View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

STATISTICAL ANALYSES OF HIGH THROUGHPUT GENETICS AND GENOMICS DATA

Zhaoyu Yin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2014

Approved by:

Fei Zou

Wei Sun

Haibo Zhou

John Preisser

Patrick Sullivan

© 2014
Zhaoyu Yin
ALL RIGHTS RESERVED

ABSTRACT

ZHAOYU YIN: Statistical Analyses of High Throughput Genetics and Genomics Data
(Under the direction of Fei Zou)

Mixed effects models are commonly used for modeling the dependence structure between twin pairs in twin studies. However, mixed effects models are extremely computationally intensive for eQTL (expression quantitative trait loci) analysis. To overcome the computational challenge, twin pairs can be randomly split into two independent groups on which multiple linear regression analysis can be performed. In my first topic, a computationally efficient score statistic is proposed to combine non-independent analysis results from the two groups.

Genome-wide association studies (GWAS) aim to identify genetic variants associated with complex traits. The standard first pass GWAS analysis where SNPs are tested one at a time may fail to detect associations due to, for example, multiple causal SNPs. Alternatively, regional SNP-set analyses have been established to test the association between a set of SNPs and a phenotype through a mixed effects model where testing the association is equivalent to testing whether one or more of the variance components are equal to 0. However, the null distribution of the likelihood ratio test (LRT) does not follow the conventional $50:50$ mixture chi-square distribution in this setting. My second topic investigates the spectral representation of LRT, based on which an empirical resampling procedure is proposed to approximate the null distribution of LRT.

When both GWAS and gene expression data are available on the same set of samples, it is natural to add gene expression as a covariate into the SNP-set analysis to jointly model the SNP and transcript association with the trait. One biologically interesting question is whether the complex phenotype is associated with the gene expression conditional on the SNP effects. My last research topic jointly models the association between the gene expression and SNP-set with the trait. Unlike traditional mixed effects models, our model allows the gene expression to be dependent on the random SNP effects since the independent assumption is likely to be violated when the gene expression is also associated with the SNP set. With relaxed independence assumption, we can make valid statistical inference and parameter estimation.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Dr. Fei Zou for her guidance throughout my doctoral research. Her knowledge and insights in biostatistics and genetics have greatly helped me to pursue my research and study with enthusiasm. I am also grateful for her generous financial support, persistent caring and encouragement during the entire period of my graduate study.

I also would like to provide my sincere appreciation to my committee members: Dr. Patrick Sullivan, Dr. Wei Sun, Dr. Haibo Zhou and Dr. John Preisser for their helpful insights, comments and suggestions.

Finally, I would like to give my special thanks to my family for supporting and encouraging me to pursue my PhD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Genome-wide Association Study (GWAS)

1.1.1 Introduction of GWAS

Genome-wide association studies (GWAS) have been conducted over the past decade and served as a standard tool to construct associations between genetic variants and complex traits, such as type 2 diabetes, breast cancer and mental disorders. Identifying genetic risk factors of complex diseases is especially beneficial in diagnosis and therapy, and personalized medicine could be developed for each individual patient according to their genetic information [Bush and Moore, 2012; Lewis, 2002]. Single Nucleotide Polymorphisms (SNPs) are the simplest and most prevalent unit of genetic markers. A SNP is a single base-pair change in a DNA sequence and typically carries two alleles. SNPs usually come about once in every 300 base pairs, and there exist roughly 10 million SNPs in the human genome. They play a very momentous role in modern genetic mapping attribute to their high frequency of occurrence and the moderate financial cost of genotyping a large number of SNPs. Genotyping technologies have developed rapidly in the past few years and the cost of genotyping has decreased significantly, thus increasing the availability and leading to a virtual explosion of whole-genome association studies [Laird and Lange, 2011, Section 1.3].

Ragoussis [2009] reviewed several high throughput SNP genotyping platforms. Among all available genotyping platforms, two are most popular: Illumina and Affymetrix. Their popularity is dominant in GWAS. As Ragoussis [2009] mentioned in his review paper, among 209 GWAS papers published until November 2008, data of 103 studies was obtained from Affymetrix Genechips and data of 83 studies was from Illumina's Infinium Beadchips.

1.1.2 Linkage Disequilibrium

Linkage disequilibrium (LD) describes the degree of no-random association between an allele of one SNP and an allele of another SNP in the same region of the genome. LD can arise in a population for many reasons, and the degree of LD is determined by several factors such as recombination rate, mutation, and demographic features. LD plays an important role in association studies, because suppose all markers are independent at the population scale, the association of every SNP with a trait needs to be investigated, which would present a significant challenge, however, due to the existence of LD, multiple variants could be highly correlated, thus only a subset of SNPs needs to be genotyped in association studies [Laird and Lange, 2011, Section 5.4].

Considering two loci A and B, there exist four possible haplotypes in a population, due to each separation for two alleles (A,a and B,b). There are several measures of LD, all of which compare the discrepancy between the observed haplotype frequencies and the haplotype frequencies expected under the null assumption of independence between the two markers. Lewontin and Kojima [1960] proposed the linkage disequilibrium coefficient D and Lewontin [1964] proposed a scaled version of D, D' where D is standardized by its maximum possible value and D' varies between 0 and 1. $D' = 0$ implies the two loci are independent from each other, whereas $D' = 1$ implies perfect

LD. Alternatively, the correlation coefficient r^2 (elsewhere, r^2 is also denoted by Δ^2) was proposed by Hill and Robertson [1968] for the purpose of genetic analysis, which ranges from 0 to 1. High r^2 values indicate that the two SNPs carry similar genetic information and an r^2 of 1 indicates perfect predictability of one SNP from another SNP, whereas an r^2 of 0 indicates the markers are in perfect equilibrium. D' and r^2 are currently the two most widely used LD measurements nowadays [Laird and Lange, 2011, Section 5.4].

Bush and Moore [2012] discussed two categories of the association between a genetic polymorphism and a trait, which are direct and indirect association due to the existence of LD. Specifically, direct association indicates the genotpyed SNP itself is the functional (causal) SNP that can affect biological mechanisms underlying a trait and result in its variation. In other circumstances, only tagSNPs are genotyped and serve as surrogates for the causal locus. Due to the presence of LD between tagSNPs and the causal SNP, the indirect association might be detected between one or more of the tagSNPs and the trait. However, the power to detect significant associations is lower with tagSNPs than that with modeling causal SNPs directly.

1.2 Association Mapping

The goal of GWAS is to investigate the association between genetic variants and complex diseases/phenotypes. The response variable in any genetic association study can be quantitative or qualitative (often binary). The association between the complex trait and the genetic variant can be measured or tested through regression models. The standard method is individual SNP analysis where the effect of each SNP on a trait is examined separately and independently. A list of candidate loci can be selected based on their $p-$values less than a given threshold after multiple testing correction [Kraft and Cox, 2008]. Specifically, a linear regression model is used for quantitative trait; a logistic regression or contingency table is used for a binary trait. For complex diseases, such as obesity and asthma, affection status is often defined by an intermediate phenotype or endpoint phenotype, such as body mass index for obesity or forced expiratory volume in one second (FEV1)for asthma [Laird and Lange, 2011, Section 7.8]. Compared to qualitative traits, association studies with a quantitative measurement as the response variable can be more reproducible, have greater power to detect genetic effects evidence, and offer better interpretations [Bush and Moore, 2012]. In addition to genetic variants, other non-genetic factors such as age or gender are also available to be adjusted for in the statistical model. For a single locus, based on modes of inheritance, genetic models can be divided into four categories: additive, recessive, dominant and multiplicative [Lewis, 2002]. Although the true genetic model is rarely known in practice, signals from both additive and dominant genetic effects can be identified with fairly good power in GWAS using additive models, thus additive models are the most popular models for GWAS data [Bush and Moore, 2012; Lettre et al., 2007].

1.2.1 Methods for Qualitative Traits

In a standard case-control GWAS, a large number of SNPs are genotyped among thousands of individuals with diseases and also for thousands of healthy individuals. The aim is to identify an initial collection of promising susceptibility loci. The frequencies of SNP alleles are compared between the case and control groups. Suppose at a given gene, the SNP has two alleles g and G with three possible genotypes gg, Gg and GG on a set of cases and controls, the observed frequencies can be summarized in the following 2×3 contingency table of Table 1.1 [Laird and Lange, 2011, Section 7.1].

If the cases carry a given SNP allele with higher frequency in contrast to control subjects, this implies that the presence of the SNP allele could raise the disease risk [Lewis, 2002]. The standard Pearson's chi-square test statistic is used to test the independence

Table 1.1: An example of 2 × 3 contingency table of genotypes in case-control studies

	gg	Gg	GG	Total
Cases	n_{11}	n_{12}	n_{13}	$n_{1.}$
Controls	n_{21}	n_{22}	n_{23}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

as [Pearson, 1909, 1910]:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} is the observed frequency of the genotype in each cell and E_{ij} is the expected value under the null hypothesis. The test statistic approximately follows χ^2_2. If the sample size is small or an expected cell count is less than 5, then the asymptotic approximation of the null distribution χ^2 is no longer valid. Fisher's exact test is used to calculate the significance of the deviation from the null hypothesis exactly, via a hypergeometric distribution [Tomlinson et al., 2007]. Baz et al. [2008] conducted a case-control study in a Turkish population and used the χ^2_2 test to determine if the polymorphisms $TNF - \alpha$ and acne are significantly associated.

For the above test, no prior genetic information is used. However, if prior information is available or a supposition exists that a greater number of allele G will increase the risk of disease, then other methods that reflect the monotone trend may be preferred. The alleles test and the trend test are two commonly used trend testing methods [Laird and Lange, 2011, Section 7.2]. For instance, Klein et al. [2005] used the alleles test to identify genetic variants strongly associated with age-related macular degeneration (AMD). However, Sasieni [1997] noted that an alleles test is only valid if the Hardy-Weinberg equilibrium (HWE) holds under the null hypothesis. If this assumption is not satisfied, the significance level of the test from the alleles test cannot be maintained. As an alternative, the Cochran-Armitage trend test [Armitage, 1955] relaxes the HWE assumption and keeps all desirable characteristics of the alleles test and thus is popular in GWAS. For example, Beecham et al. [2009] attempted to identify late-onset Alzheimer disease risk loci in a study using 492 cases and 498 controls. By Cochran-Armitage trend test, the 12q13 locus was detected to be significantly associated with late-onset Alzheimer's disease.

Logistic regression is an alternative way to evaluate genetic associations for dichotomized disease phenotypes. In addition to genetic variants, many environmental factors could contribute to variations of complex traits, such as age, gender and other demographic characteristics. Inclusion of covariates can dramatically remove their confounding effects and may improve power. Pirinen et al. [2012] studied the impact of including known covariates on the power of detecting the association in case-control studies. The inclusion of the covariates in logistic regression models generally increases power for common traits. Logistic regression can flexibly model the covariates and has become a standard tool in most GWAS packages such as PLINK by Purcell et al. [2007]. Yu et al. [2012] used logistic regressions in PLINK to perform a two-stage lgA nephropathy (lgAN) study in Han Chinese and identified genome-wide associations at 17q13 and 8p23.

1.2.2 Methods for Quantitative Traits

Quantitative measures better characterize some complex traits, such as high blood pressure, obesity and asthma. For single SNP analysis, Analysis of Variance (ANOVA) is popular. The null hypothesis is that the mean values of a phentoype are the same among all genotype groups. Linear regression is another popular approach for analyzing quantitative traits among n independent samples,

$$E(Y_i|X_i, B_i) = \beta_0 + \beta_1 X_i + \beta_2 B_i, i = 1, ..., n,$$

where for the ith individual, Y_i is the quantitative trait, X_i represents the genetic variant and B_i is a vector of other covariates to be adjusted for. The advantages of applying linear regression in GWAS include easy incorporation of covariates under the explicit parametric model and convenient conduction of hypothesis testing through likelihood ratio or score test [Laird and Lange, 2011, section 7.7]. Li et al. [2007] scanned a genome consisting of $362,129$ SNPs among $4,305$ Sardinian individuals and reported that SNPs in GLUT9 are associated with Uric Acid (UA), by applying a linear regression model. Loos et al. [2008] performed data analysis for genom-wide association data from $16,876$ subjects using a linear regression model to identify the common variants affecting body mass index.

1.3 Expression Quantitative Trait Loci (eQTL)

In many situations, the genetic variants detected by GWAS can only explain a small portion of the heritability associated with complex traits. Furthermore, the functional and regulatory consequences of the detected genetic variants are not clear, making the identification of causal genetic variants challenging [McCarthy and Hirschhorn, 2008]. Gene expression, as an intermediate molecular phenotype, can be related to genetic variants and cause variations in complex traits [Dermitzakis, 2008; Morley et al., 2004]. With widely available high throughput technologies, it is common to measure both gene expression and genetic variants on the same set of samples. To gain deep understanding of the possible regulatory role of SNPs underlying complex traits, expression quantitative trait loci (eQTL) analysis aims to determine whether genetic variants lead to the variation in gene expression which is treated as a phenotype in association studies [Gilad et al., 2008]. eQTL analyses have proven to be useful in various ways, for example, investigation of the SNP-transcript associations, discovery of eQTL hotspots (genomic regions play regulatory roles for different transcripts), classification of clinical phenotypes into a cluster of subcategories depending on the clinical characteristics and determination of lists of candidate genes based on the knowledge from GWAS for complex diseases [see reviews of Kendziorski and Wang, 2006; Wright et al., 2012].

Various analyses have become well established. Simple linear regression is one of the most commonly used models for eQTL analysis. With millions of SNPs and thousands of transcripts among thousands of individuals in modern eQTL data, eQTL analysis is extremely computationally intensive. Shabalin [2012] proposed Matrix eQTL as a tool for more computational efficient eQTL analysis as follows:

$$g = \alpha + \beta s + \gamma x + \epsilon, \quad \epsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

where g is the gene expression, s is the SNP genotype and x represents other covariates. Shabalin [2012] leveraged orthogonalization techniques for the gene expression and SNP with respect to other covariates so that the multiple linear regression model was simplified to a simple linear regression model. Within the simple model framework, test statistics, including t, F and LRT for the null hypothesis $H_0 : \beta = 0$, are functions of the sample correlation $r = \text{cor}(g, s)$. For example, $t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$. The friendly used software, such as Genevar [Yang et al., 2010] and eQTL viewer [Zou et al., 2007], is available for eQTL analysis, output visualization and association results interpretation.

1.4 Multiple Testing Correction

In GWAS analysis, a large number of SNPs are separately tested and multiple testing correction is necessary to control family-wise error. The simplest and most common approach is the Bonferroni correction, which is usually conservative and has negative impact on statistical power in GWAS [Han et al., 2009; Nicodemus et al., 2005], since it assumes all the tests are independent from each other [Sidak, 1967]. However, neighboring SNPs on a chromosome are likely to be highly correlated due to the presence of the LD and incline to be inherited together (International HapMap Consortium). Another correction method is permutation; it has been widely used in GWAS [Dudbridge, 2006; Tenesa et al., 2008]. One advantage of this method is that the correlation among SNPs is preserved and thus it is less conservative than the Bonferroni correction. However, a permutation procedure can be computationally intensive for large association studies. Zou et al. [2004] proposed an efficient resampling algorithm to determine significance threshold with overall type I error control based on a score test statistic, expressed by a sum of independent random vectors; each vector represents the contribution from one subject to the test statistic. The score test statistic only requires the calculation of estimates under the null hypothesis. Zou et al. [2004] established a detailed derivation of the score test statistic for mapping quantitative trait loci. A multiple linear regression model is considered to test the association between a single SNP and a quantitative trait with inclusion of other non-genetic covariates:

$$y_i = \beta g_i + \gamma x_i + \epsilon_i, \quad i = 1, \cdots, n,$$

where y_i is the phenotype of the ith individual, β is the genetic effect, $\gamma = (\gamma_0, \gamma_2, ..., \gamma_q)$ is a vector of coefficients for the intercept and non-gene covariates $x_i = (1, x_{i1}, ..., x_{iq})$. The log-likelihood of $\theta = (\beta, \gamma, \sigma^2)$ takes the form

$$l(\theta) = \sum_{i=1}^{n} l_i(\theta),$$

where $l_i = -\frac{1}{2}\log\sigma^2 - \frac{(y_i - \beta g_i - x_i\gamma)^2}{2\sigma^2}$. In general, we test the null hypothesis $H_0 : \beta = 0$ in presence of the nuisance parameter vector $\eta = (\gamma, \sigma^2)$.

