|P_\Omega(Y - XB - L - \mathbf{1}\mu^T)\|_F^2 + \rho\|L\|_* + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2. \tag{2.2.5}$$

2.2.2 Theory & Algorithm

To solve the optimization problem in (2.2.5) efficiently, we develop a fast and effective algorithm based on proximal gradient method (Parikh and Boyd, 2014).

We first describe the proximal gradient method for a general optimization problem

$$\underset{x}{min} \quad f(x) = g(x) + h(x), \tag{2.2.6}$$

where $g(x)$ is a convex and differentiable function, $h(x)$ is a closed proper convex which means $h(x)$ is a convex function, the epigraph of $h(x)$ is closed and $h(x) < +\infty$ for at least one x and $h(x) > -\infty$ for every x. Furthermore, we assume that $\nabla g(x)$, the gradient of $g(x)$, is Lipschitz continuous with constant ℓ, which implies that $\nabla^2 g(x) \preceq \ell \mathbf{I}$. Two symmetric matrices of the same dimensions A and B has the relationship $A \preceq B$, if $B - A$ is positive semidefinite. Then we have

$$f(x) = g(x)+h(x) \leqslant g(x_0)+\langle \nabla g(x_0), x-x_0\rangle+\frac{1}{2t}\|x-x_0\|^2+h(x), \quad t \in (0, \frac{1}{\ell}], \tag{2.2.7}$$

where x_0 is an arbitrary point in the domain of $f(x)$ and $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors. Instead using the optimization problem (2.2.6), we focus on minimizing an upper bound of the objective function, that is,

$$\underset{x}{min} \quad g(x_0) + \langle \nabla g(x_0), x - x_0\rangle + \frac{1}{2t}\|x - x_0\|^2 + h(x), \quad t \in (0, \frac{1}{\ell}], \tag{2.2.8}$$

which can be interpreted as an application of majorization-minimization algorithm (Parikh and Boyd, 2014). The optimization problem in (2.2.8) is equivalent to the following optimization problem:

$$\underset{x}{min} \quad \frac{1}{2t}\|x - (x_0 - t\nabla g(x_0))\|^2 + h(x). \tag{2.2.9}$$

Problem (2.2.9) can be solved with an iterative procedure: given the value of x at the k-th iteration, i.e., x_k, the value of x at the $k+1$-th iteration, x_{k+1} can be updated by the following formula

$$x_{k+1} = \underset{x}{argmin} \frac{1}{2t}\|x - (x_k - t\nabla g(x_k))\|^2 + h(x) = Prox_{t,h}(x_k - t\nabla g(x_k)),$$

where $Prox(\cdot)$ is called proximal operator. The iterative process is repeated until the stopping criterion is satisfied or the maximum number of iterations is reached.

To solve the optimization problem (2.2.5), we adopt an alternating optimization approach that is similar to the method in (Yang et al., 2013). Note that in the following part, t_L, t_B, and t_μ are like t used in problem (2.2.9) and correspond to the variables L, B, and μ, respectively.

First, for fixed B and μ, (2.2.5) becomes

$$\underset{L}{min} \quad \frac{1}{2}\|Y - XB - \mathbf{1}\mu^T - L\|_F^2 + \rho\|L\|_*. \tag{2.2.10}$$

In the setting of optimization problem (2.2.10), $\frac{1}{2}\|Y - XB - \mathbf{1}\mu^T - L\|_F^2$ plays the role of $g(x)$ and $\rho\|L\|_*$ plays the role of $h(x)$ in (2.2.6). By Lemma A.2.1, at the $k+1$-th iteration, we have

$$\begin{aligned} L_{k+1} &= Prox_{t_L, \rho\|\cdot\|_*}(L_k - t_L(XB_k + \mathbf{1}\mu_k^T + L_k - Y)) \\ &= S_{t_L\rho}(L_k - t_L(XB_k + \mathbf{1}\mu_k^T + L_k - Y)), \end{aligned}$$

where $S_{t_L\rho}(\cdot)$ is the singular value shrinkage operator (Appendix), t_L is the step size which can be constant or be determined by backtracking line search.

Second, for fixed L and μ, then (2.2.5) becomes

$$\underset{B}{min} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2, \tag{2.2.11}$$

where t_B is the step size which can be constant or be determined by backtracking line search. By Lemmas A.2.2 and A.2.3 and Theorem A.2.5, we can update B_{k+1} accordingly:

$$\begin{aligned} B_{k+1}^a &= B_k - t_B X^T(XB_k + \mathbf{1}\mu_k^T + L_{k+1} - Y) \\ B_{k+1}^b &= Prox_{t_B, \lambda_1\|\cdot\|_1}(B_{k+1}^a) \\ &= sign(B_{k+1}^a) \odot (|B_{k+1}^a| - \lambda_1 J)_+ \\ B_{k+1}[, j] &= Prox_{t_B, \lambda_2\|\cdot\|_2}(B_{k+1}^b[, j]) \end{aligned}$$

$$= \{1 - \frac{\lambda_2}{max\{\|B^b_{k+1}[,j]\|_2, \lambda_2\}}\}B^b_{k+1}[,j], \quad j = 1, 2, \cdots, q,$$

where J is a all-ones $p \times q$ matrix, $B[,j]$ is the j-th column of matrix B and is a p-dimensional vector, $\gamma_+ = max\{\gamma, 0\}$, the maximum of γ and 0, $|B^a_{k+1}|$, $sign(B^a_{k+1})$, and $(|B^a_{k+1}| - \lambda_1 J)_+$ are all elementwise operations.

Third, for fixed L and B, the proximal gradient method reduces to the gradient descent method with respect to μ because there is no penalty on μ. At the $k+1$-th iteration, we have

$$\mu_{k+1} = \mu_k - t_\mu(XB_{k+1} + \mathbf{1}\mu_k^T + L_{k+1} - Y)^T\mathbf{1}.$$

To accelerate the computational speed, we will use the accelerated proximal gradient method. Specifically, we apply the fast iterative shrinkage-thresholding algorithm (FISTA) (Beck and Teboulle, 2009) which keeps the simplicity of the iterative shrinkage-thresholding algorithms (ISTA) but has an improved rate $O(1/k^2)$, where k indexes the iteration. In FISTA, the step size can be constant or be determined by backtracking line search. We describe the algorithm to solve LORSEN with FISTA (see Appendix). For simplicity, here, we only describe the detailed algorithm with the constant step size, but provide the algorithms using either the constant step size determined by the backtracking in our R program `https://github.com/gaochengPRC/LORSEN`.

2.2.3 Parameter Tuning

For parameter tuning, we mainly follow the idea described in (Yang et al., 2013). Specifically, we divide the entries of Ω into training entries and testing entries such that training entries and testing entries include roughly the same number of 1's. We define two matrices Ω_1 and Ω_2 such that they have the same dimensions as Ω, Ω_1 contains all training entries and Ω_2 contains all testing entries. Furthermore, we have $\Omega = \Omega_1 + \Omega_2$ and $\Omega_1 \odot \Omega_2 = 0$. For the consistency, we re-parameterize λ_1 and λ_2 as

$\lambda \cdot \alpha$ and $\lambda \cdot (1-\alpha)$, respectively. So the optimization problem (2.2.5) becomes

$$\underset{B,L,\mu}{min} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda(\alpha\|B\|_1 + \frac{1-\alpha}{2}\|B\|_F^2). \tag{2.2.12}$$

This form is the same as that in glmnet (Friedman et al., 2010).

Given the values of parameters (ρ, α, λ), we solve the following optimization problem

$$\underset{B,L,\mu}{min} \quad \frac{1}{2}\|P_{\Omega_1}(Y - XB - L - \mathbf{1}\mu^T)\|_F^2 + \rho\|L\|_* + \lambda(\alpha\|B\|_1 + \frac{1-\alpha}{2}\|B\|_F^2). \tag{2.2.13}$$

The solutions are $B(\rho,\alpha,\lambda)$, $L(\rho,\alpha,\lambda)$ and $\mu(\rho,\alpha,\lambda)$, then we evaluate the parameters by calculating the prediction error

$$Err(\rho,\alpha,\lambda) = \frac{1}{2}\|P_{\Omega_2}(Y - XB(\rho,\alpha,\lambda) - L(\rho,\alpha,\lambda) - \mathbf{1}\mu(\rho,\alpha,\lambda)^T)\|_F^2. \tag{2.2.14}$$

The grid search over three parameters may be too computationally intensive. Therefore, we first find an optimal value for ρ, $\hat{\rho}$, which minimizes the prediction error as shown in (Yang et al., 2013) by means of Lemmas A.2.1 and A.2.4. Please refer to (Yang et al., 2013) to find the details about how to find the optimal value of ρ, $\hat{\rho}$. Once the optimal value of ρ, $\hat{\rho}$ is obtained, we pick up a value of α from a sequence sequentially, thereafter, we perform one-dimensional grid search for λ for each α. Specifically, we generate a sequence of λ values with length n_λ decreasing from $\lambda_{max}(\hat{\rho},\alpha)$ to $\epsilon\lambda_{max}(\hat{\rho},\alpha)$ on the log scale with equal space, where $\lambda_{max}(\hat{\rho},\alpha)$ is defined as the smallest λ such that $B(\hat{\rho},\alpha,\lambda(\hat{\rho},\alpha))$ is a zero matrix. $\lambda_{max}(\hat{\rho},\alpha)$ is derived as $\frac{1}{\alpha} \underset{i=1,2,\cdots,p}{max} \underset{j=1,2,\cdots,q}{max} |\langle X_i, Y_j\rangle|$ from coordinate-descent algorithm (Friedman et al., 2007), where X_i is the i-th column of X, and Y_j the j-th column of Y. In our R program, we set $\epsilon = 0.02$, $n_\lambda = 50$ and $S_\alpha := (0.2, 0.4, 0.6, 0.8, 0.9)$. The optimal parameters are $(\hat{\rho}, \alpha, \hat{\lambda}(\hat{\rho},\alpha))$ that minimize the prediction error. The optimal feasible solutions of B, L, and μ are obtained based on the set of optimal tuning parameters.

2.2.4 SNP Ranking & Joint Modeling

The procedure to select the set of optimal tuning parameters is computationally intensive. Therefore, as it is discussed in (Yang et al., 2013), it may not be computationally tractable to apply such method to the large-scale data sets that contain a large number of gene expressions and SNPs. A commonly used strategy to reduce such computational burden is to choose a subset of SNPs and then only use them in the subsequent eQTL analysis. In this paper, we will use and evaluate two existing methods for pre-selection of informative SNPs: LORS-Screening (Yang et al., 2013) and Higher Criticism Screening (HC-Screening) (Jeng et al., 2020). For LORS-Screening, we first obtain the initial estimate of β_i's by solving

$$\underset{\beta_i, L, \mu}{min} \quad \frac{1}{2}\|Y - X_i\beta_i^T - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_*, \tag{2.2.15}$$

where X_i is the i-th column of X, β_i is a q-dimensional vector for the coefficient of the i-th SNP on q genes, $i = 1, 2, \cdots, p$. For each gene, we select the top n SNPs in terms of the absolute values of association coefficients, then we take union of selected SNPs for each gene as the final set of SNPs to be involved in the joint modeling. For HC-Screening, we first obtain association coefficients as above, then calculate the standardized estimates of coefficients. For each SNP, the Higher Criticism (HC) statistic (Donoho and Jin, 2004) is calculated based on the standardized estimates of coefficients. Then we select the top n SNPs in terms of the p-values of HC statistics.

2.2.5 Simulation Design

Our simulation is similar to that described in (Jeng et al., 2020). We first download the genotype data of Chromosome 1 and Chromosome 21 for CEU samples from HapMap3, the third phase of the International HapMap Project (`https://www.genome.gov/10001688/international-hapmap-project`). CEU samples refer to Utah residents with Northern and Western European ancestry from the CEPH collection. After the quality-control (please refer to Real Data Analysis section), the genotype data of 13,815 SNPs of Chromosome 1 and 2,607 SNPs of Chromosome 21 for $n = 165$ samples are retained in analysis. To simulate gene expression levels for $q = 200$ genes on $n = 165$ samples, we first simulate non-genetic effects of $k = 15$ hidden factors. We randomly generate nk random numbers from $N(0, 1)$ to form a $n \times k$ matrix H, then let $\Sigma = HH^T$. U_j's are simulated from $N(0, 0.1 * \Sigma)$, $j = 1, 2, \cdots, q$ and stacked by column to form a $n \times q$ matrix U. e_j's are simulated from $N(0, I)$ as random noise for j-th gene expression and combined by column to form a $n \times q$ random noise matrix e. Then the expression data of q genes on n samples are simulated by $Y = XB + U + e$, where X is the $n \times p$ genotype data matrix. We set the total number of SNPs $p = 2000$, the number of causal SNPs as 60, 200, or 400. Each causal SNP randomly influences $m = 10$ (or 50) genes. We simulate nonzero genetic effects from a uniform distribution. For the "weak-dense" scenario, each causal SNP affects $m = 50$ randomly selected genes and the corresponding values in B are simulated from a uniform distribution between 0.25 and 0.75. For the "strong-sparse" scenario, each causal SNP affects $m = 10$ randomly selected genes and the corresponding values in B are simulated from a uniform distribution between 1.5 and 2. The different simulation scenarios are summarized in Table 2.1.

2.3 Results

2.3.1 Simulation Results

The number of selected SNPs and the number of selected causal SNPs from two screening methods under different simulation scenarios are summarized in Table 2.2. Several conclusions emerge from Table 2.2. First, when the number of samples is much less than the number of SNPs and the number of causal SNPs is larger than the number of samples, HC-Screening is seemingly not an appropriate screening tool. This is because the number of causal SNPs retained after the HC-Screening is much smaller than the actual number of causal SNPs, resulting in possible power loss in subsequent analysis. Second, even when the number of causal SNPs is smaller than the number of samples (n), from Table 2.2, we still observe that the LORS-Screening retains more causal SNPs than the HC-Screening. Of course, the HC-Screening reduces much computational burden especially when the number of samples is much less than the number of SNPs.

The area under the curve (AUC) is used to compare the performance between LORSEN and two existing methods, LORS (Yang et al., 2013) and FastLORS (Jeng et al., 2020). For each scenario, we replicate the simulation ten times. We consider joint modeling of multiple SNPs and multiple gene expression levels with the SNP screening and without the SNP screening. The results without the SNP screening before the eQTL mapping under different simulation scenarios are presented in Tables 2.3 and 2.4. From Tables 2.3 and 2.4, we can see that the average AUC of LORSEN is uniformly larger than those of LORS and FastLORS in the weak-dense scenarios across different number of causal SNPs no matter the SNPs are from single chromosome (Chr 1) or two chromosomes (Chr 1 + Chr 21). For the strong-sparse scenarios, FastLORS achieves the relatively larger AUC than LORS and LORSEN. For a fixed number of causal SNPs, each method achieves the larger AUC value in the stong-sparse scenario than in the weak-dense scenario. For each method under each simulation scenario, the AUCs in Tables 2.3 and 2.4 are similar, implying that each of three methods has the similar power to detect *cis*-eQTLs and *trans*-eQTLs.

The results with the SNP screening before eQTL mapping under different simulation scenarios are presented in Tables 2.5 and 2.6. As we have mentioned, the LORS-Screening keeps more SNPs in the analysis, thus retains more causal SNPs than the HC-Screening does. Each method with the LORS-Screening has the larger AUC values than it with the HC-Screening. From Tables 2.5 and 2.6, we can see that the AUC values of methods with the HC-Screening are quite close to 0.5, which indicates that the HC-Screening can essentially lead to the loss of power of methods. With the LORS-Screening, similar to the non-screening cases, LORSEN has better performance than LORS and FastLORS in the weak-dense scenarios and LORSEN and FastLORS perform similarly and slightly better than LORS in the strong-sparse scenarios. Finally, we find that for the weak-dense scenarios, each method without the SNP screening before joint modeling achieves the larger AUC values than it with the SNP screening. However, for the strong-sparse scenarios, each method with the LORS-Screening before joint modeling achieves the larger AUC values than it without the SNP screening. This may be due to that there are a large number of SNP-gene pairs with the weak association effects in the weak-dense scenarios and many causal SNPs may not be selected by the pre-screening methods. So, in the weak-dense scenarios with the use of pre-screening methods, the computational cost and the detection power can be reduced at the same time. In the strong-sparse scenarios, there are a smaller number of SNP-gene pairs with the stronger association effects than in the weak-dense scenarios, and it is expected that most of the causal SNPs will be selected by the pre-screening methods. Therefore, for the strong-sparse scenarios, the use of pre-screening methods reduce the computational cost while still retain the high detection power.

2.3.2 Real Data Analysis Results

To illustrate our method in real data analysis, we apply LORS-LORSEN (LORSEN with the LORS-Screening), LORS-LORS (LORS with the LORS-Screening) and HC-FastLORS (FastLORS with the HC-Screening) to the HapMap3 data. Specifically, we focus on Asian samples (CHB and JPT) in the HapMap3 data, and we select four chromosomes for the analysis. SNP genotype data and gene expression data are publicly available, and can be downloaded from `ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3_r3/plink_format/` and `http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264/`, respectively. Because the set of samples with the SNP genotype data and the set of samples with the gene expression data are slightly different, we only keep the samples that have both the SNP genotype data and the gene expression data in the analysis. We remove SNPs with missing values, and perform the LD pruning using Plink with its default parameters (window size: 50; moving window increment: five SNPs; cutoff value of R^2: 0.5). After the data pre-processing, a total of 160 samples (CHB: 79; JPT: 81) are included in analysis. The number of SNPs and the number of genes with the expression used in the analysis are summarized as the following: Chromosome 3: 4,086 SNPs and 1,075 genes with the expression; Chromosome 15: 2,235 SNPs and 612 genes with the expression; Chromosome 17: 2,226 SNPs and 1,098 genes with the expression; Chromosome 20: 1,863 SNPs and 606 genes with the expression. Since the significance tests generally cannot be performed for the penalization based regression models, we focus on the top 100 detected SNP-gene pairs with the largest absolute regression coefficients. From the Venn diagrams (Figures 2.1 - 2.4), we notice that there is a big overlap between the eQTLs identified by LORS-LORS and LORS-LORSEN. However, there is little overlap between the eQTLs identified by HC-FastLORS and LORS-LORS (or LORS-LORSEN). For example, among the top 100 SNP-gene pairs identified on Chromosome 3 (Figure 2.1), LORS-LORS and LORS-LORSEN share 77 SNP-gene pairs in common, while LORS-LORSEN and HC-FastLORS share four SNP-gene pairs in common and LORS-LORS and HC-FastLORS share three SNP-gene pairs in common. This observation is consistent with the observation from (Jeng et al., 2020) who also noticed that there is little overlap between the SNP-gene pairs identified by LORS-LORS and HC-FastLORS. Additionally, as adopted in (Jeng et al., 2020), we classified the detected eQTL as local if the distance between the base pair position of the SNP and the probe midpoint is less than 250kb or as distant if the distance is greater than 5mb using the method in (Westra et al., 2013). For each chromosome, we report our findings on the top ten identified SNP-gene pairs in Tables 2.7 - 2.10. The results of top ten SNP-gene pairs identified on Chromosome 3 are summarized in Table 2.7. We can see that in the top ten SNP-gene pairs identified by HC-FastLORS, the SNPs are all *trans*-eQTLs. As a comparison, in the top ten SNP-gene pairs identified by LORS-LORSEN, seven SNPs are *cis*-eQTLs and two SNPs are *trans*-eQTLs; in the top ten SNP-gene pairs identified by LORS-LORS, five SNPs are *cis*-eQTLs and four SNPs are *trans*-eQTLs. LORS-LORSEN and LORS-LORS share seven SNP-gene pairs while LORS-LORSEN and LORS-LORS do not share any SNP-gene pair with HC-FastLORS. In addition, the coefficients obtained from HC-FastLORS are ten-fold smaller than those obtained from LORS-LORSEN and LORS-LORS.

2.4 Disussion

As more human gene expression data become available, fast and efficient statistical and computational methods are needed to fully take advantage of such data to investigate the relationship between genetic variants and gene expression levels to further reveal the genetic mechanisms that underlie human complex diseases. However, most existing methods are built on small-scale samples and are not applicable to human-size datasets. In this paper, we propose a new low rank penalized regression method (LORSEN) to detect eQTLs. We develop a fast and efficient algorithm to solve optimization problems arising from our methods based on proximal gradient methods. Comprehensive simulation studies show that LORSEN outperforms two commonly used methods, LORS and FastLORS, under some simulation scenarios.

Since there are a large number of SNPs and genes to be included in the eQTL mapping and it is expected that only a small portion of SNPs will affect the gene expression levels, a number of pre-screening methods have been developed. In this paper, we used the LORS-Screening (Yang et al., 2013) and the HC-Screening (Jeng et al., 2020). We find that the HC-Screening retains much smaller number of SNPs than the LORS-Screening. Both the LORS-Screening and the HC-Screening can reduce the computational cost, but they may reduce the detection power in the eQTL mapping, depending on the association patterns between SNPs and gene expression levels. Since we do not know such association patterns in real studies, we should be cautious to apply such pre-screening methods.

There are several limitations for LORSEN. First, as a method based on the penalized regression model, we can rank the SNP-gene pairs according to the coefficients obtained from LORSEN but cannot perform the significance test. Second, we compare the performance of LORSEN with the pre-screening methods and without such methods. The use of the pre-screening methods can reduce the number of SNPs used in the analysis, thus can greatly reduce the computational cost. Third, the grid search is used to find the optimal set of tuning parameters. The grid search is easy to be implemented but is computationally intensive. It may not be feasible for large scale data. A more efficient strategy is desirable.

It has shown that the incorporation of the SNP correlation and the gene interaction network can potentially increase the power of detecting eQTLs (Kim and Xing, 2012; Cheng et al., 2014; Kim and Xing, 2009; Chen et al., 2012). We expect that our method can be improved if we use the prior knowledge of correlation structures of SNPs and genes to refine the penalty terms in optimization problems.

2.5 Tables and Figures

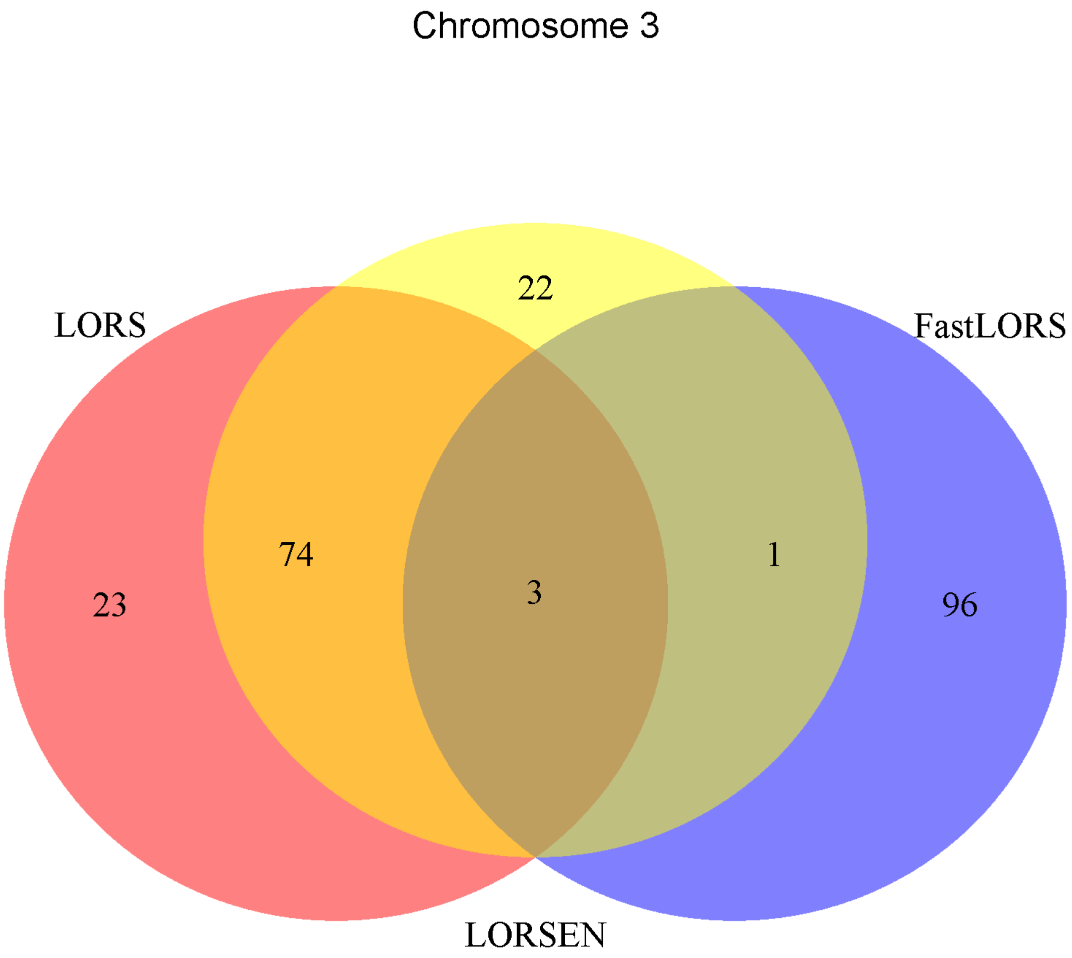


Figure 2.1: The number of SNP-gene pairs identified by FastLORS, LORSEN, and LORS on Chromosome 3.

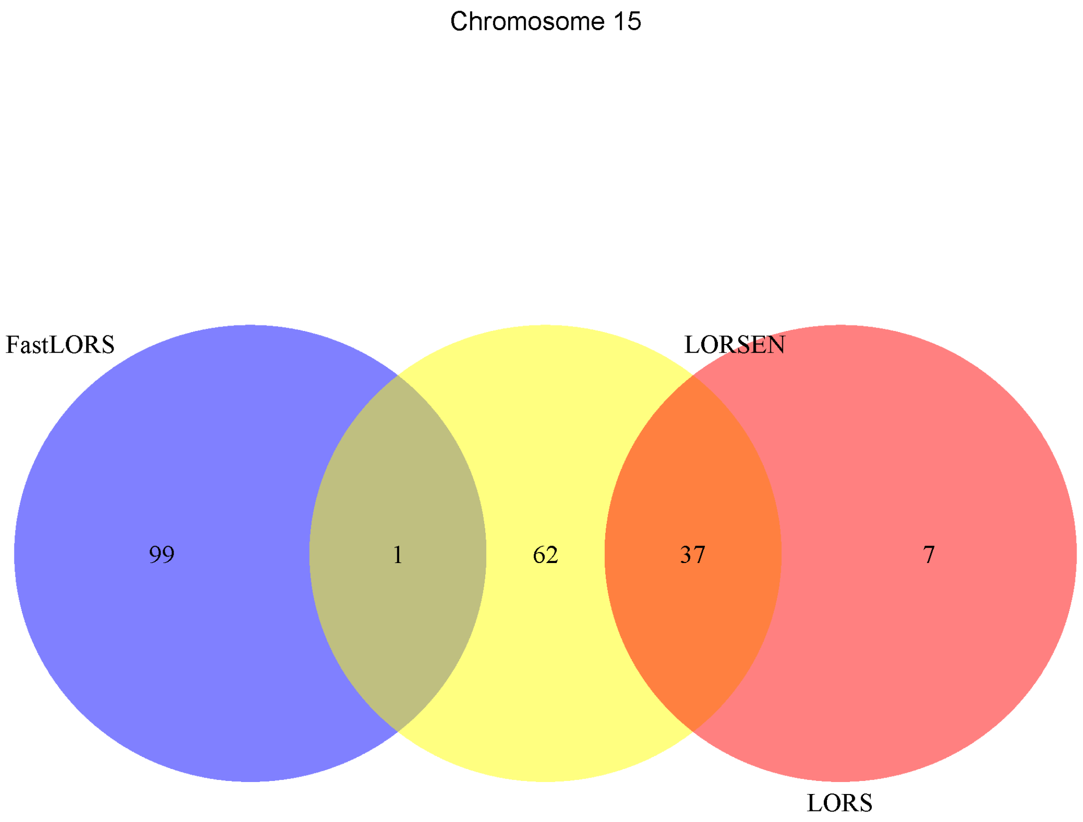


Figure 2.2: The number of SNP-gene pairs identified by FastLORS, LORSEN, and LORS on Chromosome 15.

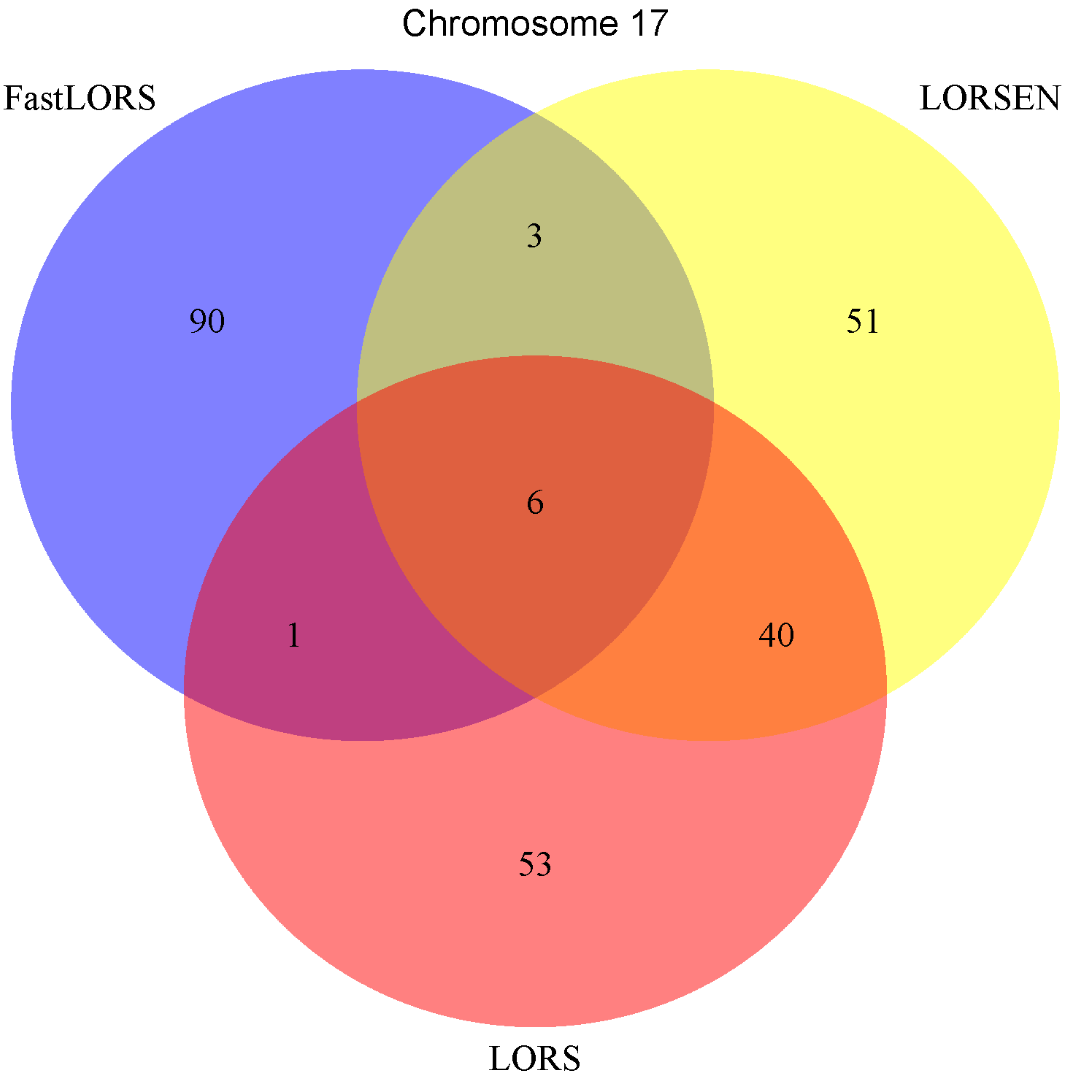


Figure 2.3: The number of SNP-gene pairs identified by FastLORS, LORSEN, and LORS on Chromosome 17.