We follow the process in Zou et al. [2004] and denote $U_{\beta,i}(\theta) = \frac{\partial l_i(\beta,\eta)}{\partial\beta}$ and $U_{\eta,i}(\theta) =$

$\frac{\partial l_i(\beta,\eta)}{\partial \eta}$. The restricted MLE $\tilde{\eta}$ is obtained by solving $\sum_{i=1}^n U_{\eta,i}(\theta) = 0$ under $\beta = 0$. $\tilde{\eta}$, the estimate of η, does not change from SNP to SNP, and thus only needs to be estimated once for each transcript in the case that multiple SNPs are examined for one transcript in eQTL analysis. Denote $U = \sum_{i=1}^n U_{\beta,i}(0,\tilde{\eta})$ as the score function for β. The Taylor expansion and law of large numbers show the equivalence of asymptotic distribution between $n^{-1/2}U$ and $n^{-1/2}\sum_{i=1}^n U_i$, where

$$U_i = U_{\beta,i}(0,\eta) - \Sigma_{\beta,\eta}(0,\eta)\Sigma_{\eta,\eta}(0,\eta)^{-1}U_{\eta,i}(0,\eta)$$

and $\Sigma_{\beta,\eta}(\beta,\eta)$ is $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \frac{\partial^2 l_i}{\partial\beta\partial\eta}$ and $\Sigma_{\eta,\eta}(\beta,\eta)$ is $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \frac{\partial^2 l_i}{\partial\eta\partial\eta}$ [Cox and Hinkley, 1974, Section 9.3]. The U_is for $i = 1,...,n$ are independent random variables with mean zero. Zou et al. [2004] claimed that $n^{-1/2}U$ is a zero-mean Gaussian process with variance Ξ, which is $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n U_i U_i'$. Substituting all unknown parameters by their sample estimator leads to

$$\hat{U}_i = U_{\beta,i}(0,\tilde{\eta}) - \Sigma_{\beta,\eta}(0,\tilde{\eta})\Sigma_{\eta,\eta}(0,\tilde{\eta})^{-1}U_{\eta,i}(0,\tilde{\eta}).$$

The consistency of MLE with the law of large numbers indicates that $n^{-1}\sum_{i=1}^n \hat{U}_i\hat{U}_i{}'$ could be used to estimate Ξ. Denote $\hat{U} = \sum_{i=1}^n \hat{U}_i$ and $\hat{V} = \sum_{i=1}^n \hat{U}_i\hat{U}_i{}'$, the score test statistic for $H_0 : \beta = 0$ is

$$W = \hat{U}'\hat{V}^{-1}\hat{U}.$$

For a single SNP analysis in GWAS, g_i represents the genotype score of one SNP in the additive model. Then U_i is a scalar rather than a vector, thus the expression of the score test statistic reduces to

$$W = \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i{}^2},$$

which is asymptotically distributed as χ_1^2 and equivalent to LRT when the sample size is large.

In contrast to a permutation test, the resampling method only needs to maximize the likelihood of the observed data once, then the significance threshold can be determined. Thus it is computationally less demanding. Alternatively, Diao and Vidyashankar [2013] proposed a modified resampling approach, where the standard normal distribution was replaced by a Rademacher distribution, that is, the random variable takes 1 or -1 with equal probability.

The resampling procedure proposed by Zou et al. [2004] is as follows:

1. Simulate G_i independently from $N(0,1)$ (Diao and Vidyashankar [2013] generated G_i from a Rademacher distribution, where G_i takes the value 1 or -1 with equal probability).

2. Let $U^*(d) = \sum_i \hat{U}_i(d) G_i$ and $W^* = U^*(d)^T(\hat{V}(d))^{-1}U^*(d)$, then set T^* to be the maximum value of $W^*(d)$ for all possible locations d.

3. Repeat the above steps N times where N is a large integer.

4. Compute the $100(1-\alpha)$th percentile of $(T_1^*, \cdots, T_N^*)$ as the threshold for a given significance level α.

Here d is the SNP location when multiple markers are tested. This resampling method preserves the correlation structure among multiple score test statistics via the standard normal random variable G_i.

False discovery rate (FDR) [Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2005] is an alternative method for multiple hypotheses testing correction. Compared to a family-wise error rate, the FDR controlling procedure is less stringent and works well in the situation where the test statistics are dependent [Benjamini and Yekutieli, 2001, 2005]. The FDR method has been found to be useful for multiple testing correction in GWAS [Jia et al., 2010; Weng et al., 2011; Wright et al., 2011].

1.5 Twin Studies

Subjects in genetic association studies may be unrelated, or from one family with high correlations such as twin studies. A twin study is a very different study design from a traditional association study, thus a specialized model is needed to account for correlations between twin pairs. Twin studies are commonly used to investigate the associations between genetic variants and complex traits [Chou et al., 2009; Park et al., 2012; Vaccarino et al., 2008]. Typical twin data includes monozygotic twins (MZ) and dizygotic twins (DZ), plus some singletons. MZ twins carry identical genetic information and are more similar than DZ twins who only share around 50% of their genes on average. Given that MZ and DZ twins grow in the same environment, the presence of a higher phenotypic correlation indicates that the phenotype is more genetically related between MZ twins than between DZ. Twin studies are often helpful to estimate the heritability by evaluating the contribution from genetic factors to the variation of a complex trait [Boomsma et al., 2002; Neale and Cardon, 1992; Silventoinen et al., 2003; True et al., 1993]. Unlike studies with independent samples, twin studies require more careful statistical modeling techniques, since neglecting genetic relatedness and shared environments among twins may result in high false positive findings.

One popular method is mixed effects models, which have been widely applied to analyze twin and family data with correlation considerations [Carlin et al., 2005; Kuna et al., 2012; Wang et al., 2011]. Another commonly used approach for twin data analysis is structural equation modeling (SEM) [Neale et al., 1989]. SEM is used to measure the contribution from genetic factors to a trait by partitioning the total variation of a trait nto four components. Specifically, Neale et al. [1989] decomposed the total variation of a phenotype into additive genetic effects (a_i), dominant genetic effect (d_i), common environment effect (c_i) and random noise (e_i), where a_i, c_i, d_i and e_i are independent from each other and are normally distributed random variables following $N(0, \sigma_a^2)$,

$N(0,\sigma_d^2)$, $N(0,\sigma_c^2)$ and $N(0,\sigma_e^2)$ respectively. The genetic model can be written as

$$y_i = \mu + g_i\beta + x_i\gamma + a_i + c_i + d_i + e_i, \quad i = 1,\cdots,n,$$

where for the ith individual, y_i is the trait of interest, μ is the grand mean, g_i represents genetic variants, and x_i denotes covariates. Referring to Falconer and Mackay [1996], the covariances from the additive genetic effects $cov(a_i,a_j)$ for MZ pairs and DZ pairs are σ_a^2 and $\sigma_a^2/2$ respectively; the covariances from dominance genetic effects $cov(d_i,d_j)$ for MZ pairs and DZ pairs are σ_d^2 and $\sigma_d^2/4$ respectively; the covariance of common environmental effect for any twin pairs is $cov(c_i,c_j) = \sigma_c^2$; the covariances from additive, dominant and common environment effects are zero for any pair of unrelated individuals. The above model is referred as the ACDE model, however, if parental data are not available, not all random effects can be estimated due to an identifiability issue, in which situation the ACE model is generally used where $\sigma_d^2 = 0$ [Feng et al., 2009; Wang et al., 2011]. The covariance structures for any pair of MZ twins, DZ twins and unrelated singletons are listed as follows,

$$\begin{aligned}
cov\begin{pmatrix} a_i \\ a_j \end{pmatrix} &= \sigma_a^2\begin{pmatrix} 1 & \rho_{ij}^a \\ \rho_{ij}^a & 1 \end{pmatrix}, \\
cov\begin{pmatrix} d_i \\ d_j \end{pmatrix} &= \sigma_d^2\begin{pmatrix} 1 & \rho_{ij}^d \\ \rho_{ij}^d & 1 \end{pmatrix}, \\
cov\begin{pmatrix} c_i \\ c_j \end{pmatrix} &= \sigma_c^2\begin{pmatrix} 1 & \rho_{ij}^c \\ \rho_{ij}^c & 1 \end{pmatrix},
\end{aligned}$$

where

$$\rho_{ij}^{a} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are MZ pairs} \\ 1/2, & \text{if } i \text{ and } j \text{ are DZ pairs} \\ 0, & \text{if } i \text{ and } j \text{ are unrelated} \end{cases}$$

$$\rho_{ij}^{d} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are MZ pairs} \\ 1/4, & \text{if } i \text{ and } j \text{ are DZ pairs} \\ 0, & \text{if } i \text{ and } j \text{ are unrelated} \end{cases}$$

$$\rho_{ij}^{c} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are MZ or DZ pairs} \\ 0, & \text{if } i \text{ and } j \text{ are unrelated.} \end{cases}$$

1.6 SNP-set Analysis in GWAS

Although single SNP analysis has proven to be quite useful for identifying many genetic variants associated with complex diseases, this analysis may fail to detect associations in some situations, due to several reasons, for example, stringent threshold for multiple testing correction when a large number of association are examined, causal SNPs not genotyped and the existence of multiple causal SNPs nearby associated with a trait [Wu et al., 2010]. SNP-set analysis is an alternative approach to address the above limitations, where multiple SNPs can be combined to be a SNP-set as one test unit in association studies, depending on the biological information and pre-specified criteria [Mayhew et al., 2013; Tzeng et al., 2011; Wu et al., 2011]. Then association test with the trait is examined for each SNP set rather than for individual SNPs. The SNP-set analysis combines information from a set of SNPs and aggregates small effects from multiple individual loci. Wu et al. [2010] discussed several grouping strategies to construct meaningful SNP sets based on their positions or correlations. Grouping SNPs in LD with the causal but untyped SNP properly and treating multiple typed SNPs as one test unit in association studies could enhance the power to identify causal effect with the trait [Schaid et al., 2002]. Moreover, SNP set analyses decrease the number of tests dramatically and thus relieve the stringent significance threshold. Furthermore, if there is more than one independent causal SNPs, their joint activities can be detected with considerable power by performing SNP-set analysis [Wu et al., 2010].

A variety of approaches have been proposed for detecting the associations between SNP-set and trait. Wu et al. [2011] proposed a sequence kernel association test (SKAT) for detecting both common and rare variants with the trait through testing for one variance component using the score test statistic. Tzeng et al. [2009] and Tzeng et al. [2011] established a similarity-based regression approach to investigate associations between a trait and multi-marker genotypes, where they regressed the trait similarities on the haplotype similarities and utilized a score test statistic to test the corresponding coefficients in the regression models. Furthermore, they showed that the score test in their similarity regression model was equivalent to the variance component method where the individual genotypes are treated as random effects under a mixed effect model framework. They claimed that this similarity-based method is superior to genotype sum methods when genetic effects have opposite directions. Mayhew et al. [2013] proposed two likelihood ratios based approaches under mixed effects model framework, which is equivalent to genotype-phenotype similarity testing and the genetic similarity is based on the correlation of multiple markers for each pair of unrelated individuals. In general, Tzeng et al. [2011] classified SNP-set analysis methods into four categories based on how information from a collection of markers is put together, including considering weighted sum of genotypes across markers, U-statistics approaches, variance component methods and all remaining methods. U-statistics approach models the association between the pairwise genetic similarity of individuals and their phenotype similarity. Variance component method treats genetic effects of each individual as random and then the relevant variance components are tested to determine the joint effect of each SNP-set on the complex trait.

A linear mixed model with one variance component can be expressed as follows

$$Y = X\beta + Zb + \varepsilon, \quad E\begin{pmatrix} b \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 0_K \\ 0_n \end{pmatrix}, \; cov\begin{pmatrix} b \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2\Sigma & 0 \\ 0_n & \sigma_\varepsilon I_n \end{pmatrix},$$

where Y is a n-dimensional response vector, $\beta_{p\times1}$ is a vector of parameters corresponding to fixed effects, K is the number of SNPs in the set, $b_{K\times1}$ is a vector of random effects from individual genetic effects, $\Sigma_{K\times K}$ is a known symmetric positive definite matrix, b and ϵ are independently and normally distributed. If Y is a trait vector and Z is a $n \times K$ matrix used to quantify genetic similarity, then testing for $H_0 : \sigma_b^2 = 0$ vs $H_1 : \sigma_b^2 \geq 0$ can help determine if SNP-set similarity is significantly associated with trait similarity. The null distribution of LRT does not follow the standard 50:50 mixture of χ_0^2 and χ_1^2 in the setting where the genetic effects of all individuals are not independent [Crainiceanu and Ruppert, 2004]. The $50 : 50$ ratio holds only if the response variable can be written as a vector including a great deal of independent random variables identically distributed under both the null and alternative hypotheses [Self and Liang, 1987; Stram and Lee, 1994]. This i.i.d. assumption is not true for SNP-set analysis using the model above where the design matrix of random effects Z cannot be written in the form of a block diagonal matrix due to the dependency of genetic effects from multiple SNPs among subjects, that is, the response vector cannot be represented as a collection of independent random variables under the alternative hypothesis [Tzeng and Zhang, 2007]. If this i.i.d. assumption is violated, Pinheiro and Bates [2000] in a simulation study found that $0.65\chi_0^2 + 0.35\chi_1^2$ mixture properly approximates the null distribution of LRT statistic. Tzeng and Zhang [2007] stated that the underpower phenomenon from the variance component method for haplotype-based similarity association analysis comes about because the standard $50:50$ ratio was used. Crainiceanu et al. [2003] investigated the null distribution of the LRT statistic and proved that the mixing proportion of χ_0^2 can be much larger than 0.5, thus the significance threshold determined by $0.5\chi_0^2 + 0.5\chi_1^2$ is too conservative for hypothesis testing. Crainiceanu and Ruppert [2004] relaxed the i.i.d. data assumption and developed an efficient resampling procedure based on the spectral decomposition of LRT statistic to derive the finite sample null distribution of the LRT statistic, which only depends on the eigenvalues from two low dimensional design matrices.

Although SNP-set analyses have proven to be successful in GWAS, it is beneficial for us to integrate other genomic data such as gene expression into association studies for assisting our understanding of biological mechanisms underlying complex traits. Gene expression, as an intermediate molecular phenotype, can be affected by SNPs and also cause variations of complex traits [Dermitzakis, 2008; Morley et al., 2004]. The presence of some diseases is the consequence of the joint effects from both genetic and expression level variation with several environmental factors, thus other valuable genomic information may be neglected if only a single genomic data type is leveraged in the association studies [Xiong et al., 2012]. With the advancement of high throughput technologies, it is common to measure both gene expression and SNPs on the same set of individuals. With the availability of multiple types of genomic data, it is natural to analyze both genetic and gene expression data jointly for a trait in a statistical model for association studies. Several methods have been proposed to study the joint relationship of genetic variants and transcripts with a trait in association studies [Huang et al., 2014; Schadt et al., 2005; Xiong et al., 2012; Zhu et al., 2008]. Huang et al. [2014] jointly modeled the effects of a SNP-set and the corresponding gene expression on the disease status. In their paper, a logistic mixed model was firstly used to regress the dichotomous outcome on both SNP-set and gene expression with their interactions after adjusting for other covariates, then a multiple linear regression model was leveraged to model the relationship between the continuous gene expression and the set of SNPs. The testing procedure for the total effect of a gene from both SNPs and the gene expression was conducted within a causal mediation analysis framework.

1.7 Outline of Thesis

The thesis starts with the literature reviews in the first chapter and the remaining parts are organized as follows. In Chapter 2, we develop a novel and computationally efficient score test statistic to perform eQTL analysis for twin data. In Chapter 3, we investigate a spectral decomposition of LRT and resampling algorithm to approximate the null distribution of LRT, which is used to test the association between a set of SNPs and a phenotype. We propose a modified version of the resampling procedure to approximate the null distribution of LRT for testing the joint effects of SNP-set and the individual SNP with most significant signal on the trait. In Chapter 4, we propose an integrative approach to model the association between the gene expression and the trait conditional on SNP effects in a mixed effects model.