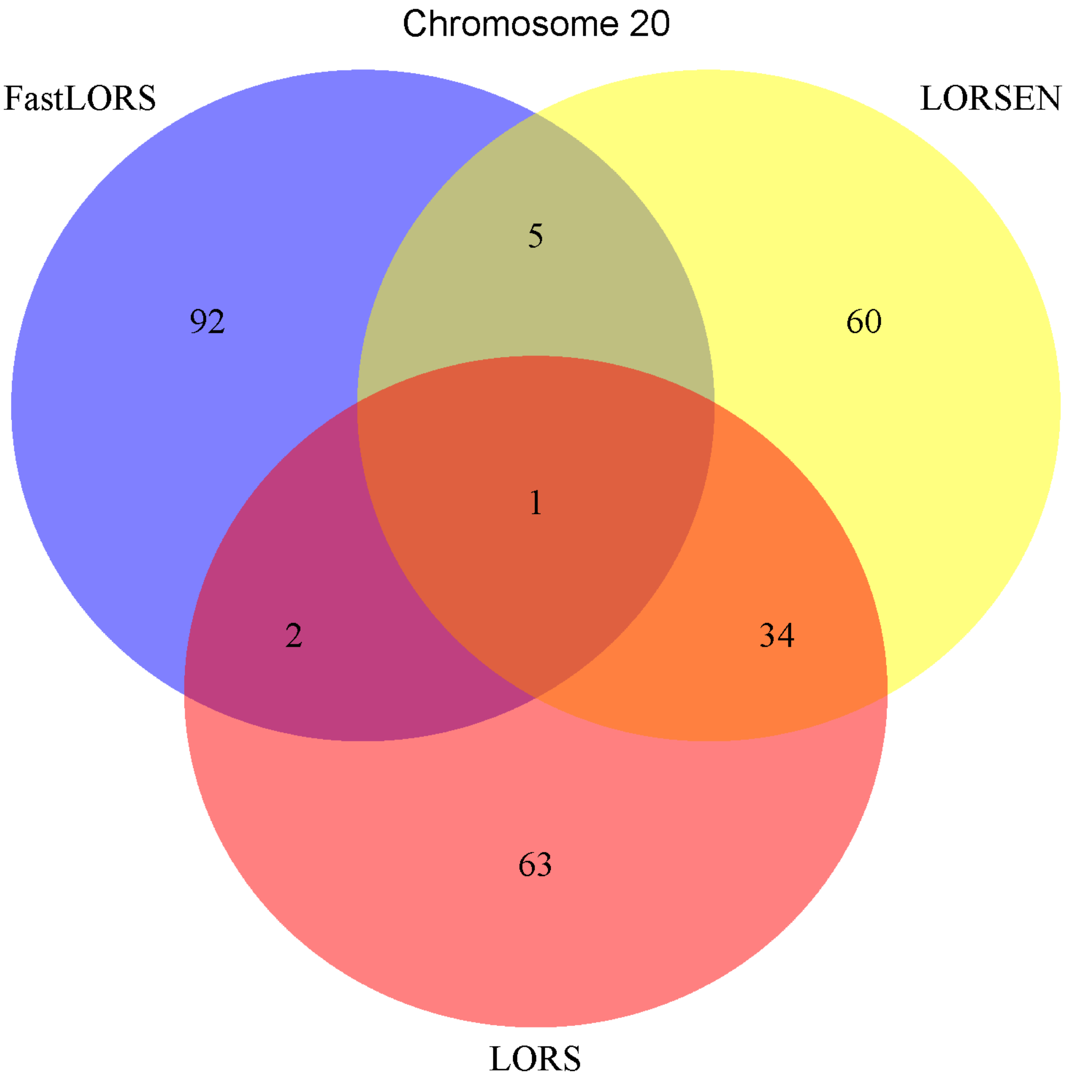


Figure 2.4: The number of SNP-gene pairs identified by FastLORS, LORSEN, and LORS on Chromosome 20.

Table 2.1: Simulation scenarios.

Chromosome	#Causal SNPs	Scenario	Method	Screening
Chr 1	60	weak-dense	FastLORS	LORS
	200	strong-sparse	LORSEN	HC
	400		LORS	
Chr 1 + Chr 21	45 + 15	weak-dense	FastLORS	LORS
	150 + 50	strong-sparse	LORSEN	HC
	300 + 100		LORS	

Table 2.2: Results of HC-Screening and LORS-Screening with ten replicates for each simulation scenario.

Chromosome	#Causal SNPs	Scenario	Screening	Average #Selected SNPs	Average #Selected Causal SNPs
Chr 1	60	weak-dense	LORS	1017	43
			HC	165	7
		strong-sparse	LORS	1023	60
			HC	165	9
	200	weak-dense	LORS	1036	130
			HC	165	20
		strong-sparse	LORS	1095	199
			HC	165	28
	400	weak-dense	LORS	1045	237
			HC	165	39
		strong-sparse	LORS	1142	346
			HC	165	44
Chr 1 + Chr 21	45 + 15	weak-dense	LORS	1044	46
			HC	165	7
		strong-sparse	LORS	1065	60
			HC	165	10
	150 + 50	weak-dense	LORS	1064	136
			HC	165	20
		strong-sparse	LORS	1123	199
			HC	165	28
	300 + 100	weak-dense	LORS	1064	244
			HC	165	37
		strong-sparse	LORS	1188	361
			HC	165	44

Table 2.3: The average AUC without the SNP screening with ten replicates for each simulation scenario. SNPs are only from chromosome 1.

Scenario	**Method**	**#Causal SNPs**		
		60	**200**	**400**
weak-dense	FastLORS	0.514	0.582	0.581
	LORSEN	**0.651**	**0.649**	**0.630**
	LORS	0.502	0.514	0.515
strong-sparse	FastLORS	0.762	**0.840**	**0.810**
	LORSEN	0.823	0.834	0.774
	LORS	**0.824**	0.819	0.754

Table 2.4: The average AUC without the SNP screening with ten replicates for each simulation scenario. SNPs are from chromosome 1 and chromosome 21.

Screening	**Method**	**#Causal SNPs**		
		60	**200**	**400**
weak-dense	FastLORS	0.530	0.567	0.575
	LORSEN	**0.658**	**0.679**	**0.625**
	LORS	0.503	0.510	0.514
strong-sparse	FastLORS	0.774	**0.826**	**0.813**
	LORSEN	**0.814**	0.810	0.788
	LORS	0.813	0.801	0.756

Table 2.5: The average AUC with the SNP screening with ten replicates for each simulation scenario. SNPs are only from chromosome 1.

Scenario	**#Causal SNPs**	**Method**	**Screening**	
			HC	**LORS**
weak-dense	60	FastLORS	0.514	0.596
		LORSEN	**0.515**	**0.618**
		LORS	0.503	0.541
	200	FastLORS	**0.512**	0.583
		LORSEN	0.511	**0.592**
		LORS	0.502	0.519
	400	FastLORS	**0.510**	**0.557**
		LORSEN	0.509	0.547
		LORS	0.502	0.511
strong-sparse	60	FastLORS	**0.565**	0.900
		LORSEN	0.558	**0.903**
		LORS	0.560	0.897
	200	FastLORS	**0.552**	**0.894**
		LORSEN	0.544	**0.894**
		LORS	0.543	0.874
	400	FastLORS	**0.536**	**0.797**
		LORSEN	0.523	0.782
		LORS	0.528	0.738

Table 2.6: The average AUC with the SNP screening with ten replicates for each simulation scenario. SNPs are from chromosome 1 and chromosome 21.

Scenario	**#Causal SNPs**	**Method**	**Screening**	
			HC	**LORS**
weak-dense	60	FastLORS	**0.518**	0.606
		LORSEN	**0.518**	**0.629**
		LORS	0.505	0.544
	200	FastLORS	**0.512**	0.591
		LORSEN	**0.512**	**0.615**
		LORS	0.503	0.524
	400	FastLORS	**0.510**	**0.563**
		LORSEN	0.507	0.556
		LORS	0.501	0.511
strong-sparse	60	FastLORS	**0.570**	0.891
		LORSEN	0.563	**0.906**
		LORS	0.564	0.891
	200	FastLORS	**0.553**	**0.904**
		LORSEN	0.547	**0.904**
		LORS	0.544	0.883
	400	FastLORS	**0.534**	**0.821**
		LORSEN	0.524	0.813
		LORS	0.525	0.765

Table 2.7: Top ten detected SNP-gene pairs for chromosome 3.

Method	SNP	Probe (Gene)	Association Coefficient	Distance	Class
HC-FastLORS	rs13084976	ILMN_1657373(LEPREL1)	0.0430	188.72mb	distant
	rs17029694	ILMN_1657373 (LEPREL1)	0.0424	188.49mb	distant
	rs12494696	ILMN_1812093 (UTS2D)	0.0322	189.72mb	distant
	rs2322212	ILMN_1756501 (ST6GAL1)	0.0310	184.74mb	distant
	rs17029694	ILMN_1708743 (NT5DC2)	0.0303	49.86mb	distant
	rs2322212	ILMN_1686920 (CCDC58)	0.0300	120.03mb	distant
	rs7647780	ILMN_1762084 (DNASE1L3)	0.0292	57.51mb	distant
	rs1516347	ILMN_1726020 (LOC652670)	0.0278	75.49mb	distant
	rs13061928	ILMN_1692261 (EPHB1)	0.0273	133.55mb	distant
	rs1377213	ILMN_1698934 (CMTM7)	0.0270	26.76mb	distant
LORS-LORSEN	rs1505587	ILMN_1657373 (LEPREL1)	0.3336	127.69mb	distant
	rs6807033	ILMN_1787750 (CD200)	0.2796	4.163kb	local
	rs11914577	ILMN_1700967 (C3orf59)	0.2245	113.51kb	local
	rs1403719	ILMN_1771599 (PLOD2)	0.1963	25.06mb	distant
	rs628267	ILMN_1760509 (EOMES)	0.1941	302.30kb	
	rs4016435	ILMN_1757350 (CTNNB1)	0.1908	27.772kb	local
	rs16839507	ILMN_1761058 (ACAD11)	0.1856	117.942kb	local
	rs693430	ILMN_1657708 (MGLL)	0.1796	86.074kb	local
	rs693430	ILMN_1707310 (MGLL)	0.1710	47.617kb	local
	rs1498090	ILMN_1793724 (C3orf31)	0.1662	58.605kb	local
LORS-LORS	rs1505587	ILMN_1657373 (LEPREL1)	1.2549	127.69mb	distant
	rs6807033	ILMN_1787750 (CD200)	0.5621	4.163kb	local
	rs4857653	ILMN_1700967 (C3orf59)	0.3640	16.16mb	distant
	rs11914577	ILMN_1700967 (C3orf59)	0.2984	113.514kb	local
	rs1403719	ILMN_1771599 (PLOD2)	0.2824	25.06mb	distant
	rs628267	ILMN_1760509 (EOMES)	0.2439	302.302kb	
	rs4016435	ILMN_1757350 (CTNNB1)	0.2404	27.772kb	local
	rs16839507	ILMN_1761058 (ACAD11)	0.2338	117.942kb	local
	rs3773014	ILMN_1762084 (DNASE1L3)	0.2268	29.187kb	local
	rs1799977	ILMN_1688392 (ZBED2)	0.2234	75.77mb	distant

Table 2.8: Top ten detected SNP-gene pairs for chromosome 15.

Method	**SNP**	**Probe (Gene)**	**Class**
HC-FastLORS	rs12594727	ILMN_1652797 (LOC400451)	distant
	rs17734920	ILMN_1652797 (LOC400451)	distant
	rs12594727	ILMN_1804277 (SPRED1)	distant
	rs4567674	ILMN_1692517 (LOC653381)	distant
	rs12440268	ILMN_1692517 (LOC653381)	distant
	rs1977035	ILMN_1710216 (AVEN)	distant
	rs11633486	ILMN_1690695 (PEX11A)	distant
	rs6606804	ILMN_1665859 (RAB27A)	distant
	rs1977035	ILMN_1693650 (FES)	distant
	rs11634559	ILMN_1748374 (LOC400304)	
LORS-LORSEN	rs6151443	ILMN_1712082 (GCNT3)	local
	rs12441559	ILMN_1712082 (GCNT3)	distant
	rs9635390	ILMN_1656899 (CIB1)	local
	rs16970801	ILMN_1749096 (BCL2L10)	distant
	rs8024414	ILMN_1813430 (TRIM69)	distant
	rs288406	ILMN_1808238 (RBPMS2)	distant
	rs7162538	ILMN_1784364 (STARD5)	local
	rs16957709	ILMN_1792173 (76P)	local
	rs1347069	ILMN_1795822 (DIS3L)	local
	rs3825946	ILMN_1667199 (SQRDL)	local
LORS-LORS	rs6151443	ILMN_1712082 (GCNT3)	local
	rs9635390	ILMN_1656899 (CIB1)	local
	rs7162538	ILMN_1784364 (STARD5)	local
	rs16957709	ILMN_1792173 (76P)	local
	rs12440502	ILMN_1805410 (C15orf48)	distant
	rs2292114	ILMN_1795524 (C15orf44)	local
	rs1347069	ILMN_1795822 (DIS3L)	local
	rs25431	ILMN_1748374 (LOC400304)	distant
	rs3825946	ILMN_1667199 (SQRDL)	local
	rs7177893	ILMN_1689274 (NIPA1)	local

Table 2.9: Top ten detected SNP-gene pairs for chromosome 17.