CHAPTER 2

A FAST EQTL ANALYSIS FOR TWIN STUDIES

2.1 Introduction

Genome-wide association studies (GWAS) have been widely used over the past decade for identifying genetic variants associated with a diversity of complex human diseases, such as type 2 diabetes, breast cancer and psychiatric disorders [Garcia-Closas et al., 2013; Hanson et al., 2014; Winham et al., 2013]. These studies have identified a large number of disease associated SNPs (single nucleotide polymorphisms). However, the majority of the SNPs detected by GWAS individually explain a very small fraction of the total heritability associated with these traits, with no immediately clear functional or regulatory roles [McCarthy and Hirschhorn, 2008]. Gene expression, as an intermediate molecular phenotype, may provide additional insight into the regulatory roles of SNPs implicated by GWAS. With widely available high throughput technologies, gene expression and genetic variant data can be collected simultaneously on the same set of individuals, and expression quantitative trait loci (eQTL) analysis can be performed to assess which genomic regions and genetics variants lead to gene expression variations [Gilad et al., 2008]. eQTL analyses have proven to be useful in various ways, for example, investigation of the SNP-gene expression associations, discovery of eQTL hotspots (genomic regions play regulatory roles for different transcripts), classification of clinical phenotypes into multiple subcategories depending on the clinical features and determination of lists of candidate genes based on the knowledge from GWAS for complex diseases [see the reviews of Kendziorski and Wang, 2006 and Wright et al., 2012]. Recent research has also shown that SNPs detected by GWAS are significantly more likely to be eQTL, which can be used to boost the discovery of genetic variants associated with the trait, and improve understanding of the molecular mechanism of complex traits [Min et al., 2011; Nica et al., 2010; Nicolae et al., 2010].

Various eQTL analytical approaches have been established where an association test is performed between one transcript and one SNP at a time by linear regression analysis, analysis of variance [Shabalin, 2012], generalized linear regression [Hernandez et al., 2012], or mixed effects models [Kang et al., 2008] depending on the type of eQTL data. More complicated analytical procedures, such as Bayesian regression [Bottolo et al., 2011; Chipman and Singh, 2011; Stegle et al., 2010] and partial least square regression [Chun and Keles, 2009] have also been applied to eQTL data. In addition, several methods have been proposed for detecting associations between a group of SNPs and the gene expression of each transcript [Hoggart et al., 2008; Michaelson et al., 2009; Zhen, 1994]. User friendly software such as Genevar [Yang et al., 2010] and eQTL viewer [Zou et al., 2007] have been developed for eQTL data analysis, output visualization and result interpretation. The high dimensionality of eQTL data makes modern eQTL analysis computationally intensive, given the fact that associations between several million SNPs and tens of thousands of transcripts need to be tested. To mitigate this heavy computational burden, analysis may be restricted to a small number of SNP-transcript pairs, however it is still computationally challenge [Ghazalpour et al., 2008]. Alternatively, Shabalin [2012] has developed Matrix eQTL as a fast eQTL analytical tool which is thousands of times faster than any existing QTL/eQTL software. Matrix eQTL is extremely computationally efficient because it expresses the association test as a function of the correlation between one SNP and one transcript, which can be realized by a quick matrix operation.

For complex psychiatric disorders, such as schizophrenia and major depressive disorder, twin studies have received attention for establishing the general extent to which genes and environment are etiologically important [Boomsma et al., 2002; Chou et al., 2009; Neale and Cardon, 1992; Park et al., 2012; Silventoinen et al., 2003; Vaccarino et al., 2008]. Typical twin data include both monozygotic twins (MZ) and dizygotic twins (DZ), plus unpaired individual twins (singletons). MZ twins are assumed to be genetically identical, while DZ twins share 50% of their genes on average. Assuming that MZ and DZ twins share the same environment, a higher phenotypic similarity between MZ twins compared to DZ twins indicates that the phenotype is genetically controlled. Unlike data with independent samples, twin data require more careful statistical modeling since ignoring genetic relatedness and shared environment among twin pairs may lead to high false and/or low true positive findings. Several statistical approaches are available for twin data. One of the most common approaches is structural equation modeling (SEM) Neale et al. [1989]. Several software programs to perform SEM are available, such as Mx [Neale et al., 1999], Mplus [Muthen and Muthen, 1998], LISREL [Jorsekog and Sorborn, 1986] and OpenMx [Boker et al., 2011, 2012]. Another popular alternative for twin and family data is the mixed-effects model where random effects are used to properly account for the correlations among subjects [Carlin et al., 2005; Kuna et al., 2012; Wang et al., 2011]. Mixed model have a well-established theory which is familiar to statisticians. Moreover, it is conveniently implemented in most statistical software and can flexibly adjust other non-genetic and genetic covariates [Feng et al., 2009; Rabe-Hesketh et al., 2008].

Though powerful, the mixed effects models are computationally intensive and impractical for modern twin eQTL analysis. Moreover, the ultra fast tool Matrix eQTL is not readily applicable to twin data since it does not model the dependence structure between twin pairs. To overcome these computational challenges, we propose a novel fast twin eQTL analysis approach. In this approach, we first randomly split the twin pairs into two groups such that within each group, the samples are unrelated. We then run a separate analysis for each group using any statistical procedure valid for independent data, such as multiple linear regression or analysis of variance. When combining the results from the two groups, we find traditional meta-analysis procedures, such as Fisher's test, is no longer applicable since the two sets of results are not independent due to the correlation between twin pairs. Naively combining the two sets of p-values without consideration of the dependence structure of the data would lead to inflated false positive findings. In this paper, we propose a novel score test which automatically adjusts the (hidden) correlation structures between twin pairs, and therefore controls the type I error accurately. To demonstrate the computational advantages and evaluate the performance of the proposed method, we conduct extensive simulations under various settings to mimic real world twin data.

Our simulation results establish that our proposed approach controls type I error rates well, with negligible power loss compared to the gold mixed effects model. Furthermore, the computational efficiency of the proposed method is dramatically improved. The proposed method is more than a thousand times faster than the mixed effects model. The fast performance of the proposed method is achieved by computing the most computationally intensive part in the score test by matrix operations, similar to what has been done in matrix eQTL. The utility of the proposed method is further illustrated by analyzing a twin eQTL data where the twin samples (~ 4,000) are from the Netherlands twin registry.

2.2 Methods

The first step of the proposed approach is to randomly split each twin pair into two groups such that all samples within each group are unrelated. Thus a statistical procedure valid for independent data can be directly applied to each group for testing SNP-transcript pair associations. For simplicity and because of its popularity for eQTL data, multiple linear regression is considered. Specifically, for each group, given a transcript and a SNP, we fit the following linear regression model:

$$y_i = \beta g_i + \boldsymbol{x}_i \boldsymbol{\gamma} + \epsilon_i, \quad i = 1, \cdots, n,$$

where for the ith individual, y_i is the gene expression, g_i is the SNP genotype which is coded 0, 1 or 2 according to the number of minor alleles in the genotype; and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_q)'$ is a vector of parameters corresponding to the vector of non-genetic covariates plus the intercept $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{iq})$. Rewriting the above model in matrix form, we get

$$Y = \beta G + X\boldsymbol{\gamma} + \epsilon,$$

where $Y = (y_1, y_2, ..., y_n)'$, $G = (g_1, g_2, ..., g_n)'$ and $X = (\boldsymbol{x}'_1, \cdots, \boldsymbol{x}'_n)'$. The log-likelihood function is therefore

$$l(\theta) = \sum_{i=1}^{n} l_i(\theta)$$

where $\theta = (\beta, \boldsymbol{\gamma}, \sigma^2)$ and $l_i = -\frac{1}{2}\log\sigma^2 - \frac{(y_i - \beta g_i - \boldsymbol{x}_\mathbf{i}\boldsymbol{\gamma})^\mathbf{2}}{2\sigma^2}$. For the eQTL analysis, the null hypothesis is $H_0 : \beta = 0$ where the vector of the nuisance parameters is $\boldsymbol{\eta} = (\boldsymbol{\gamma}, \sigma^2)$. Following the derivation procedure of Zou et al. [2004], we get the score function of the ith individual as

$$U_i = U_{\beta,i}(0, \boldsymbol{\eta}) - \Sigma_{\beta,\boldsymbol{\eta}}(0, \boldsymbol{\eta})\Sigma_{\boldsymbol{\eta},\boldsymbol{\eta}}(0, \boldsymbol{\eta})^{-1} U_{\boldsymbol{\eta},i}(0, \boldsymbol{\eta}),$$

where $U_{\beta,i}(\beta,\boldsymbol{\eta})$ and $U_{\boldsymbol{\eta},i}(\beta,\boldsymbol{\eta})$ are defined as

$$\begin{aligned} U_{\beta,i}(\beta,\boldsymbol{\eta}) &= \tfrac{\partial l_i(\beta,\boldsymbol{\eta})}{\partial\beta} = \tfrac{(y_i-\beta g_i-\boldsymbol{x}_i\boldsymbol{\gamma})g_i}{\sigma^2}, \\ U_{\boldsymbol{\eta},i}(\beta,\boldsymbol{\eta}) &= \tfrac{\partial l_i(\beta,\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} = \left(\begin{array}{c} \tfrac{(y_i-\beta g_i-\boldsymbol{x}_i\boldsymbol{\gamma})\boldsymbol{x}_i'}{\sigma^2} \\ \tfrac{(y_i-\beta g_i-\boldsymbol{x}_i\boldsymbol{\gamma})^2}{2\sigma^4}-\tfrac{1}{2\sigma^2} \end{array}\right), \text{ respectively,} \end{aligned}$$

and $\Sigma_{\beta,\boldsymbol{\eta}}(\beta,\boldsymbol{\eta})$ is the limit of $n^{-1}\sum_{i=1}^n \frac{\partial^2 l_i}{\partial\beta\partial\boldsymbol{\eta}}$, and $\Sigma_{\boldsymbol{\eta},\boldsymbol{\eta}}(\beta,\boldsymbol{\eta})$ is the limit of $n^{-1}\sum_{i=1}^n \frac{\partial^2 l_i}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}}$ as $n\to\infty$. The two Hessian matrices are

$$\begin{aligned} \tfrac{\partial^2 l(\beta,\boldsymbol{\eta})}{\partial\beta\partial\boldsymbol{\eta}} &= \textstyle\sum_{i=1}^n\left(-\tfrac{g_i z_i}{\sigma^2}, -\tfrac{(y_i-\beta g_i-\boldsymbol{x}_i\boldsymbol{\gamma})g_i}{\sigma^4}\right) \text{ and} \\ \tfrac{\partial^2 l(\beta,\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}} &= \textstyle\sum_{i=1}^n \left(\begin{array}{cc} -\tfrac{\boldsymbol{x}_i'\boldsymbol{x}_i}{\sigma^2} & -\tfrac{(y_i-\beta g_i-\boldsymbol{x}_i\boldsymbol{\gamma})\boldsymbol{x}_i'}{\sigma^4} \\ -\tfrac{(y_i-\beta g_i-\boldsymbol{x}_i\boldsymbol{\gamma})\boldsymbol{x}_i'}{\sigma^4} & \tfrac{1}{2\sigma^4}-\tfrac{(y_i-\beta g_i-\boldsymbol{x}_i\boldsymbol{\gamma})^2}{\sigma^6} \end{array}\right). \end{aligned}$$

Let the restricted MLE $\hat{\boldsymbol{\eta}}$ be the solution of $\sum_{i=1} U_{\boldsymbol{\eta},i}(\beta,\boldsymbol{\eta}) = 0$, where β is set to 0 in this equation. Specifically, we have

$$\begin{aligned} \hat{\gamma} &= (X'X)^{-1}X'Y, \\ \hat{\sigma}^2 &= \textstyle\sum_{i=1}^n (y_i-\boldsymbol{x}_i\hat{\gamma})^2/n = \|Y-X\hat{\boldsymbol{\gamma}}\|^2/n. \end{aligned}$$

Note that $\hat{\boldsymbol{\eta}}$ is estimated under H_0 and thus does not depend on the genotypes of any given SNP. Therefore it does not change from SNP to SNP, and needs only to be estimated once for each transcript. Also note that all of the off diagonal elements in $\frac{\partial^2 l(\beta,\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}}|_{\beta=0,\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$ equal to zero. Substituting all unknown parameters by their sample estimators in the score function, we get

$$\begin{aligned} \hat{U}_i &= U_{\beta,i}(0,\hat{\boldsymbol{\eta}}) - \Sigma_{\beta,\boldsymbol{\eta}}(0,\hat{\boldsymbol{\eta}})\Sigma_{\boldsymbol{\eta},\boldsymbol{\eta}}(0,\hat{\boldsymbol{\eta}})^{-1}U_{\boldsymbol{\eta},i}(0,\hat{\boldsymbol{\eta}}) \\ &= \tfrac{(y_i-\boldsymbol{x}_i\hat{\gamma})g_i}{\hat{\sigma^2}} - \textstyle\sum g_i\boldsymbol{x}_i(\sum\boldsymbol{x}_i'\boldsymbol{x}_i)^{-1}\tfrac{(y_i-\boldsymbol{x}_i\hat{\gamma})\boldsymbol{x}_i'}{\hat{\sigma^2}} + \tfrac{\sum(y_i-\boldsymbol{x}_i\hat{\gamma})g_i}{n\hat{\sigma^2}} - \tfrac{\sum(y_i-\boldsymbol{x}_i\hat{\gamma})g_i}{n\hat{\sigma^4}}(y_i-\boldsymbol{x}_i\hat{\boldsymbol{\gamma}})^2. \end{aligned}$$

For computational efficiency, we express the score function $\hat{U}_i$ in terms of a matrix operation below. Denote

$$
\begin{aligned}
a &= \textstyle\sum_{i=1}^{n} g_i \boldsymbol{x}_i = G'X, \\
A &= \textstyle\sum_{i=1}^{n_k} \boldsymbol{x}_i' \boldsymbol{x}_i = X'X, \\
c &= \tfrac{\sum (y_i - \boldsymbol{x}_i \hat{\boldsymbol{\gamma}}) g_i}{n\hat{\sigma}^2} = \tfrac{1}{n\hat{\sigma}^2}\left(G'Y - G'X(X'X)^{-1}X'Y\right) = \tfrac{1}{n\hat{\sigma}^2} G'(I - M_x)Y,
\end{aligned}
$$

we have

$$
\hat{U}_i = \frac{y_i - \boldsymbol{x}_i \hat{\boldsymbol{\gamma}}}{\hat{\sigma}^2}\left[(g_i - aA^{-1}\boldsymbol{x}_i') - c(y_i - \boldsymbol{x}_i \hat{\boldsymbol{\gamma}})\right] + c.
$$

Let $\hat{U}$ be the vector of the score function of all individuals. That is,

$$
\begin{aligned}
\hat{U} &= \begin{pmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_n \end{pmatrix} = \begin{pmatrix} \frac{y_1 - \boldsymbol{x}_1 \hat{\boldsymbol{\gamma}}}{\hat{\sigma}^2} \\ \frac{y_2 - \boldsymbol{x}_2 \hat{\boldsymbol{\gamma}}}{\hat{\sigma}^2} \\ \vdots \\ \frac{y_n - \boldsymbol{x}_n \hat{\boldsymbol{\gamma}}}{\hat{\sigma}^2} \end{pmatrix} \times \left[\begin{pmatrix} g_1 - \boldsymbol{x}_1 A'^{-1} a' \\ g_2 - \boldsymbol{x}_2 A'^{-1} a' \\ \vdots \\ g_n - \boldsymbol{x}_n A'^{-1} a' \end{pmatrix} - c \begin{pmatrix} y_1 - \boldsymbol{x}_1 \hat{\boldsymbol{\gamma}} \\ y_2 - \boldsymbol{x}_2 \hat{\boldsymbol{\gamma}} \\ \vdots \\ y_n - \boldsymbol{x}_n \hat{\boldsymbol{\gamma}} \end{pmatrix} \right] + \begin{pmatrix} c \\ c \\ \vdots \\ c \end{pmatrix} \\
&= \tfrac{(I - M_x)Y}{\hat{\sigma}^2} \times \left[(G - XA'^{-1}a') - c(I - M_x)Y)\right] + J_n c.
\end{aligned}
$$