Method	**SNP**	**Probe (Gene)**	**Class**
HC-FastLORS	rs8082184	ILMN_1747419 (PCGF2)	distant
	rs4790694	ILMN_1773352 (CCL5)	distant
	rs3213714	ILMN_1769550 (SLFN5)	distant
	rs2317668	ILMN_1769550 (SLFN5)	distant
	rs12950579	ILMN_1769550 (SLFN5)	distant
	rs17822338	ILMN_1769550 (SLFN5)	distant
	rs4985676	ILMN_1733811 (JUP)	distant
	rs4968140	ILMN_1706959 (TIMM22)	local
	rs6806	ILMN_1810486 (RAB34)	distant
	rs9915773	ILMN_1707448 (CRKRS)	distant
LORS-LORSEN	rs4794776	ILMN_1808301 (MRPL45)	local
	rs4251704	ILMN_1773352 (CCL5)	local
	rs4789267	ILMN_1782778 (FAM100B)	local
	rs3809767	ILMN_1687247 (SPATA20)	local
	rs17657522	ILMN_1697227 (USP36)	local
	rs4796817	ILMN_1697227 (USP36)	local
	rs12952713	ILMN_1750511 (NT5C3L)	distant
	rs4968140	ILMN_1706959 (TIMM22)	local
	rs6504230	ILMN_1747347 (C17orf60)	local
	rs33926631	ILMN_1738027 (BRCA1)	local
LORS-LORS	rs4794776	ILMN_1808301 (MRPL45)	local
	rs3809767	ILMN_1687247 (SPATA20)	local
	rs17657522	ILMN_1697227 (USP36)	local
	rs4796817	ILMN_1697227 (USP36)	local
	rs9905601	ILMN_1750511 (NT5C3L)	local
	rs11868362	ILMN_1733811 (JUP)	distant
	rs4791136	ILMN_1733811 (JUP)	distant
	rs4968140	ILMN_1706959 (TIMM22)	local
	rs6504230	ILMN_1747347 (C17orf60)	local
	rs33926631	ILMN_1738027 (BRCA1)	local

Table 2.10: Top ten detected SNP-gene pairs for chromosome 20.

Method	SNP	Probe (Gene)	Class
HC-FastLORS	rs692862	ILMN_1713561 (C20orf103)	distant
	rs530652	ILMN_1814247 (TCFL5)	distant
	rs6084912	ILMN_1791771 (HCK)	distant
	rs16991099	ILMN_1758146 (SIRPA)	
	rs6084217	ILMN_1804822 (SRXN1)	
	rs16991131	ILMN_1666269 (CTSZ)	distant
	rs6041750	ILMN_1702237 (FKBP1A)	local
	rs692862	ILMN_1712347 (SFRS6)	distant
	rs6052369	ILMN_1712347 (SFRS6)	distant
	rs1292244	ILMN_1670841 (CPNE1)	distant
LORS-LORSEN	rs760087	ILMN_1814247 (TCFL5)	local
	rs6115906	ILMN_1751330 (RBCK1)	local
	rs4911408	ILMN_1798014 (EIF2S2)	local
	rs16989514	ILMN_1721128 (TOMM34)	local
	rs2223246	ILMN_1666181 (SDC4)	
	rs6041750	ILMN_1702237 (FKBP1A)	local
	rs1410936	ILMN_1712347 (SFRS6)	distant
	rs2223246	ILMN_1712347 (SFRS6)	local
	rs6103330	ILMN_1712347 (SFRS6)	local
	rs13040414	ILMN_1712347 (SFRS6)	
LORS-LORS	rs6109758	ILMN_1713561 (C20orf103)	
	rs6112999	ILMN_1713561 (C20orf103)	distant
	rs6075584	ILMN_1814247 (TCFL5)	distant
	rs760087	ILMN_1814247 (TCFL5)	local
	rs16989514	ILMN_1721128 (TOMM34)	local
	rs6041750	ILMN_1702237 (FKBP1A)	local
	rs209901	ILMN_1811315 (EEF1A2)	distant
	rs1410936	ILMN_1712347 (SFRS6)	distant
	rs6103330	ILMN_1712347 (SFRS6)	local
	rs13040414	ILMN_1712347 (SFRS6)	

Chapter 3

BiNetPeR: A Bipartite Network-Based Penalized Regression Method For eQTL Mapping

Abstract

Identification and characterization of the expression quantitative trait loci (eQTLs), the genetic variations that regulate the expression of genes can greatly help us better understand the cellular mechanisms underlying human complex diseases. With the accumulation and the availability of a large volume of gene expression data and single nucleotide polymorphism (SNP) genotype data, it is important and challenging to develop more powerful statistical methods and more efficient computational tools for the eQTL mapping. In this paper, we propose a new method called BiNetPeR (Bipartite Network-based Penalized Regression) to identify the eQTLs. Most of the existing methods that use the SNP-SNP network generally construct the SNP-SNP network only from the SNP information and/or the SNP genotypes without the consideration of the gene expression data or the relationship between the SNPs and the gene expres-

sion. BiNetPeR utilizes the SNP-SNP network projected from the SNP-gene bipartite network which is constructed based on the significant marginal associations between the SNPs and the gene expression levels. BiNetPeR also uses the Laplacian matrix of SNP-SNP network to control the amount of regularization for smoothness in the penalized regression. We perform the extensive simulation studies to evaluate and compare the performance of our proposed method with two commonly used methods, FastLORS and mtLasso2G. Simulation studies show that our method outperforms FastLORS and mtLasso2G in terms of average Area Under the Curve (AUC) in most situations.

3.1 Introduction

With rapid advancements in high-throughput and sequencing technologies, a large number of single nucleotide polymorphism (SNP) genotype data and gene expression data have become available. It is necessary and important to develop effective statistical methods to investigate the associations between a set of SNPs and expression levels of a set of genes. We usually refer it to as the expression quantitative trait locus (eQTL) mapping. The eQTL mapping aims to identify the genetic variants which have impact on gene expression levels and can reveal the genetic mechanism of gene expression activities to further improve our understanding on how genetic variations are related with human complex diseases. Therefore, the eQTL mapping offers a promise for the understanding of the biological process of gene regulation and interpretation for the findings obtained from genome-wide association studies (GWAS) (Cookson et al., 2009).

A large number of statistical and computational methods have been proposed to detect the eQTLs. The most commonly used methods are based on the regression model in which gene expression levels are treated as the response variables and SNP genotypes are treated as the predictors. The simplest application of the regression model in the eQTL mapping is to use the univariate regression to detect the marginal association for each pair of SNP and gene (Shabalin, 2012). Since the eQTL mapping generally involves a large number of SNPs and a large number of genes, such univariate regression model can avoid the computational challenges from the high dimensionality of SNPs and genes, therefore can be efficiently performed. However, the univariate regression model needs the adjustment for the multiple testing and ignores the correlation structure of SNPs and genes, making it difficult to detect weak association signals between SNPs and genes. Therefore, the methods that can jointly model multiple SNPs and single gene (e.g., LSKM-LASSO (Yan et al., 2020)) or multiple genes have been developed (e.g., PANAMA (Fusi et al., 2012), JDAG (Cao et al., 2020), PEER (Stegle et al., 2010), LORSEN [1], HEFT (Gao et al., 2014), and LMM-EH-PS (Listgarten et al., 2010)). Such methods can take into consideration the SNP linkage disequilibrium (LD) structure between SNPs and the gene-gene correlations, thus are generally more powerful than the univariate regression model to detect eQTLs.

In addition, the joint modeling of multiple SNPs and multiple genes can avoid the adjustment for the multiple testing. However, the multivariate regression usually requires the intensive computation since the penalty terms are generally needed to be imposed to the regression coefficients to handle the large number of coefficients in the model and the low signal-to-noise ratio (SNR).In summary, the univariate analysis and the joint modeling have their advantages and disadvantages, respectively. Some researchers have made efforts to combine them together. Wang et al. (2011) and Yang et al. (2013) proposed to first select significant SNPs from the univariate analysis, then analyze selected SNPs using the joint modeling.

Inspired by the success of data integration, some researchers have developed the

[1] The article has been submitted to *Frontiers in Genetics*, still under review. How to cite this article: Gao, Cheng and Wei, Hairong and Zhang, Kui. LORSEN: fast and efficient eQTL mapping with low rank penalized regression.

methods that can include the external data (e.g. the protein-protein interaction networks, the summary statistics from genetic association studies) as the prior or auxiliary information in the eQTL mapping(e.g., GeP-HMRF (Wang et al., 2018a), ARCHIE (Dutta et al., 2020), GFlasso (Lee and Xing, 2012)). In particular, some researchers developed the methods, such as the two-graph guided multi-task Lasso (mtLasso2G), the graph-regularized dual lasso (GDL), and the graph-guided fused lasso (GFlasso), that can incorporate the correlation structure of SNP data and gene expression data into a penalized regression model to boost the detection power of the eQTL mapping. However, there are several drawbacks for the aforementioned methods. First, the underlying assumption of such methods is that if two SNPs are highly correlated with each other, then they should have similar genetic effects on gene expression levels; if two genes are highly correlated with each other, then SNPs should have similar overall genetic effects on expression levels of these two genes. This assumption may not hold in general. Just because two SNPs are in high LD, it does not mean that those two SNPs have the same pleiotropic effects on the gene expression levels. Similarly, the polygenic effects of SNPs on two highly correlated genes are not necessarily the same. This is mainly because that the gene expression levels are not only regulated by genetic variants but also influenced by environmental factors (e.g. random errors), confounding factors (hidden or known, e.g. batch effects), and covariates (e.g. gender, age). Second, such methods often rely on the clustering method and existing network information to obtain the structure information of SNPs and/or genes. Clustering is an unsupervised learning approach. The number of clusters and the clustering metric are usually artificially determined, which introduces the uncertainty to such methods. Third, existing network information for SNPs and genes are usually incomplete and may not be easily accessible, which restricts the use of such methods. Fourth, such methods obtain the structure information of SNPs or genes separately, which does not consider the relationship between SNPs and genes.

In this paper, we propose a novel method for the eQTL mapping, called **Bi**partite

Network-based **Pe**nalized **R**egression, abbreviated as BiNetPeR. Our method consists of the following steps. First, the significant marginal association of each pair of SNP and gene is obtained and is used to construct a SNP-gene bipartite network. Second, the SNP-SNP and/or gene-gene network is obtained by projecting the SNP-gene bipartite network to SNPs and/or genes. Third, the SNP-SNP and/or gene-gene network is used in the penalized multivariate regression to detect eQTLs. Based on the framework of the Elastic Net penalty (Zou and Hastie, 2005), BiNetPeR can be formulated as a Lasso-type problem, so the model parameters can be efficiently estimated. When no significant marginal association evidence is present, BiNetPeR reduces to LORS (or FastLORS (Jeng et al., 2020)), a commonly used method in eQTL mapping. Compared with the existing methods, BiNetPeR has several advantages. The SNP-SNP and/or gene-gene network is obtained from a SNP-gene bipartite network that is based on the marginal association of SNP genotypes and gene expression levels. Therefore, the prior information of the correlation of SNPs and/or genes is not needed. Second, since the SNP-SNP and/or the gene-gene network is based on the marginal association of SNP genotypes and gene expression levels, we expect that the corrected (or equivalently, correlated) SNPs have the similar pleiotropic effects on the gene expression levels. Third, the existing method mainly use the structure information of SNPs obtained from SNPs data only and/or the gene-gene network obtained from the gene expression levels only, we expect BiNetPeR is more powerful in the eQTL mapping.

To account for the non-genetic effects of potential hidden factors on the gene expression levels, we first apply PEER (Stegle et al., 2010) to predict the hidden factors that affect the gene expression levels, then we extract the predicted factors from the gene expression levels, thereafter we use the gene expression residuals in the subsequent analysis. Compared with existing methods such as LORS (Yang et al., 2013) and LORSEN which consider the non-genetic effects of hidden factors, BiNetPeR is more computationally efficient. This is because PEER is only applied once at the beginning while LORS and LORSEN estimate the non-genetic effects of hidden factors at each iteration, which is computationally intensive especially when a large number of SNPs and genes are used in the eQTL mapping.

We evaluate the performance of BiNetPeR and compare its performance with two commonly used methods in the eQTL mapping, FastLORS and mtLasso2G, through extensive simulation studies as well as the data from the International HapMap Project (Gibbs et al., 2003). FastLORS is a low rank penalized regression model in which the non-genetic effects of hidden factors are modeled as an unknown matrix to be estimated and L_1 [2] penalty is imposed on association coefficients to force a sparse solution. mtLasso2G considers L_1 penalty and fussed Lasso-type penalty on association coefficients, the latter one is based on the basic underlying assumption mentioned above. The design of the latter penalty term is from the structure information of SNP-SNP network and gene-gene network.

3.2 Materials and Methods

We assume that there are n samples involved in the study, and the genotype data for p SNPs and the gene expression levels for q genes are collected from the samples. Let X denote the $n \times p$ matrix of SNP genotypes coded in an additive manner, and Y denote the $n \times q$ matrix of gene expression levels. The first step of our method is to build a bipartite network $G(S_1, S_2, E)$ where the nodes in S_1 are the genetic variants (or SNPs) studied, the nodes in S_2 are the genes and E is the set of edges of which each links one SNP and one gene indicating that the SNP is associated with the expression levels of the gene. Different from the existing methods that use the SNP-SNP networks based on the relationship between the SNPs and/or the gene-gene interaction networks based on the relationship between the genes, BiNetPeR

[2] It means the L_1 norm of the matrix, that is, the sum of the absolute values of the entries of the matrix.

uses a bipartite network based on the marginal association of the SNP-gene pairs. To construct a bipartite network, we first use MatrixEQTL (Shabalin, 2012) to find all the SNP-gene pairs that have the significant marginal association and then draw an edge between the SNP and the gene if that pair of SNP and gene has the significant marginal association. MatrixEQTL is an efficient method for the eQTL mapping, thus allows us to find all the significant SNP-gene pairs from a large number of SNPs and genes efficiently. Moreover, MatrixEQTL is able to incorporate the covariates in the analysis. In the second step, we project the gene nodes onto the SNP nodes, also called the bipartite network projection (Zhou et al., 2007), to obtain the weighted connected components of SNPs. The weight w_{ij} of an edge in the connected component of SNPs is the number of genes that are marginally associated with two SNPs linked by that edge. Two SNPs lie in two different connected components if they are not marginally associated with any gene in common. In the last step, if covariates are present in the study, we first regress covariates out from gene expression data, then we standardize the gene expression residuals (still denoted by Y) and the SNP genotype data (still denoted by X); otherwise, we standardize the gene expression data (still denoted by Y) and the SNP genotype data (still denoted by X). Then we use a penalized regression model to find the associations between the SNPs and the genes. The steps for BiNetPeR are illustrated in Figure 3.1.

For the convenience of description, we first introduce the notations used in the penalized regression model. For a $p \times q$ matrix M with the elements M_{ij} ($i = 1, \cdots, p; j = 1, \cdots, q$), the Frobenius norm of M is defined as $\|M\|_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{q} M_{ij}^2}$, the square root of the sum of squared elements in the matrix; the L_1 norm of M is defined as $\|M\|_1 = \sum_{i=1}^{p} \sum_{j=1}^{q} |M_{ij}|$, the sum of absolute value of each element in the matrix. The penalized regression model is the following (Li and Li, 2010):

$$\underset{B}{min} \quad \frac{1}{2}\|Y - XB\|_F^2 + \lambda\|B\|_1 + \alpha \times pen(G^{SNP}, B), \tag{3.2.1}$$

where $pen(G^{SNP}, B) = \sum_{k=1}^{m} \sum_{i \sim j \in E_k^{SNP}} w_{ij} \sum_{s=1}^{q} (sign(\tilde{\beta}_{is}) \frac{\beta_{is}}{\sqrt{d_i}} - sign(\tilde{\beta}_{js}) \frac{\beta_{js}}{\sqrt{d_j}})^2$ is the quadratic penalty term with respect to the association coefficients, $G_k^{SNP}(V_k^{SNP}, E_k^{SNP})$ is some connected component (also a graph) of SNPs indexed by k as illustrated by "d. SNP Projection Network" in Figure 3.1. V_k^{SNP} is the set of vertices corresponding to the SNPs in the graph G_k^{SNP}. E_k^{SNP} is the set of edges in the graph G_k^{SNP}. m is the total number of connected components of SNPs. w_{ij} is the weight of edge $i \sim j$ in the k-th connected component of SNPs (G_k^{SNP}), actually the number of common genes whose expression levels are affected by the i-th SNP and the j-th SNP. We expect that if the weight of an edge is larger, then two SNPs linked by the edge should have more similarity in affecting gene expression levels. $sign(\tilde{\beta}_{is})$ is the sign of the association coefficient of the i-th SNP for the s-th gene. Similar to that described in (Li and Li, 2010), if the number of SNPs is smaller than the sample size, the signs can be obtained from a standard least square regression; otherwise, we can obtain the signs from the Elastic Net regression. λ and α are two tuning parameters and control the amount of regularization for sparsity and smoothness, respectively. When $\alpha = 0$ or the connected components of SNPs are all single nodes, the penalized regression reduces to the Lasso regression. When the normalized Laplacian matrix is the identity matrix, the penalty term reduces to the Elastic Net penalty, and the penalized regression problem becomes a special case of LORSEN.