Substituting a, A and c into the above equation, we get

$$
\begin{aligned}
\hat{U} &= \tfrac{(I - M_x)Y}{\hat{\sigma}^2} \times \left[(I - M_x)G - \tfrac{1}{n\hat{\sigma}^2}(I - M_x)YY'(I - M_x)G\right] + \tfrac{1}{n\hat{\sigma}^2} J_n Y'(I - M_x)G \\
&= \left\{\tfrac{(I - M_x)Y}{\hat{\sigma}^2} \times \left[(I - M_x) - \tfrac{1}{n\hat{\sigma}^2}(I - M_x)YY'(I - M_x)\right] + \tfrac{1}{n\hat{\sigma}^2} J_n Y'(I - M_x)\right\} G. \quad (*)
\end{aligned}
$$

Note that the elements inside $\{\ \}$ only depend on Y and the covariates X, and thus are the same across all SNPs. This motivates us to derive the above matrix operation to calculate the score vectors across a large number of SNPs for a given transcript simultaneously. Specifically, we replace the vector G in equation $(*)$ by a matrix $H = (G_1, ..., G_m)$, where G_j is the G vector corresponding to SNP j $(j = 1, ..., m)$. The score matrix $\hat{U}$ then is $n \times m$, where the jth column represents the score vector at SNP

j. Remember that twin pairs are randomly split into two groups. Notationally let's add superscripts to $\hat{U}$ above and denote $\hat{U}^{(1)}$ and $\hat{U}^{(2)}$ as the score matrices of group 1 and group 2, respectively. Also let n_k be the number of individuals in the kth group $(k = 1, 2)$ and the score function of the ith subject for the jth SNP in the kth group be $U_{i,j}^{(k)}$ $(i = 1, ..., n_k)$, or specifically we now have

$$\hat{U}^{(k)} = \begin{pmatrix} \hat{U}_{1,1}^{(k)} & \hat{U}_{1,2}^{(k)} & \cdots & \hat{U}_{1,m}^{(k)} \\ \hat{U}_{2,1}^{(k)} & \hat{U}_{2,2}^{(k)} & \cdots & \hat{U}_{2,m}^{(k)} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{U}_{n_k,1}^{(k)} & \hat{U}_{n_k,2}^{(k)} & \cdots & \hat{U}_{n_k,m}^{(k)} \end{pmatrix}.$$

The regression results for the jth SNP from the two groups may be naively combined as

$$W_{\text{naive}}(j) = \frac{(\sum_{i=1}^{n_1} \hat{U}_{i,j}^{(1)})^2}{\sum_{i=1}^{n_1} (\hat{U}_{i,j}^{(1)})^2} + \frac{(\sum_{i=1}^{n_2} \hat{U}_{i,j}^{(2)})^2}{\sum_{i=1}^{n_2} (\hat{U}_{i,j}^{(2)})^2}$$

and assumed to follow χ_2^2 asymptotically under H_0. This test is only valid when the results from the two groups are independent of each other, which is not likely to be true for twin data. To account for the dependence structure of the two groups, we derive a new score statistic to automatically adjust the correlation between the two groups:

$$W_{\text{proposed}}(j) = \frac{(\sum_{i=1}^{n_1} \hat{U}_{i,j}^{(1)} + \sum_{i=1}^{n_2} \hat{U}_{i,j}^{(2)})^2}{\sum_{i=1}^{n_1} (\hat{U}_{i,j}^{(1)})^2 + \sum_{i=1}^{n_2} (\hat{U}_{i,j}^{(2)})^2 + 2\sum_{i=1}^{n_{twin}} \hat{U}_{i,j}^{(1)} \times \hat{U}_{i,j}^{(2)}},$$

where n_{twin} is the total number of twin pairs, and groups 1 and 2 individuals are arranged in such way that the first n_{twin} samples in groups 1 and 2 are paired twins, and the remaining samples are singletons. The proposed score statistic is assumed to follow χ_1^2 asymptotically under H_0.

2.3 Simulations and Real Data Analysis

2.3.1 Simulation Studies

The proposed method is applied to simulated twin data to evaluate its performance. Each dataset includes 900 or 1800 individuals consisting of MZ twins, DZ twins and singletons in the ratio $2 : 2 : 1$. Two continuous covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ with effects $\boldsymbol{\gamma} = (0.3, 0.1)$ are correlated to the response variable y_i, where x_{i1} and x_{i2} are generated from $N(3,1)$ and $N(5,2)$, respectively. The response variable y is generated from the model [Wang et al., 2011],

$$y_i = \mu + g_i\beta + \boldsymbol{x}_i\boldsymbol{\gamma} + a_i + c_i + d_i + e_i, \quad i = 1, \cdots, n,$$

where for the ith individual, y_i is the gene expression, μ is the grand mean, g_i is the SNP genotype, and $\boldsymbol{x}_i$ is a vector of non-genetic covariates. The random terms in the above model a_i,d_i, c_i, e_i are the additive, dominant, common environment effects and random error, respectively, which are mutually independent and normally distributed with mean zero and variance σ_a^2, σ_c^2, σ_d^2 and σ_e^2, respectively. For subjects i and j who are a twin pair, we have $cov(a_i, a_j) = \sigma_a^2$ and $cov(d_i, d_j) = \sigma_d^2$ if they are MZ pair; $cov(a_i, a_j) = \sigma_a^2/2$ and $cov(d_i, d_j) = \sigma_d^2/4$ if they are DZ pair, while $cov(c_i, c_j) = \sigma_c^2$ for all twin pairs [Falconer and Mackay, 1996]. According to Neale et al. [1989], the above model is referred to as the ACE or ACDE model depending $\sigma_d^2 = 0$ or not. For each simulation set up, 1000 datasets are generated. We set $\sigma_e^2 = 1$, $\sigma_a^2 = 1.5$ and $\sigma_c^2 = 0.75$, resulting in the additive heritability h_a^2 of 0.462 and the variance explained by the shared environmental c^2, of 0.231 under the ACE model. For the ACDE model, we set σ_d^2 to be 1.5, leading to $h_a^2 = 0.316$ and $c^2 = 0.158$. Since there are 1000 association tests for each simulated dataset, 1000 datasets give a total of $1000 * 1000 = 1$ million tests, which are used for the type I error evaluation. The type I errors of the proposed method were compared with three other methods at different significance levels: 1) a multiple linear regression model on the full twin data, where the dependence between the twin pairs is ignored; 2) the naive score approach; and 3) the mixed effects model: $y_{ij} = g_{ij}\beta + x_{ij}\gamma + a_{ij} + d_{ij} + c_{ij} + \epsilon_{ij}$, where i and j are family id and individual index respectively, x_{ij} is the vector of non-genetic covariates plus intercept, g_{ij} is the SNP genotype, and the definition of a_{ij}, c_{ij}, d_{ij} and their covariance structures are described in the above ACE and ACDE models.

Results from Tables 2.1 and 2.2 demonstrate that the type I error rates of both the proposed method and the mixed effects model are well controlled. In contrast, if the linear regression model (lm) is directly applied to the full twin data, the type I error rates have been dramatically inflated. For example, under the settings of $n = 1800$ and the targeted type I error rate $\alpha = 0.05$, the type I errors for the data from the ACE model and ACDE model are 0.098 and 0.102, respectively. The type I error inflation in the naive method is also clear.

For power comparisons among the proposed, naive and mixed effect model approaches, we generated 1000 datasets under the alternative hypothesis, where we set β to $0.32, 0.37, 0.45$, and 0.50 in the ACE model and β to $0.40, 0.45, 0.50$, and 0.55 in the ACDE model. For both of the models, the non-genetic effects $\boldsymbol{\gamma}$ and the variance components σ_a^2, σ_e^2, σ_c^2 and σ_d^2 were kept the same values as those in the above type I error investigation. The results in Tables 2.3 and 2.4 show that the proposed method has negligible power loss compared to the gold standard mixed effects models regardless of the sample size. All simulations and data analyses are conducted in R (a programming language, R Core Team [2014]).

Table 2.1: Type I error comparison for data from the ACE model

	n = 900				n = 1800			
α	0.05	0.01	0.001	0.0001	0.05	0.01	0.001	0.0001
proposed	0.0522	0.0108	0.0011	0.0001	0.0507	0.0100	0.0010	0.0001
naive	0.0590	0.0159	0.0027	0.0005	0.0577	0.0149	0.0023	0.0004
mixed	0.0509	0.0103	0.0011	0.0001	0.0502	0.0100	0.0010	0.0001
lm	0.0988	0.0301	0.0057	0.0011	0.0984	0.0298	0.0053	0.0010

proposed: proposed score test statistic
naive: naive score test statistic
mixed: mixed effects model
lm: multiple linear regression

Table 2.2: Type I error comparison for data from the ACDE model

	n = 900				n = 1800			
α	0.05	0.01	0.001	0.0001	0.05	0.01	0.001	0.0001
proposed	0.0524	0.0107	0.0011	0.0001	0.0510	0.0104	0.0010	0.0001
naive	0.0604	0.0164	0.0028	0.0005	0.0586	0.0160	0.0027	0.0005
mixed	0.0509	0.0103	0.0010	0.0001	0.0502	0.0101	0.0010	0.0001
lm	0.1024	0.0317	0.0061	0.0012	0.1020	0.0317	0.0061	0.0012

proposed: proposed score test statistic
naive: naive score test statistic
mixed: mixed effects model
lm: multiple linear regression

Table 2.3: Power comparison for data from the ACE model

		$\beta = 0.32$			$\beta = 0.37$			$\beta = 0.45$			$\beta = 0.50$		
	Method	proposed	naive	mixed	pro	naive	mixed	pro	naive	mixed	pro	naive	mixed
$n = 900$	$\alpha = 0.05$	0.733	0.721	0.761	0.834	0.834	0.865	0.933	0.934	0.951	0.970	0.964	0.986
	0.01	0.503	0.557	0.557	0.664	0.689	0.710	0.830	0.851	0.882	0.899	0.917	0.933
	0.001	0.225	0.318	0.286	0.363	0.473	0.434	0.629	0.702	0.696	0.745	0.811	0.811
	0.0001	0.103	0.157	0.130	0.166	0.281	0.224	0.368	0.531	0.466	0.530	0.670	0.630
$n = 1800$	$\alpha = 0.05$	0.945	0.944	0.966	0.984	0.982	0.996	1.000	1.000	0.999	1.000	1.000	1.000
	0.01	0.831	0.859	0.889	0.940	0.951	0.963	0.989	0.991	0.999	0.999	1.000	0.999
	0.001	0.592	0.687	0.690	0.784	0.849	0.867	0.953	0.975	0.983	0.988	0.991	0.999
	0.0001	0.344	0.489	0.459	0.567	0.708	0.693	0.864	0.932	0.928	0.951	0.976	0.983

proposed: proposed score test statistic
naive: naive score test statistic
mixed: mixed effects model

Table 2.4: Power comparison for data from the ACDE model

		β = 0.40			β = 0.45			β = 0.50			β = 0.55		
	Method	proposed	naive	mixed	pro	naive	mixed	pro	naive	mixed	pro	naive	mixed
n = 900	α = 0.05	0.740	0.749	0.786	0.842	0.847	0.873	0.904	0.906	0.932	0.949	0.949	0.963
	0.01	0.501	0.554	0.561	0.633	0.683	0.688	0.755	0.793	0.804	0.844	0.877	0.890
	0.001	0.231	0.298	0.268	0.33	0.433	0.395	0.459	0.576	0.538	0.597	0.694	0.672
	0.0001	0.094	0.173	0.122	0.151	0.260	0.201	0.237	0.359	0.299	0.334	0.498	0.426
n = 1800	α = 0.05	0.964	0.961	0.980	0.986	0.985	0.995	0.995	0.996	0.999	0.999	0.999	0.999
	0.01	0.859	0.884	0.913	0.948	0.960	0.971	0.981	0.984	0.990	0.994	0.994	0.998
	0.001	0.664	0.740	0.745	0.800	0.858	0.869	0.895	0.944	0.948	0.969	0.981	0.985
	0.0001	0.431	0.575	0.511	0.605	0.730	0.703	0.756	0.846	0.844	0.865	0.940	0.927

proposed: proposed score test statistic

naive: naive score test statistic

mixed: mixed effects

2.3.2 Computational Efficiency

Our proposed method combines the advantages of multiple linear regression and the matrix operation to achieve fast computational performance. The proposed method took 361 seconds with one 2.93GHz Intel processor to perform 1 million association tests whereas the mixed effects model took $21,000$ seconds using 100 Intel processors simultaneously to run the same number of tests. The R function that implements the proposed method can be found at `www.bios.unc.edu/~feizou/software`.

2.3.3 Resampling

A common issue in most statistical genetics analysis is the multiple comparison problem. Due to the structure of the genome, test statistics performed at different loci are likely to be correlated, thus Bonferroni correction (which assumes independence of the test statistics) might be too conservative. Permutation procedure [Doerge and Churchill, 1996] is thus widely used for assessing genome-wide significance but the procedure is time consuming. The resampling method for assessing genome-wide statistical significance proposed by Zou et al. [2004] is much more computationally efficient than the permutation procedure, because it only requires to maximize the likelihood of the observed data once. Other advantages of this resampling procedure over the permutation procedure include the flexibility to adjust non-genetic covariates and being more robust to the violation of normality assumption for the data. In our simulation, we borrowed the idea from this resampling procedure and modified the proposed test statistic as follows :

$$W^*_{\text{proposed}}(j) = \frac{\sum_{i=1}^{n_1} \hat{U}^{(1)}_{i,j} G_i + \sum_{i=1}^{n_2} \hat{U}^{(2)}_{i,j} G'_i}{\sum_{i=1}^{n_1} (\hat{U}^{(1)}_{i,j})^2 + \sum_{i=1}^{n_2} (\hat{U}^{(2)}_{i,j})^2 + 2 \times \sum_{i=1}^{n_{twin}} \hat{U}^{(1)}_{i,j} \times \hat{U}^{(2)}_{i,j}}$$

where G_i (i=1,...,n) were generated from i.i.d. standard normal distribution. Diao and Vidyashankar [2013] generated their G_i from the Rademacher distribution, where G_i takes value 1 or -1 with equal probability. For each of the two possible distributions of G_i, we performed resampling 10000 times and calculated their type I error rates under different significance levels. The results show that the type I error rates are well controlled, which are 0.0526(0.0523) and 0.01147(0.0109) under the significance level $\alpha = 0.05$ and 0.01 respectively for Rademacher (standard normal) distribution. The power is almost the same, with only a slightly increase, as those from the proposed method W_{proposed}. Table 2.5 summarizes the detailed power comparisons based on the modified test statistic involving resampling procedure for the ACE model with 900 individuals.

Table 2.5: Resampling method power comparison for data generated from the ACE model with different sampling distributions

	Rademacher Distribution				Standard Normal Distribution			
	$\beta = 0.32$	$\beta = 0.37$	$\beta = 0.45$	$\beta = 0.50$	$\beta = 0.32$	$\beta = 0.37$	$\beta = 0.45$	$\beta = 0.50$
$\alpha = 0.05$	0.734	0.834	0.933	0.970	0.733	0.834	0.933	0.969
0.01	0.509	0.671	0.833	0.903	0.503	0.664	0.830	0.899
0.001	0.242	0.383	0.644	0.760	0.229	0.374	0.632	0.745
0.0001	0.115	0.186	0.411	0.574	0.100	0.169	0.374	0.538

2.3.4 Netherlands Twin Registry (NTR) eQTL Study

A total of $2,752$ individuals from the Netherlands Twin Registry (NTR) had their SNP genotypes and gene expression data measured [Wright et al., 2014] on Affymetrix 6.0 SNP arrays and U219 expression arrays. The goals of this project include quantifying the heritability of each transcript and building a detailed list of eQTL in peripheral blood. After quality control [see Wright et al., 2014, for details], $686,895$ SNPs and $47,585$ transcripts on $2,494$ individuals were used for the eQTL analysis. The $2,494$ individuals include 638 MZ pairs, 529 DZ pairs and 160 singletons. A total of 96 covariates are included in the eQTL analysis, for example, plate number, hybridization well position, gender,and 5 principal components (PCs) from the gene expression data, and 3 PCs derived from the SNP data [see Wright et al., 2014].