To solve the problem 3.2.1, we reformulate it as

$$\min_{B} \quad \frac{1}{2}\sum_{s=1}^{q} \|Y_s - X\beta_s\|_2^2 + \lambda \sum_{s=1}^{q} \|\beta_s\|_1 + \alpha \sum_{s=1}^{q}\sum_{k=1}^{m}\sum_{i\sim j \in E_k^{SNP}} w_{ij}(sign(\tilde{\beta}_{is})\frac{\beta_{is}}{\sqrt{d_i}} - sign(\tilde{\beta}_{js})\frac{\beta_{js}}{\sqrt{d_j}})^2, \tag{3.2.2}$$

equivalently, we have

$$\min_{B} \quad \frac{1}{2}\sum_{s=1}^{q} \|Y_s - X\beta_s\|_2^2 + \lambda \sum_{s=1}^{q} \|\beta_s\|_1 + \alpha \sum_{s=1}^{q} \beta_s^T K_s^T L K_s \beta_s, \tag{3.2.3}$$

where $K_s = diag(sign(\tilde{\beta}_{\ell s}))$, $\ell = 1, 2, \cdots, p$, L is the sum of L_k, $k = 1, \cdots, m$. L_k is the $p \times p$ normalized Laplacian matrix corresponding to the graph G_k^{SNP}, defined as

$$L_k(i,j) = \begin{cases} 1 & \text{if } i = j \text{ and } d_k^i \neq 0, \\ -w_{ij}/\sqrt{d_k^i d_k^j} & \text{if } i \text{ and } j \text{ are adjacent}, \\ 0 & \text{otherwise}, \end{cases}$$

where $d_k^i = \sum_{j \sim i \in E_k^{SNP}} w_{ij}$ is the degree of the vertex i in the graph G_k^{SNP}.

The optimization problem 3.2.3 can be decomposed into q independent subproblems. We can parallelly solve the problems

$$\underset{\beta_s}{min} \quad \frac{1}{2}\|Y_s - X\beta_s\|_2^2 + \lambda\|\beta_s\|_1 + \alpha\beta_s^T K_s^T L K_s \beta_s, \tag{3.2.4}$$

where $s = 1, 2, \cdots, q$. There are two approaches to solve the optimization problem 3.2.4. One is based on the coordinate-descent algorithm (Li and Li, 2010). The other one is to reformulate the problem as a Lasso problem with augmented data matrices (Li and Li, 2008), then the R package Glmnet (Friedman et al., 2010) can be used to solve the problem.

Because gene expression activities are not only regulated by genetic variants, but also influenced by some hidden non-genetic factors. To account for the effects induced by the non-genetic factors , we can firstly apply PEER (Stegle et al., 2010, 2012) to predict the unknown factors, then extract the factors from gene expression. We use the gene expression residuals obtained from PEER to replace the gene expression levels in the subsequent analysis as described above.

3.3 Simulation Design

We perform the extensive simulations to evaluate the performance of BiNetPeR. We first download the genotype data of Chromosome 1 for 165 CEU samples from the third phase of the International HapMap Project (HapMap3) (`https://www.genome.gov/10001688/international-hapmap-project`). The CEU samples refer to Utah residents with Northern and Western European ancestry from the CEPH collection. After the quality-control (window size: 50; moving window increment: five SNPs; cutoff value of R^2: 0.5), the genotype data of 13,815 SNPs for $n = 165$ samples are retained. For our simulations, we randomly choose p SNPs. To simulate the gene expression levels for $q = 200$ genes on $n = 165$ samples, we use the following regression model: $Y = XB + U + e$, where Y is an $n \times q$ matrix that represents the gene expression levels for q genes on n samples, X is an $n \times p$ matrix that represents the SNP genotypes for p SNPs on n samples, B is a $p \times q$ matrix that represents the effects of p SNP genotypes on the gene expression levels of q genes, U is an $n \times q$ matrix that represents the non-genetic effects of k hidden factors in the gene expression levels of q genes, and e is an $n \times q$ matrix that represents the random errors for the gene expression levels. We first simulate non-genetic effects of k hidden factors. We independently generate nk random numbers from the standard normal distribution to form a $n \times k$ matrix H, then independently generate each column of U from the multivariate normal distribution with mean 0 and the covariance matrix τHH^T, where τ is the variance component and τ is set as either 0.1 or 1.2 in our simulations. We then independently generate nq random numbers from the standard normal distribution to form the error matrix e. Because the expression of one gene may be regulated by multiple genetic variants and one genetic variant may affect the expression levels of multiple genes. We consider the following scenarios to simulate B, the matrix for the regression coefficients.

- Scenario 1: To evaluate the false positive rates, we assume that the SNP genotypes do not affect the expression level of any gene considered, that is, $B = 0$ for this simulation scenario. We use $\tau = 0.1$ and $k = 15$ hidden factors for this simulation scenario.

- Scenario 2: To assess the performance of our method when we vary the influence of non-genetic effects of hidden factors on gene expression levels, we set two different values (15 and 30) for k and two different values (0.1 and 1.2) for τ, we consider four different combinations: (A) $k = 15$, $\tau = 0.1$; (B) $k = 30$, $\tau = 0.1$; (C) $k = 15$, $\tau = 1.2$; (D) $k = 30$, $\tau = 1.2$. We expect that our method has robust performance for these four combinations where $p = 300$, the number of causal SNPs is 80. For each causal SNP, we first randomly choose $m = 50$ genes among $q = 200$ gene as the genes whose expression levels are affected by that SNP. We either generate the regression coefficient for a SNP-gene pair from a uniform distribution between 0.5 and 1 if the gene expression is affected by the SNP or set the regression coefficient as 0 otherwise.

- Scenario 3: To assess the performance of our method when we keep a constant ratio of the number of causal SNPs and the total number of SNPs used in the simulations, we consider the following three cases: (1) the number of causal SNPs is 40 and the total number of SNPs is 150; (2) the number of causal SNPs is 80 and the total number of SNPs is 300; and (3) the number of causal SNPs is 160 and the total number of SNPs is 600. Again, for each causal SNP, we first randomly choose $m = 50$ genes among $q = 200$ gene as the genes whose expression levels are affected by that SNP. We either generate the regression coefficient for a SNP-gene pair from a uniform distribution between 0.5 and 1 if the gene expression is affected by the SNP or set the regression coefficient as 0 otherwise. For this simulation scenario, we set $\tau = 0.1$ and $k = 15$.

- Scenario 4: To evaluate the performance of the proposed method with different correlation structures of the causal SNPs, we consider the following simulations. We first randomly choose n_a SNPs as the primary causal SNPs. For each primary causal SNP, we find additional n_b causal SNPs such that, first, the distance between any one of those n_b causal SNPs and the primary causal SNP is less than 100 kb; second, the correlation between any one of those n_b causal SNPs and the primary causal SNP is large. In our simulations, we use $r^2 > 0.7$ as the criteria. Therefore, we first find n_a sets of causal SNPs and the total number of causal SNPs is $n_a(n_b+1)$. We then randomly choose $(300-n_a(n_b+1))$ SNPs that are far away from and nearly independent of causal SNPs to be non-causal SNPs. We use the following combinations of n_a and n_b: $n_a = 40$, $n_b = 1$; $n_a = 20$, $n_b = 3$. For each set of $(n_b + 1)$ SNPs, we generate the regression coefficients in the following way: for primary causal SNP, we randomly choose 50 genes whose expression levels are affected by the primary causal SNP and generate each regression coefficient from a uniform distribution between 0.5 and 1. For each of n_b non-primary causal SNPs, we randomly choose 40 genes among 50 genes affected by the primary causal SNP and 10 genes from the rest of 150 genes. By doing this, we actually conduct our simulations based on the assumption that the highly correlated SNPs may have the similar effects on the gene expression levels. Again, for a SNP-gene pair, we generate the corresponding regression coefficient from a uniform distribution between 0.5 and 1 if the gene expression levels are affected by the SNP. We set the corresponding regression coefficient as zero if the gene expression levels are not affected by the SNP. For this simulation scenario, we set $\tau = 0.1$ and $k = 15$.

For each simulation scenario, we repeat the simulations ten times and use the average Area Under the Curve (AUC) as the criteria to compare different methods.

3.4 Simulation Results

In simulation scenario 1, the SNP genotypes have no effect on the gene expression levels. We expect that the false positive rate (FPR) should be close to zero for each method. Here, we use two thresholds (0 and 0.001) for the absolute value of the estimated regression coefficient to determine if a SNP is significantly associated with a gene. First, we consider a SNP is significantly associated with the expression levels of a gene if the absolute value of the estimated regression coefficient is greater than 0. For BiNetPeR, the average FPR of ten replicates is 4.67×10^{-5}; for mtLasso2G, the average FPR is 1; for FastLORS, the average FPR is 6.81×10^{-3}. Second, we consider a SNP is significantly associated with the expression levels of a gene if the absolute value of the estimated regression coefficient is greater than 0.001. Then we calculate the average FPR for each method. For BiNetPeR, the average FPR of ten replicates is 4.33×10^{-5}; for mtLasso2G, the average FPR is 3.13×10^{-2}; for FastLORS, the average FPR is 2.17×10^{-5}. We conclude that BiNetPeR is a valid method. It can be seen that our method (BiNetPeR) maintains the low FPR in two situations while mtLasso2G has the largest FPR. FastLORS has a much larger FPR than BiNetPeR when 0 is used as the threshold, while it has a similar FPR to BiNetPeR when 0.001 is used as the threshold.

In simulation scenario 2, we consider four different combinations for the values of τ and k to evaluate the influence of non-genetic effects of hidden factors on the detection of eQTLs by BiNetPeR, FastLORS, and mtLasso2G (Figure 3.2). In the simulations, we consider a SNP is significantly associated with the expression levels of a gene if the absolute value of the estimated regression coefficient is greater than 0.001. For $k = 15$ and $\tau = 0.1$, the average AUCs for BiNetPeR, mtLasso2G, and FastLORS are 0.815, 0.619, and 0.734, respectively. For other combinations of τ and k, the AUCs for BiNetPeR, mtLasso2G, and FastLORS are similar to those from $k = 15$ and $\tau = 0.1$. So, the values of τ and k do not have influence on the detection of eQTLs for three methods considered here.

In simulation scenario 3, we keep the ratio of the number of causal SNPs and the total number of SNPs used as a constant. From the simulation results (Figure 3.3), we observe that our method has the best performance. Specifically, when the number of causal SNPs is 40 and the total number of SNPs used in the simulation is 150, the average AUCs for BiNetPeR, mtLasso2G, and FastLORS are 0.850, 0.827, and 0.587, respectively. When the number of causal SNPs is 80 and the total number of SNPs used in the simulation is 300, the average AUCs for BiNetPeR, mtLasso2G, and FastLORS are 0.815, 0.619, and 0.734, respectively. When the number of causal SNPs is 160 and the total number of SNPs used in the simulation is 600, the average AUCs for BiNetPeR, mtLasso2G, and FastLORS are 0.710, 0.635, and 0.711, respectively. We can see that BiNetPeR outperforms mtLasso2G and FastLORS in terms of average AUC in all the three cases. Of course, as expected, as the number of non-causal SNPs becomes larger, the average AUC of BiNetPeR becomes smaller.

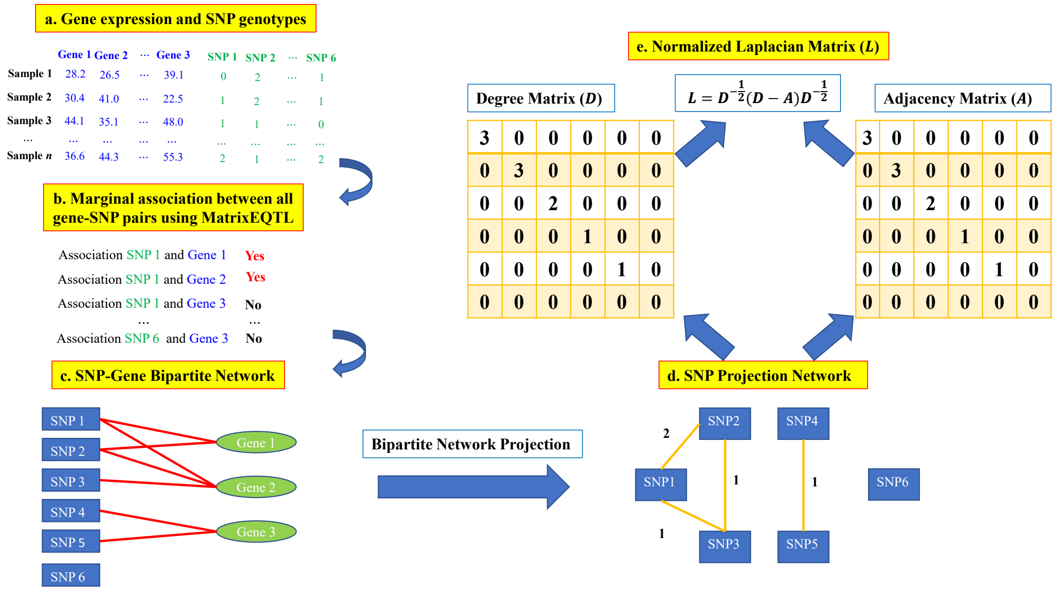
In simulation scenario 4, we explore the influence of the relationship of causal SNPs on the performance of three methods considered here (Figure 3.4). When we randomly choose 300 SNPs in the analysis and randomly choose 80 SNPs out of 300 SNPs to be causal, it is expected that the correlation between the causal SNPs is not strong and the causal SNPs independently influence the expression levels of genes. In this case, BiNetPeR has the best performance with an average AUC of 0.815 while the AUCs of mtLasso2G and FastLORS are 0.619 and 0.734, respectively (Figure 3.4). When the number of primary SNPs is 40 and the number of non-primary causal SNPs is one, the average AUCs for BiNetPeR, mtLasso2G, and FastLORS are 0.751, 0.682, and 0.817, respectively. When the number of primary causal SNPs is 20 and the number of non-primary causal SNPs is three, the average AUCs for BiNetPeR, mtLasso2G, and FastLORS are 0.690, 0.730, and 0.811, respectively. We can see that when tightly linked causal SNPs have the similar effects on the expression levels of genes, BiNetPeR does not perform well: the average AUC of BiNetPeR is lower than the AUC of FastLORS in two situations and the average AUC of BiNetPeR decreases from 0.751 to 0.690 when the ratio of the number of non-primary causal SNPs and the number of primary causal SNPs increases from 1 to 3. In contrast, the average AUC of FastLORS is the highest and does not change much with the ratio of the number of non-primary causal SNPs and the number of primary causal SNPs.