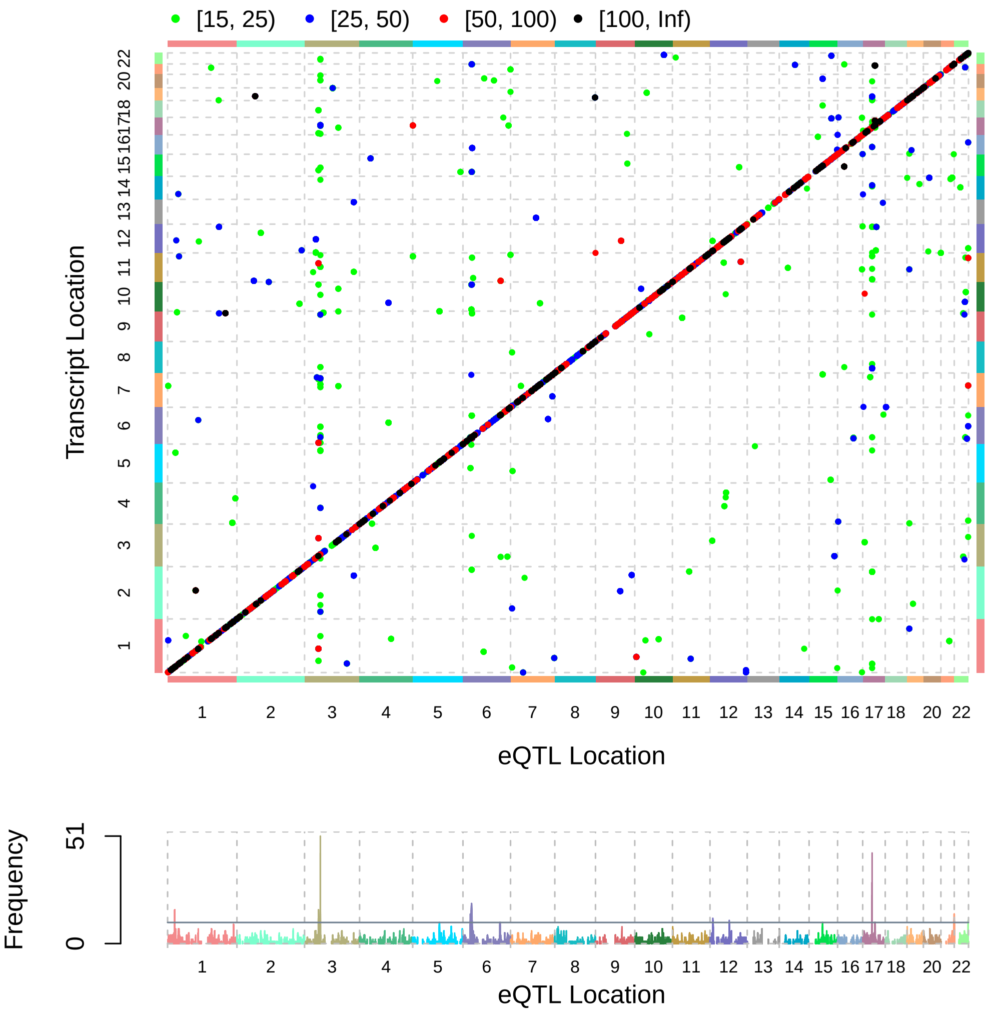
In contrast to controlling the family-wise error, controlling the false discovery rate (FDR) is less stringent and handles the dependence between test statistics reasonably well [Benjamini and Yekutieli, 2001]. Instead of saving the $p-$values for all the SNP-transcript pairs, Shabalin [2012] only saved the $p-$values that pass a pre-specified significance threshold, on which the Benjamini and Hochberg procedure was applied. Following the procedure of Shabalin [2012], we identified $184,474$ and $127,891$ SNP-transcript pairs at $FDR = 0.01$ and $FDR = 0.001$, respectively, with the proposed method. In contrast, the naive method identified $225,419$ and $136,509$ significant SNP-transcript pairs at $FDR = 0.01$ and $FDR = 0.001$, which are significantly more than what have been detected by the proposed method. These additional findings are likely to be false positives, given the fact that the type I error inflation of the naive method. Based on the results from the proposed method, SNP-transcripts pairs with $p-$values passing certain thresholds are shown in the heat map (Figure 2.1). It clearly suggests that there are two master regulator regions on chromosomes 3 and 17 that affect a large number of transcripts, which deserve further investigation.

The top 20 SNP-transcript pairs detected by the proposed approach are summarized in Table 2.6. Note that the p-values of these pairs from the naive score method and Matrix eQTL method are too small to be reported precisely and thus are reported as 0 in the software R.

Table 2.6: Top 20 SNP-transcript pairs identified from the proposed method

	SNP	chr(SNP)	start(SNP)	end(SNP)	GENE	chr(GENE)	start(GENE)	end(GENE)	$log_{10}(pval)$	q value
1	rs3909451	chr5	96295120	96295121	11745601_a_at	chr5	96215266	96225192	228.4077	1.278352e-218
2	rs27300	chr5	96363406	96363407	11745601_a_at	chr5	96215266	96225192	227.2055	8.544353e-218
3	rs12229020	chr12	10117690	10117691	11729479_a_at	chr12	10124007	10138056	227.1056	8.544353e-218
4	rs7313235	chr12	10132274	10132275	11729479_a_at	chr12	10124007	10138056	226.8456	1.166032e-217
5	rs2235918	chr1	17422605	17422606	11727597_at	chr1	17393255	17445948	226.6566	1.441486e-217
6	rs2076607	chr1	17422659	17422660	11727597_at	chr1	17393255	17445948	225.8381	7.908774e-217
7	rs9272346	chr6	32604371	32604372	11750528_x_at	chr6	32410960	32411702	225.5650	1.271430e-216
8	rs4808485	chr19	16439497	16439498	11764269_at	chr19	16438314	16438642	224.0849	3.360241e-215
9	rs27290	chr5	96350093	96350094	11745601_a_at	chr5	96215266	96225192	221.9954	3.670607e-213
10	rs4698634	chr4	17630191	17630192	11736024_at	chr4	17616279	17627249	221.6546	7.240198e-213
11	rs27300	chr5	96363406	96363407	11747137_x_at	chr5	96215265	96253609	221.4142	1.145032e-212
12	rs3909451	chr5	96295120	96295121	11747137_x_at	chr5	96215265	96253609	221.2025	1.708595e-212
13	rs2549794	chr5	96244540	96244541	11747137_x_at	chr5	96215265	96253609	221.0543	2.218743e-212
14	rs12229020	chr12	10117690	10117691	11753241_s_at	chr12	10124195	10137625	220.9615	2.550835e-212
15	rs7673500	chr4	17621378	17621379	11736024_at	chr4	17616279	17627249	220.8615	2.939277e-212
16	rs12229020	chr12	10117690	10117691	11762101_at	chr12	10124033	10136838	220.8420	2.939277e-212
17	rs2549794	chr5	96244540	96244541	11745601_a_at	chr5	96215266	96225192	220.7022	3.816818e-212
18	rs2076608	chr1	17422301	17422302	11727597_at	chr1	17393255	17445948	220.1213	1.373340e-211
19	rs2235914	chr1	17424844	17424845	11727597_at	chr1	17393255	17445948	220.0841	1.417414e-211
20	rs9902260	chr17	62399869	62399870	11757545_x_at	chr17	62399863	62400230	219.5656	4.443247e-211

Figure 2.1: eQTL hotspot for NTR twin data



2.4 Discussion

eQTL analysis usually involves thousands of transcripts and millions of SNPs assayed in thousands of individuals. The challenges for eQTL analysis for twin data are the heavy computational burden and the dependency of twins within pairs. The mixed effects model is the gold standard for analyzing twin data. However, it is extremely time-consuming for modern eQTL data. Although Matrix eQTL is an ultra fast tool, it is not applicable to twin data for eQTL analysis, since the covariance matrix V is assumed to be known. This assumption is not true for twin data, where the covariance matrix can be decomposed into several parts. For example, $V = \sigma_a^2(A + \frac{\sigma_e^2}{\sigma_a^2}I)$ where σ_a^2 and σ_e^2 are unknown. If the heritability is much larger than random error, covariance for the additive genetic effect A can be used to approximate the covariance V, otherwise dropping the second term in the parentheses may result in an invalid inference. Alternatively, the heritability for each transcript needs to be estimated from the mixed effects model before the matrix eQTL is applied, and the computation advantage will be lost dramatically. In contrast to the mixed effects model, we randomly split the twin pairs into two groups and then applied multiple linear regression in each group to calculate the score vector for examining the association of each SNP-transcript pair. The simplicity of linear regression models mitigates the heavy computational burden dramatically. Simulation studies demonstrate the computational advantages of the proposed method. Specifically, the proposed method is almost 6000 times faster than the mixed effect model. The proposed method achieves the fast performance for two reasons: the simplicity of the linear regression model and large matrix operations (especially the matrix multiplication), expressing the most computationally intensive part in an efficient way.

To combine the analytical results from two groups, we propose a novel score statistic which automatically adjusts the correlation between the two groups in a natural way.

Results from simulations show that the proposed method controls type I error rates much better at the target levels than the linear regression model and the naive method. Moreover, this proposed method is competitive and does not lose much power compared to the mixed effects model but is much more computationally efficient. The similar idea based on a score statistic has been applied to meta-analysis of GWAS with overlapping samples, where a robust estimator for variance-covariance matrix was proposed [Lin and Sullivan, 2009]. Although we have focused our attention on twin studies, the proposed method could be applied and extended to multivariate phenotypes, where multiple correlated phenotypes are measured for each subject.

In simulation studies, we have also tried other test statistics such as

$$T_2 = \begin{pmatrix} \sum_{i=1}^{n_1} \hat{U}_{i,j}^{(1)} \\ \sum_{i=1}^{n_2} \hat{U}_{i,j}^{(2)} \end{pmatrix}^T \begin{pmatrix} \sum_{i=1}^{n_1} (\hat{U}_{i,j}^{(1)})^2 & \sum_{i=1}^{n_{twin}} \hat{U}_{i,j}^{(1)} \times \hat{U}_{i,j}^{(2)} \\ \sum_{i=1}^{n_{twin}} \hat{U}_{i,j}^{(1)} \times \hat{U}_{i,j}^{(2)} & \sum_{i=1}^{n_2} (\hat{U}_{i,j}^{(2)})^2 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n_1} \hat{U}_{i,j}^{(1)} \\ \sum_{i=1}^{n_2} \hat{U}_{i,j}^{(2)} \end{pmatrix} \sim \chi_2^2,$$

however, none performs better than the proposed one. In addition, we examined the robustness of the proposed method under two model misspecification scenarios where (1) if the IBD value for DZ pairs is not fixed at 0.5, but is randomly generated from $N(0.50, 0.04)$ to reflect the IBD variation among DZ samples that are commonly observed; (2) the random error e_i is non-normally distributed with a skewed distribution or a distribution with heavier tails (see Table 2.7). Except these changes, the simulation set ups are kept the same as those in Table 2.1 with sample size $n = 900$. Clearly, both the proposed method and the computationally intensive mixed effects model are very robust to the model misspecifications, further suggesting broad applications of the proposed method.

Table 2.7: Type I error for data from misspecified ACE model

	$\rho_{dz} \sim N$		$E \sim Gamma^1$		$E \sim Gamma^2$		$E \sim t^1$		$E \sim t^2$	
	pro	mix	pro	mix	pro	mix	pro	mix	pro	mix
$\alpha = 0.05$	0.0520	0.0509	0.0524	0.0502	0.0524	0.0508	0.0525	0.0519	0.0545	0.0491
$\alpha = 0.01$	0.0107	0.0103	0.0110	0.0111	0.0108	0.0101	0.0110	0.0113	0.0116	0.0118
$\alpha = 0.001$	0.0011	0.0011	0.0011	0.0011	0.0011	0.0010	0.0012	0.0011	0.0013	0.0018
$\alpha = 0.0001$	0.0001	0.0001	0.0001	0.0004	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001

pro: proposed score test statistic

mix: mixed effects

N: Normal(0.501,0.0389)

$Gamma^1$: $Gamma(2, \sqrt{2})$ and $Gamma^2$: $Gamma(6, \sqrt{6})$

t^1: $t_{df=4}$ and t^2: $t_{df=3}$

CHAPTER 3

EMPIRICAL PROCEDURE FOR ASSESSING SIGNIFICANCE OF SNP-SET AND PHENOTYPE ASSOCIATION

3.1 Introduction

Many common diseases and traits are the consequence of the interactions among multiple genetic and environmental factors. GWAS aims to identify the genetic variants associated with complex traits [Bush and Moore, 2012]. With the advent of large GWAS studies, hundreds of genetic variants associated with a variety of traits were identified which have provided invaluable biological insights. The standard approach in GWAS is individual SNP-based analysis where phenotype is regressed on one SNP at a time conditional on other covariates, and the $p-$value for each SNP is compared with a pre-specified threshold. However, this method has limitations due to several reasons, for example, stringent threshold, small to modest effect of a single SNP, causal SNPs not genotyped and etc [Wu et al., 2010].

SNP-set analysis has thus been developed as a promising alternative strategy to bypass the aforementioned issues. A collection of SNPs are combined to be a SNP-set according to biology knowledge with pre-specified criteria and then statistical analysis is leveraged to examine whether a trait is associated with each SNP set rather than with individual SNPs. SNP-set analyses aggregate information from a collection of SNPs and accumulate signals from multiple individual SNPs, thus enhancing the power to capture the true effect of the untyped causal SNP or evidence from interactions if multiple causal SNPs exist and locate very close to each other [Schaid et al., 2002; Wu et al., 2010]. Moreover, the significance threshold become less stringent for SNP-set analyses since the number of tests are decreased dramatically. Thus, SNP-set analysis is valuable for investigating associations between genetic variants and complex traits and has received increasing attention recently.

Several SNP-set based analysis approaches have been well established. Tzeng et al. [2009] proposed a similarity-based regression method, where the model regresses the trait similarity for each pair of independent samples on their haplotype similarity. The significance of the trait-haplotype association is detected by testing the corresponding regression coefficient using a score test which has a weighted χ^2 limiting distribution. Mayhew et al. [2013] proposed a likelihood ratio test (LRT) approach in mixed effects models for phenotype-genotype similarity testing using individual genetic effects as random effects. In his work, a fast algorithm was applied to identify LD blocks to form non-overlapping SNP-sets. He employed a correlation-based genotype similarity instead of a haplotype-based similarity. Testing genotype-phenotype association is equivalent to testing whether the variance components are equal to 0. However, under the set up of Mayhew et al. [2013], the response vector cannot be divided into a collection of independent components under the alternative hypothesis, thus a conventional $50 : 50$ mixture of χ^2_q and χ^2_{q+1} is not appropriate as the null distribution of the LRT [Crainiceanu and Ruppert, 2004]. Mayhew et al. [2013] empirically estimated the mixing proportion, but this empirical process does not work if only a small number of LD blocks are available. In this chapter, we investigate the spectral representation of the LRT and propose an empirical resampling procedure to approximate the finite sample null distribution of the LRT.

3.2 Methods

Considering one LD block with M SNPs, let Y_i $(i = 1, \cdots, n)$ be the continuous trait for the ith individual, and A_{ij} is the $(i,j)_{th}$ element of matrix A and represents the measure of genetic similarity between individuals i and j, using all or partial SNPs in the current set. Mayhew et al. [2013] assumed the $n \times 1$ vector Y followed a multivariate normal distribution with mean $E(Y) = \beta_0$ and covariance matrix $var(Y) = \sigma^2[I(1 - \rho) + \rho A]$. To determine genotype-phenotype associations, he proposed the LRT_1 to test $H_0 : \rho = 0$ VS $H_0 : \rho > 0$. Recognizing the potential signal from an individual SNP, he proposed an extended model including a single SNP from the SNP set as a fixed covariate. For the chosen SNP, denote g_i as the genotype for the ith individual and let $G = (g_1, \cdots, g_n)$. The model becomes $E(Y) = \beta_0 + \beta_1 G$ and $var(Y) = \sigma^2[I(1-\rho)+\rho A]$. For jointly testing trait-genotype similarity and the individual SNP with most significant association with the trait, he proposed the LRT_2 to test $H_0 : \rho = 0$ and $\beta_1 = 0$ VS $H_1 : \rho > 0$ or/and $\beta_1 \neq 0$. He assumed that both tests follow $\pi\chi^2_q + (1 - \pi)\chi^2_{q+1}$ where $q = 0$ for LRT_1 and $q = 1$ for LRT_2. They estimated the mixing proportion π if a large number of LD blocks are tested and majority of them blocks are from null model. Clearly, this empirical procedure will not work for data with only a small number of LD blocks. Crainiceanu and Ruppert [2004] employed a spectral representation of the LRT and developed a resampling procedure to approximate the null distribution of LRT. We implement this algorithm with necessary modifications to derive the null distribution of the LRT and propose an improved version of LRT_2, LRT^*, which is defined as the maximum value of LRT_2 in each LD block and the details are provided later.

We begin by describing the spectral representation of LRT with its resampling algorithm of Crainiceanu and Ruppert [2004]. Consider the linear mixed model with one variance component

$$Y = X\beta + Zb + \varepsilon, \quad E\begin{pmatrix} b \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 0_K \\ 0_n \end{pmatrix}, \quad \mathrm{cov}\begin{pmatrix} b \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2 I_K & 0 \\ 0 & \sigma_\varepsilon I_n \end{pmatrix},$$

where I_K and I_n are two identity matrices, β is a $p-$dimensional vector of fixed effect parameters, b is a $K-$dimensional vector of random effects, and (b, ε) are normally distributed. In our genetic problem setting, Z is the scaled genotype matrix of K SNPs, and the similarity matrix A is determined by Z. The marginal distribution of Y has mean $X\beta$ and covariance matrix $\sigma_\varepsilon V_\lambda$, where $\lambda = \sigma_b^2/\sigma_\varepsilon^2$, $V_\lambda = I_n + \lambda ZZ'$. Note that $\sigma_b^2 = 0$ if and only if $\lambda = 0$. In Mayhew et al. [2013], $\rho = \frac{\lambda}{1+\lambda}$. They proposed to test the null hypothesis:

$$H_0 : \lambda = 0, \text{ and the last } q \text{ elements of } \beta \text{ equal to } 0$$

against the alternative hypothesis:

$$H_A : \lambda > 0 \text{ or at least one } \beta_i \neq 0, i = p - q + 1, \cdots, p.$$

The above hypothesis is the general form for testing one variance component $\sigma_b^2 = 0$ with $q > 0$ constraints on the fixed effects. We only consider $q = 0$ and $q = 1$ cases. In particular, if $q = 0$ then we have the important case corresponding to LRT_1 of Mayhew et al. [2013]. Following the notations in Crainiceanu and Ruppert [2004], the marginal distribution of $Y \sim N(X\beta, \sigma_\varepsilon^2 V_\lambda)$ is proportional to

$$|\sigma_\varepsilon^2 V_\lambda|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Y - X\beta)'(\sigma_\varepsilon^2 V_\lambda)^{-1}(Y - X\beta)\right\}.$$

Treating λ as a known variable, MLEs for β and σ_ε^2 can be expressed as functions of λ

by solving two score equations

$$\begin{aligned}\hat{\beta}(\lambda) &= (X'V_\lambda^{-1}X)^{-1}X'V_\lambda^{-1}Y,\\ \hat{\sigma_\varepsilon^2}(\lambda) &= \tfrac{1}{n}\left[(Y - X\hat{\beta}(\lambda))'V_\lambda^{-1}(Y - X\hat{\beta}(\lambda))\right].\end{aligned}$$

Replacing β and σ by $\hat{\beta}(\lambda)$ and $\hat{\sigma_\varepsilon^2}(\lambda)$ in the log-likelihood function, the profile log-likelihood function of λ is obtained and then LRT is the difference of the profile likelihood under the alternative and null hypothesis:

$$LRT = \sup_{\lambda \geq 0}\left\{-\log|V_\lambda| - n\log(Y'P_\lambda'V_\lambda^{-1}P_\lambda Y)\right\} + n\log(Y'S_1Y),$$

where $S_1 = I_n - X_1(X_1'X_1)^{-1}X_1'$ and $P_\lambda = I_n - X(X'V_\lambda^{-1}X)^{-1}X'V_\lambda^{-1}$. X_1 consists of the first $p-q$ columns of X, given that we are interested in testing one variance component with q fixed effects simultaneously.

Crainiceanu and Ruppert [2004] employed a spectral decomposition of the LRT to approximate the null distribution of LRT as follows:

$$LRT = n\left(1 + \frac{\sum_{s=1}^{q} u_s^2}{\sum_{s=1}^{n-p}\omega_s^2}\right) + \sup_{\lambda \geq 0}\{f(\lambda)\},$$

where

$$\begin{aligned}f(\lambda) &= n\log\{1 + \tfrac{N(\lambda)}{D(\lambda)}\} - \textstyle\sum_{s=1}^{K}\log(1 + \lambda\xi_s),\\ N(\lambda) &= \textstyle\sum_{s=1}^{K}\frac{\lambda\mu_s}{1+\lambda\mu_s}\omega_s^2,\\ D(\lambda) &= \textstyle\sum_{s=1}^{K}\frac{\omega_s^2}{1+\lambda\mu_s} + \sum_{s=K+1}^{n-p}\omega_s^2, \text{ and}\end{aligned}$$

μ_s and ξ_s are the K eigenvalues of the $K \times K$ matrices $Z'P_0Z$ and $Z'Z$ respectively; and two independent vectors are expressed by $\omega = (\omega_1, ..., \omega_{n-p}) = \frac{W^TY}{\sigma_\epsilon}$ and $u = (u_1, ..., u_q) = \frac{U^TY}{\sigma_\epsilon}$. There exist two matrices, $W_{n\times p}$ and $U_{n\times q}$, satisfying the following conditions

[Crainiceanu and Ruppert, 2003; Kuo, 1999; Patterson and Thompson, 1971],

$$\begin{aligned}
W^TW &= I_{n-p}, \\
WW^T &= P_0, \\
W^TV_\lambda W &= \text{diag}\{(1+\lambda\mu_s)\}, \\
Y^TP_\lambda^TV_\lambda^{-1}P_\lambda Y &= Y^TW\text{diag}\{(1+\lambda\mu_s)\}^{-1}W^TY,
\end{aligned}$$

and

$$\begin{aligned}
UU^T &= S_X - S_{X_1}, \\
U^TU &= I_q, \text{ respectively,}
\end{aligned}$$

where $P_0 = I_n - X(X'X)^{-1}X'$, $S_X = X(X'X)^{-1}X'$ and $S_{X_1} = X_1(X_1'X_1)^{-1}X_1'$. Notice that the approximated distribution only depends on the eigenvalues μ_s and ξ_s of the two $K \times K$ matrices $Z'P_0Z$ and $Z'Z$. Once they are calculated, the following algorithm can be used to approximate the null distribution of LRT [Crainiceanu and Ruppert, 2004]: Define a grid of possible values of λ: $0 = \lambda_1 < \lambda_2 < ... < \lambda_m$ and repeat the following steps until a pre-specified number of simulations is complete,

1. Simulate $n-p$ independent χ_1^2 random variables $\omega_1^2, \cdots, \omega_K^2, \omega_{K+1}^2, \cdots, \omega_{n-p}^2$.

2. Independently from the above step, simulate independent random variables $u_1^2, \cdots, u_q^2$ from χ_1^2.

3. Calculate $S_K = \sum_{s=1}^K \omega_s^2$, $X_{n,K,p} = \sum_{s=K+1}^{n-p} \omega_s^2$ and $X_q = \sum_{s=1}^q u_s^2$.

4. For every value of λ_i compute

$$\begin{aligned}
N(\lambda) &= \textstyle\sum_{s=1}^K \frac{\lambda_i\mu_s}{1+\lambda_i\mu_s}\omega_s^2, \\
D(\lambda) &= \textstyle\sum_{s=1}^K \frac{\omega_s^2}{1+\lambda_i\mu_s} + X_{n,K,p}.
\end{aligned}$$

5. Find $\lambda_{\max}$ that maximizes $f(\lambda_i)$ and plug $\lambda_{\max}$ into LRT to output

$$LRT = f(\lambda_{\max}) + n\left(1 + \frac{X_q}{S_K + X_{n,K,p}}\right).$$

The resampling procedure of Crainiceanu and Ruppert [2004] can be directly applied to the LRT_1. However, the procedure needs to be modified for the LRT_2 of Mayhew et al. [2013]. For a given LD block with M SNPs, for the jth SNP, let the fixed covariates be represented by the matrix $X(d_j)$ and the remaining $(M-1)$ SNPs are used to form matrix $Z(d_j)$. That is, for the jth SNP,

$$X(d_j) = \begin{pmatrix} 1 & g_{j,1} \\ 1 & g_{j,2} \\ \vdots & \vdots \\ 1 & g_{j,n} \end{pmatrix} = (J, G_j),$$

and

$$Z(d_j) = \begin{pmatrix} g_{1,1} & \cdots & g_{j-1,1} & g_{j+1,1} & \cdots & g_{M,1} \\ g_{1,2} & \cdots & g_{j-1,2} & g_{j+1,2} & \cdots & g_{M,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ g_{1,n} & \cdots & g_{j-1,n} & g_{j+1,n} & \cdots & g_{M,n} \end{pmatrix} = (G_{(-j)}),$$

where the $g_{j,i}$ is the genotype of the ith individual at the jth SNP. The expanded model becomes

$$Y = X(d_j)(\beta_0, \beta_1)^T + Z(d_j)b + \epsilon.$$

Let the corresponding LRT test for

$$H_0 : \beta_1 = 0 \text{ and } \sigma_b^2 = 0 \quad \text{VS} \quad H_1 : \beta_1 \neq 0 \text{ or } \sigma_b^2 > 0$$

be $LRT(d_j)$. It is thus reasonable to test the overall effect of the LD block as

$$LRT^* = \max_{1 \le j \le M} \{LRT(d_j)\}.$$

However, for computational reasons, Mayhew et al. [2013] only calculated LRT_2 which equals $LRT(d^*)$ where d^* corresponds to the location of the SNP in the block that has the minimal $p-$value from the single SNP analysis. This heuristic procedure is used to approximate LRT^*. Both simulation and real data have shown that LRT_2 approximates LRT^* well. Thus it is natural to use the null distribution of LRT^* to approximate the null distribution of LRT_2. However, for the null distribution of LRT^*, the resampling procedure above no longer applies. For a given SNP j, let's redefine W, U, ω and u as $W(d_j), U(d_j), \omega(d_j)$ and $u(d_j)$ to emphasize the fact that they are related to $X(d_j)$ and $Z(d_j)$.

Note that each element of $\omega(d_j)$ and $u(d_j)$ marginally follows a standard normal distribution. However, for two SNPs located in the same LD block, say at d_j and d_l, $\omega(d_j)$ and $\omega(d_l)$, similarly $u(d_j)$ and $u(d_l)$ are correlated. The correlation structures need to be considered in the resampling procedure, which we describe below. Under the null hypothesis, $cov(Y) = \sigma_\epsilon^2$ and we have

$$\begin{aligned} \text{cov}(\omega(d_l), \omega(d_j)) &= \text{cov}\left(\frac{W(d_l)^T Y}{\sigma_\epsilon}, \frac{W(d_j)^T Y}{\sigma_\epsilon}\right) &= W(d_l)^T W(d_j), \\ \text{cov}(u(d_l), u(d_j)) &= \text{cov}\left(\frac{U(d_l)^T Y}{\sigma_\epsilon}, \frac{U(d_l)^T Y}{\sigma_\epsilon}\right) &= U(d_l)^T U(d_j). \end{aligned}$$

Taking the dependence structure among multiple test statistics into consideration, we separately simulate two long vectors $(\omega(d_1), ..., \omega(d_M))$ and $(u(d_1), ..., u(d_M))$ from two multivariate normal densities, instead of independently simulating the two independent random vectors $\omega(d_j)$ and $u(d_j)$ $(j = 1, ..., M)$. The covariance matrices for the two long vectors are as follows:

$$\begin{pmatrix} \omega_1(d_1) \\ \omega_2(d_1) \\ \vdots \\ \omega_{n-p}(d_m) \end{pmatrix} \sim N\left(0, \begin{bmatrix} W(d_1)^T W(d_1) & W(d_1)^T W(d_2) & \cdots & W(d_1)^T W(d_m) \\ W(d_2)^T W(d_1) & W(d_2)^T W(d_2) & \cdots & W(d_2)^T W(d_m) \\ \vdots & \vdots & \ddots & \vdots \\ W(d_m)^T W(d_1) & W(d_m)^T W(d_2) & \cdots & W(d_m)^T W(d_m) \end{bmatrix}\right),$$

$$\begin{pmatrix} u_1(d_1) \\ u_2(d_1) \\ \vdots \\ u_q(d_m) \end{pmatrix} \sim N\left(0, \begin{bmatrix} U(d_1)^T U(d_1) & U(d_1)^T U(d_2) & \cdots & U(d_1)^T U(d_m) \\ U(d_2)^T U(d_1) & U(d_2)^T U(d_2) & \cdots & U(d_2)^T U(d_m) \\ \vdots & \vdots & \ddots & \vdots \\ U(d_m)^T U(d_1) & U(d_m)^T U(d_2) & \cdots & U(d_m)^T U(d_m) \end{bmatrix}\right).$$

After the two long vectors are completely generated, the corresponding subsets $\omega(d_j)$ and $u(d_j)$ are extracted for SNP j and used in steps of the resampling algorithm. The maximum of the resampling LRT tests across all M SNPs is recorded and formed as one of the realized LRT^* under H_0.

3.3 Simulations and Real Data Analysis

3.3.1 Simulation Studies

Mayhew et al. [2013] developed a efficient algorithm to identify multiple LD blocks as SNP-sets based on HapMap 3 CEU genotype data. We choose the first 10 LD blocks consisting of 10 and 30 SNPs respectively to evaluate the performance of the proposed method. The testing procedure is conducted on each block separately and the results are roughly similar across all blocks.

For each LD block, $100,000$ response vectors are generated from the null model $Y = \alpha + \epsilon$ and we calculate the true maximal likelihood ratios for each dataset. For the same LD block, the proposed algorithm is applied with $100,000$ resamplings. To evaluate the performance of this fast approximation to the null distribution of the LRT_1 and LRT^*, we compare the distribution of the true likelihood ratios with those obtained from the resampling procedure by QQ plot. Moreover, the type I error tables for both LRT_1 and LRT^* are also provided. The QQ plots for LRT_1 indicate that the resampling procedure approximates the true distribution quite well. In addition, Table 3.1 shows that type I error rates for LRT_1 approximation are well controlled. For LRT^*, the empirical distribution is slightly overestimated. However, the QQ plots show that the bias in the LRT^* distribution is almost constant across the range of LRT^* values, which motivates us to employ the following permutation correction procedure for LRT^* [Doerge and Churchill, 1996]. The procedure in details is as follows:

1. Permute y several times (a small number) and calculate the mean of the test statistics from the permuted data, denoted by $\bar{t}_{perm}$.