3.5 Discussion

As a large volume of human gene expression data and SNP genotype data becomes available, it is desirable to develop more powerful and efficient statistical methods and computational tools to fully take advantage of such data to investigate the relationship between genetic variants and gene expression levels to help us further understand the genetic mechanism underlying human complex diseases. A number of methods that can jointly model the multiple genetic variants and the expression levels of multiple genes and can incorporate the correlation between SNPs and/or genes have been developed. In this paper, we propose a novel bipartite network-based penalized regression method (BiNetPeR) to detect eQTLs. Our method constructs and uses the SNP-SNP network and/or the gene-gene network projected from the SNP-gene bipartite network based on the significant marginal associations of each SNP-gene pair. Thus, our method actually incorporates the marginal association of SNPs and gene expression levels and does not require the prior information of the SNP-SNP network and/or the gene-gene network. We conduct comprehensive simulations to evaluate and compare the performance of BiNetPeR with two commonly used methods, mt-Lasso2G and FastLORS, in the eQTL mapping. Several conclusions are emerged from our simulation studies. First, BiNetPeR has the appropriate false positive rate, thus is a valid method for the eQTL mapping. Second, BiNetPeR is robust to the non-genetic effects of hidden-factors on the gene expression levels and has the highest average AUC in four cases under consideration in simulation scenario 2. Third, BiNetPeR has the highest average AUC in three cases under consideration in simulation scenario 3, when the ratio of the number of causal SNPs and the total number of SNPs is constant. In summary, BiNetPeR is a valid and more powerful method to detect eQTLs.

There are several limitations for BiNetPeR. In one of simulation scenarios, we consider that some SNPs that are in high linkage disequilibrium have the similar effects on the gene expression levels. FastLORS has the better performance than BiNetPeR in such situation. We are planning to conduct more extensive simulations to find why BiNetPeR does not perform as well as FastLORS. In addition, in our current model, we only use the SNP-SNP network projected from the SNP-gene bipartite network. We are in the process to explore how to also incorporate the gene-gene network projected from the SNP-gene bipartite network to further improve the power for the eQTL mapping.

3.6 Tables and Figures


a. Gene expression and SNP genotypes
b. Marginal association between all gene-SNP pairs using MatrixEQTL
c. SNP-Gene Bipartite Network
SNP 1
SNP 2
SNP 3
SNP 4
SNP 5
SNP 6
Gene 1
Gene 2
Gene 3
Bipartite Network Projection
d. SNP Projection Network
SNP2
SNP4
2
SNP1
1
1
SNP6
1
SNP3
SNP5
e. Normalized Laplacian Matrix (L)
Degree Matrix (D)
Adjacency Matrix (A)

	Gene 1	Gene 2	···	Gene 3	SNP 1	SNP 2	···	SNP 6
Sample 1	28.2	26.5	···	39.1	0	2	···	1
Sample 2	30.4	41.0	···	22.5	1	2	···	1
Sample 3	44.1	35.1	···	48.0	1	1	···	0
···	···	···	···	···	···	···	···	···
Sample *n*	36.6	44.3	···	55.3	2	1	···	2

Association SNP 1 and Gene 1	Yes
Association SNP 1 and Gene 2	Yes
Association SNP 1 and Gene 3	No
···	···
Association SNP 6 and Gene 3	No

$$L = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}$$

Degree Matrix (D)

3	0	0	0	0	0
0	3	0	0	0	0
0	0	2	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	0

Adjacency Matrix (A)

3	0	0	0	0	0
0	3	0	0	0	0
0	0	2	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	0

Figure 3.1: Schematic of BiNetPeR: a. SNP genotypes and gene expression levels are obtained; b. MatrixEQTL is used to identify the significant marginal associations between SNPs and genes; c. A SNP-gene bipartite network is constructed based on the marginal associations; d. A weighted SNP-SNP network is obtained by projecting the SNP-gene bipartite network onto the SNPs; e. The degree matrix and adjacency matrix are obtained and used in the regularized regression model.

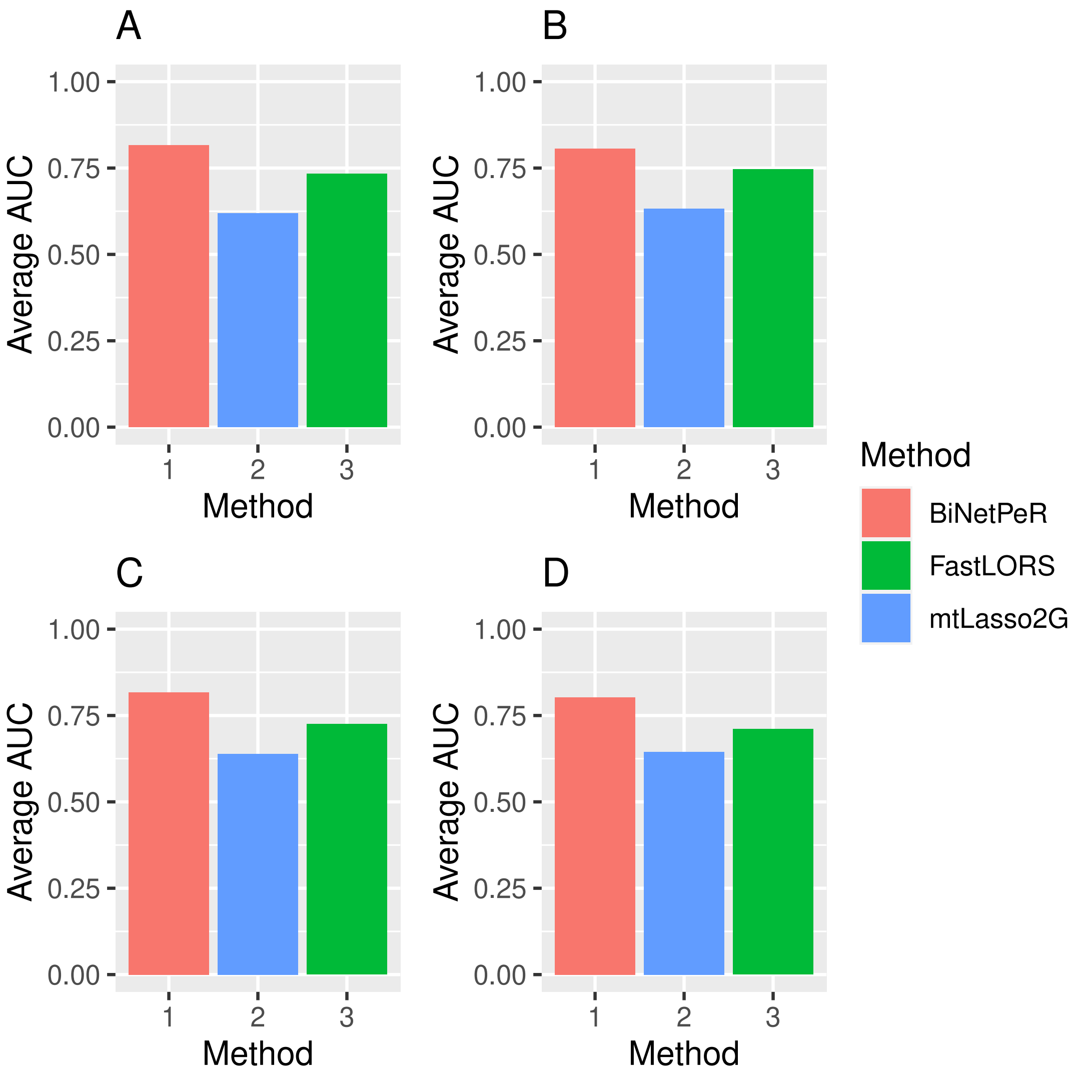


Figure 3.2: Average AUC of three methods in simulation scenario 2. (A): $k = 15$, $\tau = 0.1$; (B): $k = 30$, $\tau = 0.1$; (C): $k = 15$, $\tau = 1.2$; (D): $k = 30$, $\tau = 1.2$.

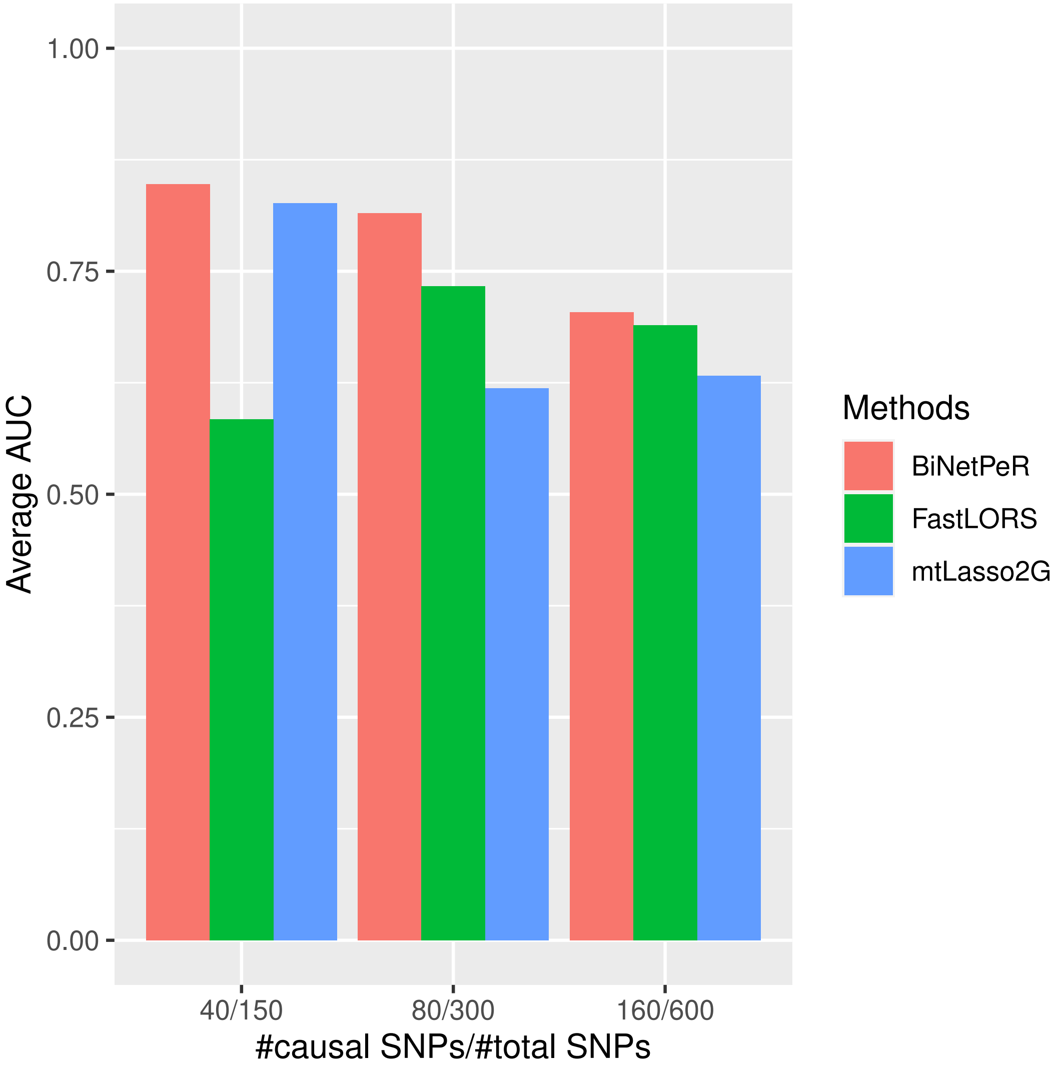


Figure 3.3: Average AUC of three methods in simulation scenario 3.

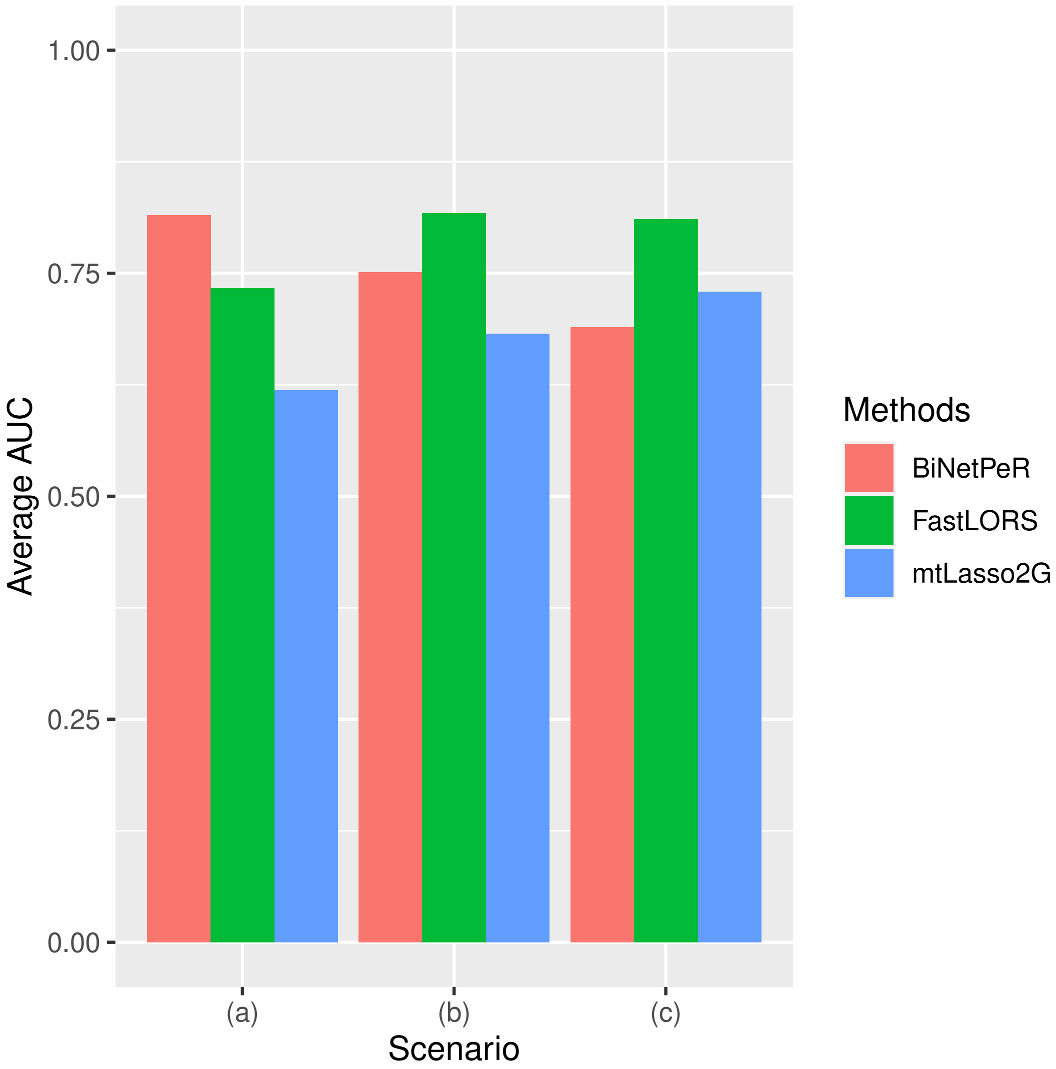


Figure 3.4: Average AUC of three methods in simulation scenario 4. (a) corresponds to case (A) in simulation scenario 2; (b) corresponds to the case in which $n_a = 40$, $n_b = 1$; (c) corresponds to the case in which $n_a = 20$, $n_b = 3$.

References

Albert, F. W. and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat Rev Genet*, 16(4):197–212.

Aschard, H., Vilhjalmsson, B. J., Greliche, N., Morange, P. E., Tregouet, D. A., and Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet*, 94(5):662–76.