2. Calculate the mean of the test statistics calculated from the resampling method, denoted by $\bar{t}_{resamp}$.

3. Find the difference between the two means, diff $= \bar{t}_{resamp} - \bar{t}_{perm}$.

4. The adjusted resampling test statistic $=$ the resampling test statistic $-$ diff.

Figures 3.2 and 3.3 compare the QQ plots of the original proposed statistic versus the true test statistics with that of the adjusted proposed versus the true test statistics for the LD block with different sizes. We can see that this permutation adjustment performs well; the type I errors are improved dramatically and much less conservative. For example, at the significance level α = 0.05, the type I error for one block with size 10 changes from 0.0315 to 0.0495 after the permutation adjustment. For another block with size 30, we observe a similar trend. The permutation adjustment corrects the type

I error from 0.0326 to 0.0515. For other significance levels, type I errors for one block from both LRT^* and the adjusted LRT^* are summarized in Table 3.2.

Table 3.1: Type I error of LRT_1 for one block including 10 or 30 SNPs

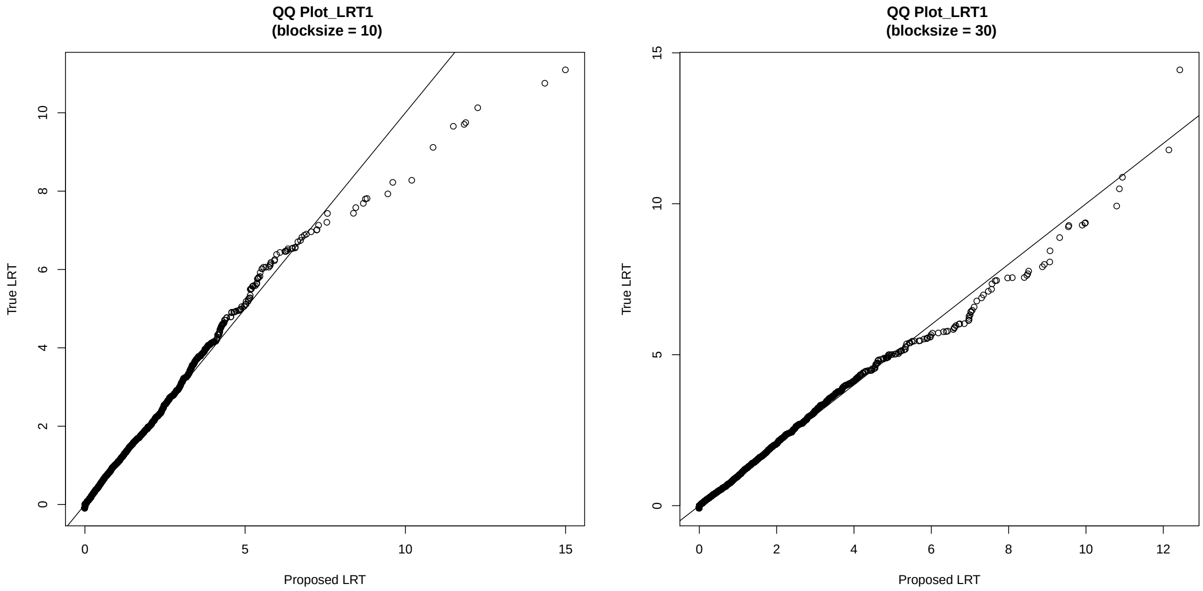
α	blocksize = 10	blocksize = 30
0.05	0.0492	0.0503
0.01	0.0095	0.0100
0.001	0.0013	0.0012

Table 3.2: Type I error of LRT^* for one block including 10 or 30 SNPs

	blocksize = 10		blocksize = 30	
α	unadjusted	adjusted	unadjusted	adjusted
0.05	0.0315	0.0495	0.0326	0.0515
0.01	0.0074	0.0119	0.0057	0.0084
0.001	0.0005	0.0008	0.0007	0.0010

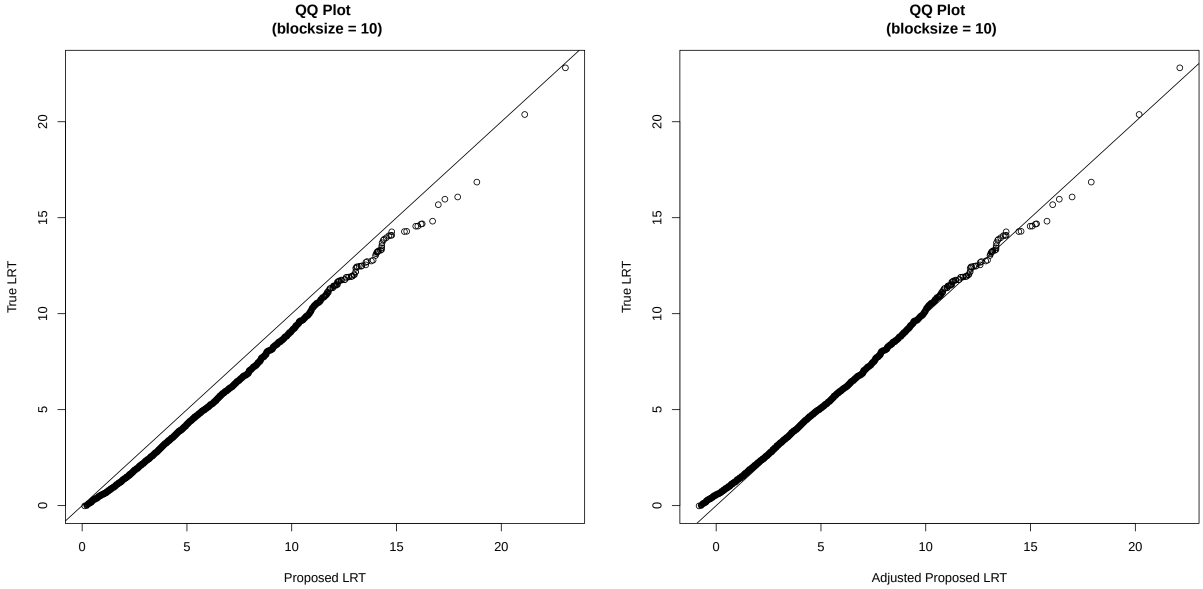
unadjusted: original LRT^*
adjusted: LRT^* with permutation adjustment

Figure 3.1: QQ plot between true LRT_1 VS proposed LRT_1.



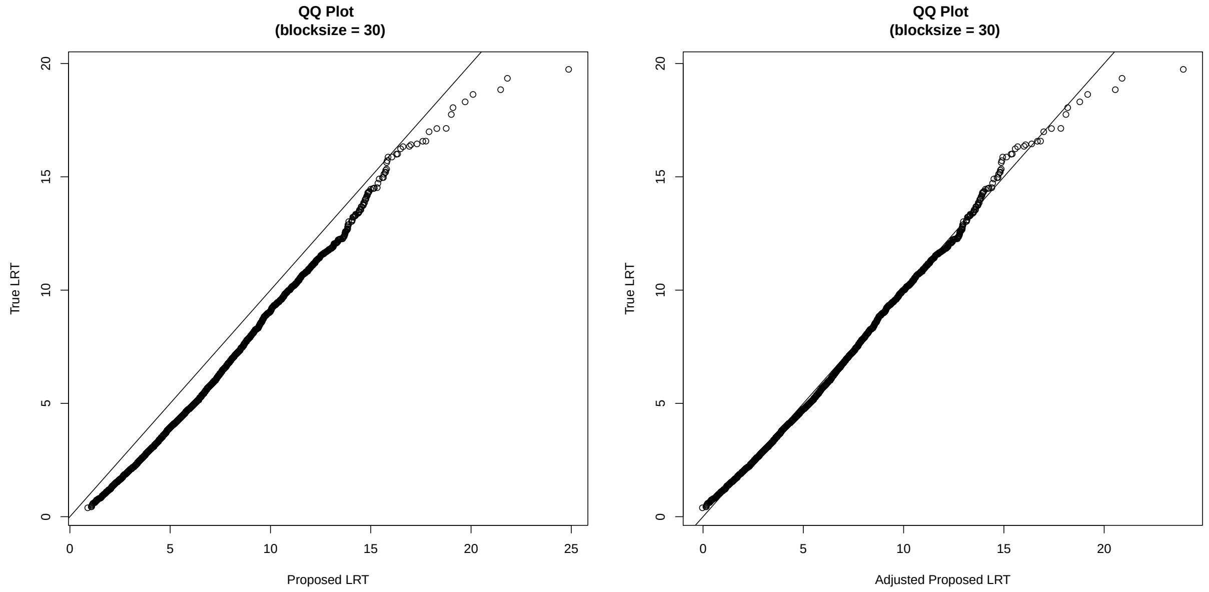
The left graph is for the LD block with size 10;
The right graph is for the LD block with size 30.

Figure 3.2: QQ plot between true LRT^* VS proposed LRT^* for the LD block with size 10.



The proposed LRT^* on the left graph is from the original proposed resampling algorithm; whereas the proposed LRT^* on the right graph is adjusted by permutation adjustment.

Figure 3.3: QQ plot between true LRT^* VS proposed LRT^* for the LD block with size 30.



The proposed LRT^* on the left graph is from the original proposed resampling algorithm; whereas the proposed LRT^* on the right graph is adjusted by permutation adjustment.

3.3.2 Application to CF Data

Our proposed method is applied to real data analysis from the cystic fibrosis genome wide association study of Wright et al. [2011], which includes $n = 1,978$ independent individuals. In this study, the quantitative lung disease was the phenotype. Before the implementation of the proposed resampling procedure, we regress the trait and multiple SNPs on several covariates including gender and genotype PCs and use their residualized values for further SNP-set analysis. Several top blocks are chosen for analyses and the resampling algorithm is used to approximate the null distributions of the LRT_1 and LRT^* with the permutation adjustment and corresponding $p-$values are calculated. Our results are compared to the reports of Mayhew et al. [2013] and Wright et al. [2011]. Although ad hoc, the empirical approach proposed in Mayhew et al. [2013] works reasonably well and their results are empirically justified by our approach.

Table 3.3: Application to CF data - LRT_1

	LRT_1			Mayhew LRT_1	Wright GWAS	
Chr	Begin BP	End BP	$p-$value	$p-$value	Best SNP	$p-$value
11	34617194	34767019	$1.30e-07$	$1.07e-07$	rs10836312	$1.56e-07$
12	91853140	91914388	$3.71e-06$	$5.22e-06$	rs7302601	$6.31e-06$
9	124542690	124625730	$1.02e-05$	$1.65e-05$	rs10818752	$1.09e-02$
11	34769763	35033028	$1.50e-05$	$2.60e-05$	rs12793173	$3.34e-08$
6	32320963	32412009	$5.30e-05$	$7.26e-05$	rs498422	$4.89e-05$
17	75349284	75375359	$3.01e-05$	$1.07e-04$	rs3829574	$5.93e-04$
7	12057085	12178552	$7.10e-05$	$1.20e-04$	rs7789227	$2.18e-05$
4	163550013	163575483	$1.04e-05$	$1.26e-04$	rs2060670	$8.37e-05$
2	184992674	185397916	$1.42e-04$	$1.31e-04$	rs7586860	$1.14e-05$
8	17602799	17629409	$8.00e-05$	$1.52e-04$	rs388769	$3.90e-05$

Top LD block regions with top LRT_1 results from Mayhew et al. [2013]
LRT_1: Results based on the resampling procedure of Crainiceanu and Ruppert [2004]
Mayhew LRT_1: Similarity-based SNP-set analysis results reported by Mayhew et al. [2013]
Wright GWAS: Individual SNP analysis results reported by Wright et al. [2011]

Table 3.4: Application to CF data - LRT^*

	LRT^*			Mayhew LRT_2	Wright GWAS	
Chr	Begin BP	End BP	p–value	p–value	Best SNP	p–value
11	34617194	34767019	$5.02e-08$	$8.63e-08$	rs10836312	$1.56e-07$
11	34769763	35033028	$2.93e-06$	$6.06e-06$	rs12793173	$3.34e-08$
16	60923063	60944750	$3.32e-05$	$3.53e-05$	rs11645366	$1.23e-05$
6	32487484	32696978	$4.41e-05$	$6.53e-05$	rs2516049	$5.78e-06$
11	9572035	9784447	$6.34e-05$	$6.85e-05$	rs93139	$4.01e-06$
12	91853140	91914388	$2.60e-05$	$6.95e-05$	rs7302601	$6.31e-06$
14	69582947	69632776	$6.61e-05$	$7.34e-05$	rs12883884	$1.20e-06$
8	17602799	17629409	$6.52e-05$	$8.75e-05$	rs388769	$3.90e-05$
9	124542690	124625730	$1.24e-05$	$1.28e-04$	rs10818752	$1.09e-02$
3	182966964	182993270	$1.91e-04$	$1.93e-04$	rs10513780	$2.83e-05$

Top LD block regions with top LRT_2 results from Mayhew et al. [2013]
LRT^*: Results based on the proposed resampling procedure with permutation adjustment
Mayhew LRT_2: Similarity-based SNP-set analysis results reported by Mayhew et al. [2013]
Wright GWAS: Individual SNP analysis results reported by Wright et al. [2011]

3.4 Discussion

The use of a SNP-set as a test unit instead of individual SNPs in regression models establishes that the advantage of accumulating evidences from multiple loci to enhance the power to detect genotype-phenotype associations and the improvement would be substantial if more than one causal variants exist in one LD block [Mayhew et al., 2013; Tzeng et al., 2009, 2011; Wu et al., 2010]. Mayhew et al. [2013] proposed two LRT approaches, LRT_1 and LRT_2 under the mixed effects model framework to perform SNP-set analysis in GWAS, where the individual SNP effects are treated as random. LRT_1 is used to test if one variance component is 0, which is equivalent to testing whether genotype similarity is associated with phenotype similarity. LRT_2 jointly tests the variance component and one fixed effect of a single SNP in the block that has the most significant association with the trait. However, the response vector cannot be written as a collection of components independently from each other, thus conventional $50:50$ mixture of χ^2 distributions is no longer appropriate as the null distributions of the LRT_1 and LRT_2 [Crainiceanu and Ruppert, 2004].

Mayhew et al. [2013] empirically estimated the mixing proportions. Their approach assumes that the majority of the LD blocks tested are under the null hypothesis, and the mixing proportions are the same across all the LD blocks. This empirical method is not applicable to the case in which only one LD block is available. Besides, it is unclear if the assumption is valid that the mixing proportions across all the LD blocks are truly constant or not. Based on the work by Crainiceanu and Ruppert [2004], we apply the original algorithm to LRT_1 directly and propose a modified version to estimate the null distributions of the LRT^* for each LD block. For LRT^*, Mayhew et al. [2013] only chose the SNP with the minimal $p-$value as a main effect for computational reasons. Alternatively, we applied the proposed resampling algorithm to LRT^* where each SNP in the LD block is modeled as a fixed covariate. The proposed resampling algorithm is quite computationally efficient because it is essentially based on the eigenvalues of two low dimensional design matrices, which are not repeatedly computed in each iteration of resampling and only need to be calculated once.

The simulations with different block sizes demonstrate that the proposed resampling method for LRT_1 controls type I error rates quite well but is somewhat conservative for LRT^*. The performance of the resampling method for LRT^* is improved greatly after the permutation adjustment. For the CF data, we chose several top blocks to calculate $p-$values for LRT_1 and LRT^* respectively. The results are very similar to those in Mayhew et al. [2013], which empirically confirm the validity of their empirical procedure. For the CF data, for all LD blocks investigated, $LRT^* = LRT_2$, thus the $p-$values estimated from the proposed resampling procedure are valid for LRT_2. For LD blocks where $LRT^* \neq LRT_2$, the resampling $p-$values may be slightly conservative for LRT_2.

CHAPTER 4

INTEGRATIVE ANALYSIS OF SNP AND GENE EXPRESSION DATA IN GWAS

4.1 Introduction

Traditional GWAS have been successful in detecting associations between SNPs and complex traits, where a large number of SNP markers are tested individually with the trait. However, this single SNP analysis may fail to detect true associations due to several reasons, for example, the causal SNPs are not genotyped, and are in poor LD with each genotyped SNP [Wu et al., 2010]. Moreover, the identified SNPs from GWAS typically only explain a small fraction of the heritability of the phenotype [McCarthy and Hirschhorn, 2008]. To overcome these limitations of single SNP analysis, several SNP-set analyses have been proposed to enhance the power to detect the true causal genetic effects by accumulating biological information from a collection of SNPs and aggregating small effects across these multiple markers in association studies [Mayhew et al., 2013; Tzeng et al., 2011; Wu et al., 2010]. Although SNP-set analyses improve the ability to identify genetic risk factors, information from other genomic data is usually ignored and that information may assist our understanding of underlying biological mechanisms of complex diseases [Huang et al., 2014; Nicolae et al., 2010]. Gene expression as an intermediate molecular phenotype can be affected by the SNP genotype [Morley et al., 2004] and also associated with phenotypic variation [Dermitzakis, 2008].