Banerjee, S., Simonetti, F. L., Detrois, K. E., Kaphle, A., Mitra, R., Nagial, R., and Söding, J. (2020). Reverse regression increases power for detecting trans-eqtls. *bioRxiv*, page 2020.05.07.083386.

Bays, H. E., Chapman, R. H., and Grandy, S. (2007). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. *Int J Clin Pract*, 61(5):737–47.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., Bielak, L. F., Zhao, W., Smith, J. A., Peyser, P. A., Kardia, S. L. R., Ghosh, D., and Epstein, M. P. (2016). A statistical approach for testing cross-phenotype effects of rare variants. *Am J Hum Genet*, 98(3):525–540.

Bůžková, P. (2013). Linear regression in genetic association studies. *PLoS One*, 8(2):e56976.

Cai, J., Candès, E., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20:1956–1982.

Cao, X., Ding, L., and Mersha, T. B. (2020). Joint variable selection and network modeling for detecting eqtls. *Statistical applications in genetics and molecular biology*, 19(1).

Chatla, S. B. and Shmueli, G. (2017). An extensive examination of regression models with a binary outcome variable. *Journal of the Association for Information Systems*, 18.

Chen, H., Meigs, J. B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol*, 37(2):196–204.

Chen, W. M., Manichaikul, A., and Rich, S. S. (2009). A generalized family-based association test for dichotomous traits. *Am J Hum Genet*, 85(3):364–76.

Chen, X., Shi, X., Xu, X., Wang, Z., Mills, R., Lee, C., and Xu, J. (2012). A two-graph guided multi-task lasso approach for eqtl mapping. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 208–217, La Palma, Canary Islands. PMLR.

Cheng, W., Zhang, X., Guo, Z., Shi, Y., and Wang, W. (2014). Graph-regularized dual lasso for robust eqtl mapping. *Bioinformatics*, 30(12):i139–48.

Chun, H. and Keles, S. (2009). Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, 182(1):79–90.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat Rev Genet*, 10(3):184–94.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994.

Dutta, D., He, Y., Saha, A., Arvanitis, M., Battle, A., and Chatterjee, N. (2020). Novel aggregative trans-eqtl association analysis of known genetic variants detect trait-specific target gene-sets. *medRxiv*.

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

Feng, T., Elston, R. C., and Zhu, X. (2011). Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (spwss, orwss). *Genet Epidemiol*, 35(5):398–409.

Fischer, S. T., Jiang, Y., Broadaway, K. A., Conneely, K. N., and Epstein, M. P. (2018). Powerful and robust cross-phenotype association test for case-parent trios. *Genet Epidemiol*, 42(5):447–458.

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., Garcia Giron, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martinez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigo, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2019). Gencode reference annotation for the human and mouse genomes. *Nucleic Acids Res*, 47(D1):D766–d773.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22.

Fusi, N., Stegle, O., and Lawrence, N. D. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput Biol*, 8(1):e1002330.

Gao, C., Tignor, N. L., Salit, J., Strulovici-Barel, Y., Hackett, N. R., Crystal, R. G., and Mezey, J. G. (2014). Heft: eqtl analysis of many thousands of expressed genes while simultaneously controlling for hidden factors. *Bioinformatics*, 30(3):369–76.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P. K.-H., Tsui, L.-C., Waye, M. M. Y., Wong, J. T.-F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., Murray, S. S., Oliphant, A. R., Montpetit, A., Hudson, T. J., Chagnon, F., Ferretti, V., Leboeuf, M., Phillips, M. S., Verner, A., Kwok, P.-Y., Duan, S., Lind, D. L., Miller, R. D., Rice, J. P., Saccone, N. L., Taillon-Miller, P., Xiao, M., Nakamura, Y., Sekine, A., Sorimachi, K., Tanaka, T., Tanaka, Y., Tsunoda, T., Yoshino, E., Bentley, D. R., Deloukas, P., Hunt, S., Powell, D., Altshuler, D., Gabriel, S. B., Zhang, H., Zeng, C., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Aniagwu, T., Marshall, P. A., Matthew, O., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Stein, L. D., Cunningham, F., Kanani, A., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Donnelly, P., Marchini, J., McVean, G. A. T., Myers, S. R., Cardon, L. R., Abecasis, G. R., Morris, A., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Altshuler, D., Daly, M. J., Schaffner, S. F., Qiu, R., Kent, A., Dunston, G. M., Kato, K., Niikawa, N., Knoppers, B. M., Foster, M. W., Clayton, E. W., Wang, V. O., Watkin, J., Gibbs, R. A., Belmont, J. W., Sodergren, E., Weinstock, G. M., et al. (2003). The international hapmap project. *Nature*, 426(6968):789–796.

Gomila, R. (2019). Logistic or linear? estimating causal effects of treatments on binary outcomes using regression analysis.

Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*, 70(1):42–54.

He, Q., Avery, C. L., and Lin, D. Y. (2013). A general framework for association tests with multivariate traits in large-scale genomics studies. *Genet Epidemiol*, 37(8):759–67.

Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality Quantity*, 43(1):59–74.

Hu, Y. J., Sun, W., Tzeng, J. Y., and Perou, C. M. (2015). Proper use of allele-specific expression improves statistical power for cis-eqtl mapping with rna-seq data. *J Am Stat Assoc*, 110(511):962–974.

Huyghe, J. R., Bien, S. A., Harrison, T. A., Kang, H. M., Chen, S., Schmit, S. L., Conti, D. V., Qu, C., Jeon, J., Edlund, C. K., Greenside, P., Wainberg, M., Schumacher, F. R., Smith, J. D., Levine, D. M., Nelson, S. C., Sinnott-Armstrong, N. A., Albanes, D., Alonso, M. H., Anderson, K., Arnau-Collell, C., Arndt, V., Bamia, C., Banbury, B. L., Baron, J. A., Berndt, S. I., Bezieau, S., Bishop, D. T., Boehm, J., Boeing, H., Brenner, H., Brezina, S., Buch, S., Buchanan, D. D., Burnett-Hartman, A., Butterbach, K., Caan, B. J., Campbell, P. T., Carlson, C. S., Castellvi-Bel, S., Chan, A. T., Chang-Claude, J., Chanock, S. J., Chirlaque, M. D., Cho, S. H., Connolly, C. M., Cross, A. J., Cuk, K., Curtis, K. R., de la Chapelle, A., Doheny, K. F., Duggan, D., Easton, D. F., Elias, S. G., Elliott, F., English, D. R., Feskens, E. J. M., Figueiredo, J. C., Fischer, R., FitzGerald, L. M., Forman, D., Gala, M., Gallinger, S., Gauderman, W. J., Giles, G. G., Gillanders, E., Gong, J., Goodman,

P. J., Grady, W. M., Grove, J. S., Gsur, A., Gunter, M. J., Haile, R. W., Hampe, J., Hampel, H., Harlid, S., Hayes, R. B., Hofer, P., Hoffmeister, M., Hopper, J. L., Hsu, W. L., Huang, W. Y., Hudson, T. J., Hunter, D. J., Ibanez-Sanz, G., Idos, G. E., Ingersoll, R., Jackson, R. D., Jacobs, E. J., Jenkins, M. A., Joshi, A. D., Joshu, C. E., Keku, T. O., Key, T. J., Kim, H. R., Kobayashi, E., Kolonel, L. N., Kooperberg, C., Kuhn, T., Kury, S., et al. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet*, 51(1):76–87.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet*, 92(6):841–53.

Jeng, X. J., Rhyne, J., Zhang, T., and Tzeng, J.-Y. (2020). Effective snp ranking improves the performance of eqtl mapping. *Genetic Epidemiology*, 44(6):611–619.

Jiang, D. and McPeek, M. S. (2014). Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet Epidemiol*, 38(1):10–20.

Jiang, Y., Conneely, K. N., and Epstein, M. P. (2014). Flexible and robust methods for rare-variant testing of quantitative traits in trios and nuclear families. *Genet Epidemiol*, 38(6):542–51.

Kendziorski, C. M., Chen, M., Yuan, M., Lan, H., and Attie, A. D. (2006). Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics*, 62(1):19–27.

Kichaev, G., Bhatia, G., Loh, P. R., Gazal, S., Burch, K., Freund, M. K., Schoech, A., Pasaniuc, B., and Price, A. L. (2019). Leveraging polygenic functional enrichment to improve gwas power. *Am J Hum Genet*, 104(1):65–75.

Kim, J., Zhang, Y., and Pan, W. (2016). Powerful and adaptive testing for multi-trait and multi-snp associations with gwas and sequencing data. *Genetics*, 203(2):715–31.

Kim, S. and Xing, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet*, 5(8):e1000587.

Kim, S. and Xing, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *Ann. Appl. Stat.*, 6(3):1095–1117.

Lasky-Su, J., Murphy, A., McQueen, M. B., Weiss, S., and Lange, C. (2010). An omnibus test for family-based association studies with multiple snps and multiple phenotypes. *Eur J Hum Genet*, 18(6):720–5.

Lauc, G., Huffman, J. E., Pucic, M., Zgaga, L., Adamczyk, B., Muzinic, A., Novokmet, M., Polasek, O., Gornik, O., Kristic, J., Keser, T., Vitart, V., Scheijen, B., Uh, H. W., Molokhia, M., Patrick, A. L., McKeigue, P., Kolcic, I., Lukic, I. K., Swann, O., van Leeuwen, F. N., Ruhaak, L. R., Houwing-Duistermaat, J. J., Slagboom, P. E., Beekman, M., de Craen, A. J., Deelder, A. M., Zeng, Q., Wang, W., Hastie, N. D., Gyllensten, U., Wilson, J. F., Wuhrer, M., Wright, A. F., Rudd, P. M., Hayward, C., Aulchenko, Y., Campbell, H., and Rudan, I. (2013). Loci associated with n-glycosylation of human immunoglobulin g show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet*, 9(1):e1003225.

Lee, M. K., Shaffer, J. R., Leslie, E. J., Orlova, E., Carlson, J. C., Feingold, E., Marazita, M. L., and Weinberg, S. M. (2017a). Genome-wide association study of facial morphology reveals novel associations with frem1 and park2. *PLoS One*, 12(4):e0176566.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*, 95(1):5–23.

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

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–75.

Lee, S. and Xing, E. P. (2012). Leveraging input and output structures for joint mapping of epistatic and marginal eqtls. *Bioinformatics*, 28(12):i137–46.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83(3):311–21.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4(3):1498.

Listgarten, J., Kadie, C., Schadt, E. E., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470.

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

Maity, A., Sullivan, P. F., and Tzeng, J. Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet Epidemiol*, 36(7):686–95.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322.

Mueller, P. W., Rogus, J. J., Cleary, P. A., Zhao, Y., Smiles, A. M., Steffes, M. W., Bucksa, J., Gibson, T. B., Cordovado, S. K., Krolewski, A. S., Nierras, C. R., and Warram, J. H. (2006). Genetics of kidneys in diabetes (gokind) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. *J Am Soc Nephrol*, 17(7):1782–90.

Mägi, R., Suleimanov, Y. V., Clarke, G. M., Kaakinen, M., Fischer, K., Prokopenko, I., and Morris, A. P. (2017). Scopa and meta-scopa: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. *BMC Bioinformatics*, 18(1):25.

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M. R., and Coin, L. J. (2012). Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PLoS One*, 7(5):e34861.

Pan, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genet Epidemiol*, 33(6):497–507.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239.

Pezzolesi, M. G., Poznik, G. D., Mychaleckyj, J. C., Paterson, A. D., Barati, M. T., Klein, J. B., Ng, D. P., Placha, G., Canani, L. H., Bochenski, J., Waggott, D., Merchant, M. L., Krolewski, B., Mirea, L., Wanic, K., Katavetin, P., Kure, M., Wolkow, P., Dunn, J. S., Smiles, A., Walker, W. H., Boright, A. P., Bull, S. B., Doria, A., Rogus, J. J., Rich, S. S., Warram, J. H., and Krolewski, A. S. (2009). Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. *Diabetes*, 58(6):1403–10.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). Plink:

a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75.

Rakitsch, B. and Stegle, O. (2016). Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biol*, 17:33.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, 15(11):1576–83.

Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L., Peyser, P. A., and Lin, X. (2012). Snp set association analysis for familial data. *Genet Epidemiol*, 36(8):797–810.

Service, S. K., Verweij, K. J., Lahti, J., Congdon, E., Ekelund, J., Hintsanen, M., Räikkönen, K., Lehtimäki, T., Kähönen, M., Widen, E., Taanila, A., Veijola, J., Heath, A. C., Madden, P. A., Montgomery, G. W., Sabatti, C., Järvelin, M. R., Palotie, A., Raitakari, O., Viikari, J., Martin, N. G., Eriksson, J. G., Keltikangas-Järvinen, L., Wray, N. R., and Freimer, N. B. (2012). A genome-wide meta-analysis of association studies of cloninger's temperament scales. *Transl Psychiatry*, 2(5):e116.

Shabalin, A. A. (2012). Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–8.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, 6(5):e1000770.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

van der Sluis, S., Posthuma, D., and Dolan, C. V. (2013). Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*, 9(1):e1003235.

Wang, J., Zheng, J., Wang, Z., Li, H., and Deng, M. (2018a). Inferring gene-disease association by an integrative analysis of eqtl genome-wide association study and protein-protein interaction data. *Human heredity*, 83(3):117–129.

Wang, J. G., Staessen, J. A., Franklin, S. S., Fagard, R., and Gueyffier, F. (2005). Systolic and diastolic blood pressure lowering as determinants of cardiovascular outcome. *Hypertension*, 45(5):907–13.

Wang, L., Lee, S., Gim, J., Qiao, D., Cho, M., Elston, R. C., Silverman, E. K., and Won, S. (2016). Family-based rare variant association analysis: A fast and efficient method of multivariate phenotype association analysis. *Genet Epidemiol*, 40(6):502–11.

Wang, P., Dawson, J. A., Keller, M. P., Yandell, B. S., Thornberry, N. A., Zhang, B. B., Wang, I. M., Schadt, E. E., Attie, A. D., and Kendziorski, C. (2011). A model selection approach for expression quantitative trait loci (eqtl) mapping. *Genetics*, 187(2):611–21.

Wang, Z., Sha, Q., Fang, S., Zhang, K., and Zhang, S. (2018b). Testing an optimally weighted combination of common and/or rare variants with multiple traits. *PLoS One*, 13(7):e0201186.

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., t Hoen, P. A. C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez, D. G., Nalls, M. A., Homuth, G., Nauck, M., Radke, D., Völker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S. A., Enquobahrie, D. A., Lumley, T., Montgomery, G. W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R. C., Visscher, P. M., Knight, J. C., Psaty, B. M., Ripatti, S., Teumer, A., Frayling, T. M., Metspalu, A., van Meurs, J. B. J., and Franke, L. (2013). Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243.

Won, S., Kim, W., Lee, S., Lee, Y., Sung, J., and Park, T. (2015). Family-based association analysis: a fast and efficient method of multivariate association analysis with multiple variants. *BMC Bioinformatics*, 16:46.

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

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *Am J Hum Genet*, 86(6):929–42.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 89(1):82–93.

Yan, K. K., Zhao, H., Wu, J. T., and Pang, H. (2020). An enhanced machine learning tool for cis-eqtl mapping with regularization and confounder adjustments. *Genetic Epidemiology*, 44(8):798–810.

Yan, Q., Weeks, D. E., Celedon, J. C., Tiwari, H. K., Li, B., Wang, X., Lin, W. Y., Lou, X. Y., Gao, G., Chen, W., and Liu, N. (2015). Associating multivariate quantitative phenotypes with genetic variants in family samples with a novel kernel machine regression method. *Genetics*, 201(4):1329–39.

Yang, C., Wang, L., Zhang, S., and Zhao, H. (2013). Accounting for non-genetic factors by low-rank representation and sparse regression for eqtl mapping. *Bioinformatics*, 29(8):1026–34.

Yu, Y.-L. (2013). On decomposing the proximal map. *Advances in neural information processing systems*, 26:91–99.

Yue, F., Zhu, H., Song, Y., Peng, Q., and Zhou, X. (2020). Efficient and effective control of confounding in eqtl mapping studies through joint differential expression and mendelian randomization analyses. *Bioinformatics*.

Zhang, L. and Sun, L. (2018). On 'reverse' regression for robust genetic association studies and allele frequency estimation with related individuals. *bioRxiv*, page 470328.

Zhou, T., Ren, J., Medo, M., and Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical review E*, 76(4):046115.

Zhu, Y. and Xiong, M. (2012). Family-based association studies for next-generation sequencing. *Am J Hum Genet*, 90(6):1028–45.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Appendix A

Details & Proofs

A.1 Theoretical Details In Chapter 1

A.1.1 Lemmas

Lemma A.1.1. *Assume A is an $n \times n$ symmetric matrix, x an $n \times 1$ vector, B an $n \times n$ nonsingular symmetric matrix. Then,*

$$\max_{x \neq 0} \frac{x^T A x}{x^T B x} = \max_{x^\star \neq 0} \frac{x^{\star T} Q x^\star}{x^{\star T} x^\star} = \max_{\|x^\star\|_2 = 1} x^{\star T} Q x^\star = \lambda_{max}(Q), \tag{A.1.1}$$

where $x^\star = B^{\frac{1}{2}} x$ and $Q = (B^{-\frac{1}{2}})^T A (B^{-\frac{1}{2}})$.

Proof. From linear algebra, we know there exists an orthogonal matrix P such that $Q = P^T \Lambda P$, Λ is a diagonal matrix. Then

$$x^{\star T} Q x^\star = x^{\star T} P^T \Lambda P x^\star = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2 \leq \lambda_{max}(Q) \left\| y \right\|_2^2, \tag{A.1.2}$$

where $y = Px^\star$, $\|y\| = \|Px^\star\| = 1$. Without loss of generality, assume $\lambda_{max}(Q) = \lambda_1$, then equality holds if and only if $y_1 = 1, y_i = 0$ for $i \geq 2$. ■

Lemma A.1.2. *Assume A is an $m \times n$ matrix, and B is an $n \times m$ matrix. Then,*

$$\lambda_{max}(AB) = \lambda_{max}(BA). \tag{A.1.3}$$

Proof. Without loss of generality, assume $\lambda_1 = \lambda_{max}(AB)$, $\alpha_1 = \lambda_{max}(BA)$. Then we have $(AB)v = \lambda_1 v$, for some eigenvector v corresponding to eigenvalue λ_1, and $(BA)u = \alpha_1 u$, for some eigenvector u corresponding to eigenvalue α_1.

Furthermore, we have

$$B(AB)v = (BA)(Bv) = \lambda_1(Bv), \quad A(BA)u = (AB)(Au) = \alpha_1(Au). \tag{A.1.4}$$

So λ_1 is also an eigenvalue of BA, $\lambda_1 \leq \alpha_1$. Similarly, α_1 is also an eigenvalue of AB, $\alpha_1 \leq \lambda_1$. Therefore, $\lambda_1 = \alpha_1$. ■

A.1.2 Derivation Of Test Statistics

Under the linear model

$$x_i = \beta^T y_i + (\gamma^T z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2\Phi). \tag{A.1.5}$$

In the above formula: $Y = (y_1^T, y_2^T, ..., y_n^T)^T$: $n \times K$ matrix of phenotypes
$X_i = (x_{i1}, x_{i2}, ..., x_{iM})^T, i = 1, 2, ..., n$: genotypes at M markers of the i-th individual
$X = (X_1^T, X_2^T, ..., X_n^T)^T$: $n \times M$ matrix of genotypes of all n individuals
$x_i = w^T X_i$, $x = (x_1, x_2, ..., x_n)^T$: weighted combination of genotypes of all n individuals
$Z = (z_1^T, z_2^T, ..., z_n^T)^T$: $n \times L$ matrix of covariates, L is the number of covariates
Φ: $n \times n$ kinship matrix of all n individuals

A.1.2.1 Without Covariates

$$l = exp\{-\frac{1}{2\sigma^2}(x - Y\beta)^T\Phi^{-1}(x - Y\beta)\}(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}|\Phi|^{-\frac{1}{2}}, \tag{A.1.6}$$

$$logl = -\frac{1}{2\sigma^2}(x - Y\beta)^T\Phi^{-1}(x - Y\beta) - \frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{1}{2}log|\Phi|, \tag{A.1.7}$$

$$\frac{\partial logl}{\partial \beta} = \frac{1}{\sigma^2} Y^T \Phi^{-1}(x - Y\beta), \tag{A.1.8}$$

$$\frac{\partial^2 logl}{\partial \beta^2} = -\frac{1}{\sigma^2} Y^T \Phi^{-1} Y. \tag{A.1.9}$$

So, under the null hypothesis, the score test statistic is

$$T_{score} = \frac{w^T X^T \Phi^{-1} Y (Y^T \Phi^{-1} Y)^{-1} Y^T \Phi^{-1} X w}{\hat{\sigma}^2}. \tag{A.1.10}$$

Notice that MLE of σ^2 is $\hat{\sigma}^2 = \frac{1}{n} x^T \Phi^{-1} x$. We use D to represent $\frac{1}{n} X^T \Phi^{-1} X$ in the score test statistic. So, we have

$$T^0_{score} = \frac{w^T X^T \Phi^{-1} Y (Y^T \Phi^{-1} Y)^{-1} Y^T \Phi^{-1} X w}{w^T D w}. \tag{A.1.11}$$

Using lemmas A.1.1 and A.1.2, the final test statistic is

$$T_{MF-TOWmuT} = \max_w T^0_{score} = \lambda_{max}((Y^T \Phi^{-1} Y)^{-1} Y^T \Phi^{-1} B \Phi^{-1} Y), \quad B = X D^{-1} X^T. \tag{A.1.12}$$

A.1.2.2 With Covariates

$$l = exp\{-\frac{1}{2\sigma^2}(x - Y\beta - Z\alpha)^T \Phi^{-1}(x - Y\beta - Z\alpha)\}(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}|\Phi|^{-\frac{1}{2}}, \tag{A.1.13}$$

$$logl = -\frac{1}{2\sigma^2}(x - Y\beta - Z\alpha)^T \Phi^{-1}(x - Y\beta - Z\alpha) - \frac{n}{2} log 2\pi - \frac{n}{2} log \sigma^2 - \frac{1}{2} log|\Phi|, \tag{A.1.14}$$

$$\frac{\partial logl}{\partial \beta} = \frac{1}{\sigma^2} Y^T \Phi^{-1}(x - Y\beta - Z\alpha), \tag{A.1.15}$$

$$\frac{\partial logl}{\partial \alpha} = \frac{1}{\sigma^2} Z^T \Phi^{-1}(x - Y\beta - Z\alpha). \tag{A.1.16}$$

Under the null hypothesis, MLE of α is $\hat{\alpha} = (Z^T \Phi^{-1} Z)^{-1} Z^T \Phi^{-1} x$, and MLE of σ^2 is $\hat{\sigma}^2 = \frac{1}{n} x^T \Phi^{-1}(I - P)x$, $P = Z(Z^T \Phi^{-1} Z)^{-1} Z^T \Phi^{-1}$. Notice that $(I - P)^2 = I - P$, $(\Phi^{-1}(I - P))^T = (I - P)^T \Phi^{-1} = \Phi^{-1}(I - P)$.

$$\left.\begin{pmatrix} \frac{\partial logl}{\partial \alpha} \\ \frac{\partial logl}{\partial \beta} \end{pmatrix}\right|_{\beta=0, \alpha=\hat{\alpha}, \sigma^2=\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2} \begin{pmatrix} Z^T \\ Y^T \end{pmatrix} \Phi^{-1}(I - P)x, \tag{A.1.17}$$

$$\frac{\partial^2 logl}{\partial\alpha\partial\alpha^T} = -\frac{1}{\hat{\sigma}^2} Z^T\Phi^{-1}Z \qquad \frac{\partial^2 logl}{\partial\beta\partial\beta^T} = -\frac{1}{\hat{\sigma}^2} Y^T\Phi^{-1}Y \tag{A.1.18}$$

$$\frac{\partial^2 logl}{\partial\alpha\partial\beta^T} = -\frac{1}{\hat{\sigma}^2} Z^T\Phi^{-1}Y \qquad \frac{\partial^2 logl}{\partial\beta\partial\alpha^T} = -\frac{1}{\hat{\sigma}^2} Y^T\Phi^{-1}Z, \tag{A.1.19}$$

so Fisher information matrix is

$$I = \frac{1}{\hat{\sigma}^2}\begin{pmatrix} Z^T\Phi^{-1}Z & Z^T\Phi^{-1}Y \\ Y^T\Phi^{-1}Z & Y^T\Phi^{-1}Y \end{pmatrix}. \tag{A.1.20}$$

So, under null hypothesis, score test statistic is

$$\tilde{T}_{score} = \frac{w^TX^TA^TY(Y^TAY)^{-1}Y^TAXw}{\hat{\sigma}^2}, \quad A = \Phi^{-1}(I-P). \tag{A.1.21}$$

Let $\tilde{X} = (I-P)X$,$\tilde{Y} = (I-P)Y$, $\tilde{D} = \frac{1}{n}\tilde{X}^T\Phi^{-1}\tilde{X}$. So, we have

$$\tilde{T}^0_{score} = \frac{w^TX^TA^TY(Y^TAY)^{-1}Y^TAXw}{w^T\tilde{D}w}. \tag{A.1.22}$$

Similarly, using lemmas A.1.1 and A.1.2, the final test statistic is

$$\tilde{T}_{MF-TOWmuT} = \max_w \tilde{T}^0_{score} = \lambda_{max}((\tilde{Y}^T\Phi^{-1}\tilde{Y})^{-1}\tilde{Y}^T\Phi^{-1}\tilde{B}\Phi^{-1}\tilde{Y}), \quad \tilde{B} = \tilde{X}\tilde{D}^{-1}\tilde{X}^T. \tag{A.1.23}$$

A.2 Theoretical Details In Chapter 2

A.2.1 Lemmas and Theorems

Lemma A.2.1. *For each $\tau \geqslant 0$ and $Y \in \mathbb{R}^{n_1\times n_2}$, the solution of*

$$\underset{X}{min} \quad \frac{1}{2}\|X-Y\|_F^2 + \tau\|X\|_* \tag{A.2.24}$$

is $S_\tau(Y) := US_\tau(\Sigma)V^T(= Prox_{\tau\|\cdot\|_*}(Y))$, where $S_\tau(\Sigma) = diag(\{(\sigma_i-\tau)_+\})$, $Y = U\Sigma V^T$, the singular value decomposition of matrix Y, $\Sigma = diag(\{\sigma_i\}_{1\leqslant i\leqslant r})$, r is the rank of Y. $S_\tau(\cdot)$ is called singular value shrinkage operator.

Proof. see (Cai et al., 2010) or (Mazumder et al., 2010). ■

Lemma A.2.2. *For each fixed non-negative λ and $v \in \mathbb{R}^n$, the solution of*

$$\underset{x}{min} \quad \frac{1}{2}\|x - v\|_2^2 + \frac{\lambda}{2}\|x\|_2^2 \tag{A.2.25}$$

is $(Prox_{\frac{\lambda}{2}\|\cdot\|_2^2}(v))_i = sign(v_i)(|v_i| - \lambda)_+$, $i = 1, 2, \cdots, n$, known as the (elementwise) soft thresholding operator.

Proof. see (Parikh and Boyd, 2014). ■

Lemma A.2.3. *For each fixed non-negative ρ and $v \in \mathbb{R}^n$, the solution of*

$$\underset{x}{min} \quad \frac{1}{2}\|x - v\|_2^2 + \rho\|x\|_1 \tag{A.2.26}$$

is $Prox_{\rho\|\cdot\|_1}(v) = (1 - \frac{\rho}{max\{\|v\|_2, \rho\}})v$.

Proof. see (Parikh and Boyd, 2014). ■

Lemma A.2.4. *For the optimization problem*

$$\begin{aligned}
&\underset{X}{min} \quad \frac{1}{2}\|P_\Omega(Y - X)\|_F^2 + \tau\|X\|_* \\
&= \underset{X}{min} \quad \frac{1}{2}\|[P_\Omega(Y) + P_{\Omega^\perp}(X)] - X\|_F^2 + \tau\|X\|_*,
\end{aligned}$$

the optimization solution can be obtained via updating X using $X \leftarrow S_\tau(P_\Omega(Y) + P_{\Omega^\perp}(X))$ with an arbitrary initialization.

Proof. see (Mazumder et al., 2010). ■

Theorem A.2.5. *A sufficient condition for $Prox_{f+g} = Prox_f \circ Prox_g$ is $\forall\ x \in \mathcal{H}$, $\partial g(Prox_f(x)) \supseteq \partial g(x)$, where $\mathcal{H}$ represents Hilbert space and $\circ$ represents composition of two operators.*

Proof. see (Yu, 2013). ■

A.2.2 Algorithms

Algorithm 1: FISTA with constant step size

```
Input: $t_L$, $t_B$, $t_\mu$, $\widetilde{L}_1 = L_0 \in \mathbb{R}^{n\times q}$, $\widetilde{B}_1 = B_0 \in \mathbb{R}^{p\times q}$, $\widetilde{\mu}_1 = \mu_0 \in \mathbb{R}^{q\times 1}$, $t_1 = 1$,
the maximum number of iterations $N$, $\Omega$

Output: optimal feasible solutions $L^*$, $B^*$, $\mu^*$

    for $k = 1$ to $N$
        $L_k \leftarrow S_{t_L\rho}(\widetilde{L}_k - t_L\Omega \odot (X\widetilde{B}_k + \mathbf{1}\widetilde{\mu}_k^T + \widetilde{L}_k - Y))$
        $B_k^1 \leftarrow \widetilde{B}_k - t_B X^T(\Omega \odot (X\widetilde{B}_k + \mathbf{1}\widetilde{\mu}_k^T + L_k - Y))$
        $B_k^2 \leftarrow sign(B_k^1) \odot (|B_k^1| - \lambda_1 J)_+$
        for $j = 1$ to $q$
            $B_k[,j] \leftarrow \{1 - \frac{\lambda_2}{max\{\|B_k^2[,j]\|_2,\lambda_2\}}\}B_k^2[,j]$
        end
        $\mu_k \leftarrow \widetilde{\mu}_k - t_\mu(\Omega \odot (XB_k + \mathbf{1}\widetilde{\mu}_k^T + L_k - Y))^T\mathbf{1}$
        $t_{k+1} \leftarrow (1 + \sqrt{1 + 4t_k^2})/2$
        $\widetilde{L}_{k+1} \leftarrow L_k + \frac{t_k - 1}{t_{k+1}}(L_k - L_{k-1})$
        $\widetilde{B}_{k+1} \leftarrow B_k + \frac{t_k - 1}{t_{k+1}}(B_k - B_{k-1})$
        $\widetilde{\mu}_{k+1} \leftarrow \mu_k + \frac{t_k - 1}{t_{k+1}}(\mu_k - \mu_{k-1})$
        if stopping criteria is satisfied
            break;
        end
    end
```