Huang et al. [2014] partitioned the total effect of SNPs on a disease trait into two categories: direct or indirect, depending on whether SNPs influence the trait through the gene expression or not. Specifically, they stated that partial variability of the effect of SNPs on complex traits can be attributed to gene expression if SNPs affect traits in an indirect way via gene expression. Given that both genetic variants and transcripts may play important roles in the development of diseases, it is disadvantageous to leverage information from a single genomic data type in association studies since valuable information from other types of genomic data may be lost [Xiong et al., 2012]. Studies have revealed that genetic variants associated with a trait are more likely to show significant signals in eQTL analysis [Nicolae et al., 2010] and help prioritize the results from GWAS [Hsu et al., 2010]. These findings with the availability of multiple types of genomic data motivate us to analyze associations of a trait jointly with gene expression data and SNP data in a statistical model. Several methods have been developed to jointly model the association of both genetic and gene expression with a trait to improve our understanding of biological mechanisms and networks [Huang et al., 2014; Schadt et al., 2005; Xiong et al., 2012; Zhu et al., 2008].

In this chapter, we propose to add gene expression into a SNP-set analysis under a linear mixed effects model framework. In this model, gene expression is treated as a fixed covariate and SNP genotypes of individuals are treated as random effects. One interesting biological question is to determine whether the effect of gene expression on the complex trait is significant conditional on the SNP-set effects, which is equivalent to performing hypothesis testing for the corresponding regression coefficient. Our model also allows us to detect the joint effects of the SNP-set and gene expression on the complex trait.

A common assumption for estimation in a mixed effects model is the independence between the fixed and random effects. In a single statistical framework, this assumption is likely to be violated when gene expression is related to the SNPs investigated, for example, the SNP-set is eQTL SNPs for the given gene. When this independence assumption is not satisfied and ignored, we will show the gene expression effect can be biasedly estimated. Another linear mixed model is proposed to deal with the correlation between the gene expression and the random SNP effects. The joint distribution of complex trait and gene expression helps relax the independence assumption for a valid effect estimate and inference, and handles the correlation between gene expression and a set of SNPs properly.

Extensive simulation studies are used to evaluate the performance of the proposed method as compared to the naive method. The naive method is solely based on one mixed effect model, where the dependence of gene expression on the SNP-set is ignored. Comparisons in terms of point estimates, type I errors and power for testing the conditional effect of gene expression on disease phenotype illustrate the superiority of the proposed method. A permutation test is used to approximate the null distribution of LRT for testing the joint effect of the SNP-set and gene expression on a disease.

4.2 Methods

The proposed model jointly examines the association of the gene expression and SNP-set with a phenotype. Let Y_i and X_i denote a continuous trait of interest and gene expression respectively, for individual i ($i = 1, ..., n$), and let $G_{K \times n}$ be the scaled genotype matrix of K SNPs across all samples. For simplicity, we start with considering that the complex trait depends on multiple SNP genotypes and a gene expression without other covariates adjustment. If there are a set of covariates that need to be adjusted, the trait, gene expression and SNP genotypes can be residualized separately for them at first. Specifically, our proposed approach assumes that the trait mean is associated with both gene expression and K SNPs in a SNP-set and models the n vector

Y from a multivariate normal distribution as

$$Y = \mu_1 J + X\beta + b_1 + \epsilon_1, \quad b_1 \sim N(0, \sigma_1^2 A), \quad \epsilon_1 \sim N(0, \sigma_{e1}^2 I_n), \tag{4.1}$$

where $Y = (Y_1, ..., Y_n)^T$ is the vector of the continuous phenotype, $X = (X_1, ..., X_n)^T$ is the vector of the gene expression, J is a $n \times 1$ vector with all elements 1s, μ_1 is the grand mean, β is the regression coefficient for the gene expression, b_1 is the random effects from K SNPs whose covariance matrix is the similarity matrix $A_{n\times n} = G^T G$. Specifically, A_{ij} is the $(i,j)th$ element of A, representing the measure of genetic similarity between individuals i and j. In the above model, the choice of multiple SNPs could be either based on eQTL results or their positions with respect to the corresponding gene as a fixed covariate [Huang et al., 2014].

Regarding the association between genetic variants and gene expression [Morley et al., 2004], we next employ a linear mixed effects model to handle the correlation between a set of SNPs and the continuous gene expression

$$X = \mu_2 J + b_2 + \epsilon_2, \quad b_2 \sim N(0, \sigma_2^2 A), \quad \epsilon_2 \sim N(0, \sigma_{e2}^2 I_n), \tag{4.2}$$

where μ_2 is the grand mean of the gene expression. Here, we let $cov(b_1, b_2) = \rho\sigma_1\sigma_2 A$ which induces a correlation between X and b_1.

We are interested in making a valid statistical inference about the regression coefficient for the gene expression β and test the effect of the gene expression on the trait, conditional on all SNPs in the current set. Besides, we will examine the joint effects of the gene expression and SNP-set on the trait. The null hypotheses corresponding to our two goals can be written below respectively

$$H_0 : \beta = 0 \quad VS \quad H_A : \beta \neq 0 \tag{4.3}$$

$$H_0 : \beta = 0, \quad \sigma_1^2 = 0 \quad VS \quad H_A : \beta \neq 0 \quad or \quad \sigma_1^2 \neq 0 \tag{4.4}$$

Instead of using the linear mixed model (4.1) solely to make statistical inference, we consider the joint distribution of (Y, X) where the correlation between X and b_1 is incorporated. The joint distribution of (Y, X) becomes

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 + \mu_2\beta \\ \mu_2 \end{bmatrix}, \Sigma \right),$$

where the covariance matrix Σ is

$$\Sigma = \begin{pmatrix} \beta^2(\sigma_2^2 A + \sigma_{e2}^2) + \sigma_1^2 A + \sigma_{e1}^2 I_n + 2\rho\beta\sigma_1\sigma_2 A & \beta(\sigma_{e2}^2 I_n + \sigma_2^2 A) + \rho\sigma_1\sigma_2 A) \\ \beta(\sigma_{e2}^2 I_n + \sigma_2^2 A) + \rho\sigma_1\sigma_2 A) & \sigma_2^2 A + \sigma_{e2}^2 I_n \end{pmatrix} = M_1 \otimes I_n + M_2 \otimes A,$$

with M_1 and M_2 defined as below,

$$\begin{aligned} M_1 &= \begin{pmatrix} \beta^2\sigma_{e2}^2 + \sigma_{e1}^2 & \beta\sigma_{e2}^2 \\ \beta\sigma_{e2}^2 & \sigma_{e2}^2 \end{pmatrix}, \text{and} \\ M_2 &= \begin{pmatrix} \beta^2\sigma_2^2 + \sigma_1^2 + 2\rho\sigma_1\sigma_2\beta & \beta\sigma_2^2 + \rho\sigma_1\sigma_2 \\ \beta\sigma_2^2 + \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \end{aligned}$$

Note that the dimension of Σ is $2n \times 2n$, thus if the sample size n is relatively large, calculating the inverse and determinant of Σ can be time consuming. However, the Woodbury matrix identity and Sylvester's determinant theorem can be applied here to mitigate the computation burden. By the Kronecker product property, we know that

$$\Sigma = M_1 \otimes I_n + M_2 \otimes G^T G = M_1 \otimes I_n + (M_2 \otimes G^T) I_{2K} (I_2 \otimes G).$$

Thus the inverse and determinant of Σ can be calculated as:

$$\begin{aligned}
\Sigma^{-1} &= (M_1 \otimes I_n)^{-1} - (M_1 \otimes I_n)^{-1}(M_2 \otimes G^T) \\
&\quad [I_{2K} + (I_2 \otimes G)(M_1 \otimes I_n)^{-1}(M_2 \otimes G^T)]^{-1} (I_2 \otimes G)(M_1 \otimes I_n)^{-1}, \\
\det(\Sigma) &= \det(M_1 \otimes I_n) \times \det\left[I_{2K} + (I_2 \otimes G)(M_1 \otimes I_n)^{-1}(M_2 \otimes G^T)\right].
\end{aligned}$$

where $(M_1 \otimes I_n)^{-1} = M_1^{-1} \otimes I_n$. The dimension of the matrix in [ ] of Σ^{-1} is $2K \times 2K$, which is substantially reduced from $2n \times 2n$. Point estimate $\hat{\beta}$ is obtained from the maximal likelihood estimate. Two test statistics are provided for testing (4.3). One is a t test statistic where the standard error of $\hat{\beta}$ is calculated from a hessian matrix, the other is a likelihood ratio test which follows the distribution χ_1^2.

To test the joint effect of the SNP-set and gene expression on the trait Y ($H_0 : \beta = 0$ and $\sigma_1^2 = 0$), the null distribution of LRT does not follow a $50:50$ mixture of χ^2 distribution since the two nuisance parameters ρ and σ_2 disappear under the null hypothesis indicating the existence of non-identifiability issue [Davies, 1977, 1987]. However, the null distribution for (4.4) can be approximated by a permutation procedure where LRT is calculated for each permuted data, depending on the assumption of exchangeability when the null hypothesis in (4.4) holds [Doerge and Churchill, 1996].

4.3 Simulations and Real Data Analysis

4.3.1 Simulation Studies

To evaluate the performance of our proposed method, we conduct extensive simulations under several scenarios. Complex continuous trait, SNP-set and gene expression data are generated in each simulation for this integrative analysis. The SNP-set is the linkage disequilibrium block identified from a fast algorithm developed by Mayhew et al. [2013]. The scaled genotype matrix G including K SNPs are calculated on the

LD block. We assume that the continuous trait depends on both gene expression and SNP-set through model (4.1). The correlation between gene expression and SNP-set is considered through model (4.2). which is induced from the correlation between random effect b_1 in model (4.1) and random effect b_2 in model (4.2). For simplicity, μ_1 and μ_2 are set to 0 and the four variance parameters: σ_1, σ_2, σ_3 and σ_4 are set to 1. The random effects b_1 and b_2 are jointly simulated from a multivariate density with covariance $\rho\sigma_1\sigma_2 G'G$ and variances $\sigma_1^2 G'G, \sigma_2^2 G'G$ respectively. We choose three scenarios of the correlation parameter ρ that controls the magnitude of correlation between gene expression and SNP-set: $\rho = 0.3$ or $\rho = 0.5$ or $\rho = 0.8$. Three different sizes of the LD block are also considered as the scaled SNP genotype matrix G of K SNPs: $K = 10$, or $K = 30$, or $K = 50$. Thus, our simulation studies contain nine scenarios in total for all possible combination values of K and ρ under the null hypothesis (4.3). The number of simulations is 5000 and we let sample size n equal 1000 in each simulation. For our first interest in the effect of gene expression on the trait conditional on the SNP-set effect, we calculate the point estimate, confidence interval and type I error rates at the significance level $\alpha = 0.05$ and $\alpha = 0.01$ and then compare the results among three approaches: 1) the proposed method where a joint distribution of the trait and gene expression is considered from two mixed models (4.1) and (4.2); 2) the naive method based on the mixed model (4.1) only, where the correlation between gene expression and SNP-set is ignored; 3) a multiple linear regression model where only the complex trait and gene expression data are included in the association study. In addition to type I error rates comparisons, we generated 1000 data sets under the alternative hypothesis of (4.3) to perform power comparisons. The regression coefficient of gene expression β is set to $0.05, 0.10, -0.05$ and -0.10. All other parameters in the models (4.1) and (4.2) are kept the same as those in the above type I error investigation. For our secondary interest, 5000 data sets are generated in various scenarios with different block sizes and strengths of correlation between the gene expression and SNP-set under the null hypothesis (4.4). The other parameters and sample size are not changed. The permutation procedure [Doerge and Churchill, 1996] is exploited to approximate the null distribution for testing the joint effects of the SNPs and gene expression on the trait value. The type I error rates are calculated by comparing the true LRTs from 5000 data sets to the significant threshold determined by the permutation test.

With respect to our first interest, a large number of spurious associations are presented if we consider the effect of gene expression on the complex trait without integrating SNP genotypes data into the model. The point estimates of β from the naive method are biased and the proposed method gives unbiased estimates of β. Specifically, the point estimates from the proposed method is much closer to the true value as compared to those from the naive method for all scenarios under the null hypothesis $H_0 : \beta = 0$. The Hessian matrix from the *optim* function in R is used to estimate the standard error of $\hat{\beta}$ for each simulated data set and its mean is very close to the Monte Carlo error $\sqrt{var(\hat{\beta})}$. This implies that the Hessian matrix provides an accurate estimate for the standard error of $\hat{\beta}$ which is used for the calculation of the t test. In addition to the t test, LRT is also performed. The results from both tests are close to each other. Type I errors of the naive model are inflated and such inflation is more severe for larger values of ρ and block size K. For example, type I errors are 0.054, 0.061 and 0.115 for $K = 10$ with $\rho = 0.3$, $K = 10$ with $\rho = 0.8$ and $K = 50$ with $\rho = 0.8$, respectively at the significance level $\alpha = 0.05$. However, the proposed method has well controlled type I errors for all scenarios. This confirms our concern that ignoring the correlation structure may lead to biased parameter estimates. In other words, the superiority of the proposed method over the naive method is more apparent for larger block size K or/and higher strength of correlation ρ. In addition to the type I error inflation, the power comparisons demonstrate that the power loss from the naive method sometimes can be large compared to the proposed method depending on the sign of β. In the events where the naive method has a high power gain, the gain is due to the inflated type I errors. Regarding our second interest in testing $H_0 : \beta = 0$ and $\sigma_1^2 = 0$, we utilize the QQ plot to evaluate the performance of the approximation of the null distribution through a permutation procedure. The QQ plots of the true LRT statistics versus permuted LRT statistics reveal that this approximation performs well for different block sizes with various correlation strengths. For significance levels $\alpha = 0.05$ and $\alpha = 0.01$, type I errors for all scenarios are calculated, where the 95% and 99% percentiles of the permuted LRTs from 5000 simulations are treated as the thresholds. All simulations results are summarized in the Tables 4.1 to 4.8.

Table 4.1: Performance comparison under $H_0 : \beta = 0$ for $K = 10$

		$\hat{\beta}$	$\overline{s.e.}(\hat{\beta})$	T test	*LRT*	T test	*LRT*
α				0.05	0.05	0.01	0.01
$\rho = 0.3$	pro	-0.0011	0.0317	0.0470	0.0472	0.0110	0.0110
	naive	0.0017	0.0316	0.0488	0.0486	0.0110	0.0106
	lm	0.1365	0.0313	0.7946		0.7410	
$\rho = 0.5$	pro	0	0.0317	0.0508	0.0506	0.0118	0.0118
	naive	0.0041	0.0317	0.0540	0.0540	0.0134	0.0134
	lm	0.2274	0.0307	0.8632		0.8172	
$\rho = 0.8$	pro	-0.0006	0.0318	0.0560	0.0552	0.0122	0.0112
	naive	0.0058	0.0318	0.0608	0.0606	0.0138	0.0138
	lm	0.3637	0.0292	0.9682		0.9567	

pro: proposed method
naive: naive method
lm: multiple linear regression model
$\overline{s.e.}(\hat{\beta})$: mean of standard errors of $\hat{\beta}$

Table 4.2: Performance comparison under $H_0 : \beta = 0$ for $K = 30$

		$\hat{\beta}$	$\overline{s.e.}(\hat{\beta})$	T test	LRT	T test	LRT
α				0.05	0.05	0.01	0.01
$\rho = 0.3$	pro	-0.0006	0.0320	0.0494	0.0488	0.0076	0.0076
	naive	0.0059	0.0317	0.0534	0.0530	0.0088	0.0088
	lm	0.1403	0.0315	0.7984		0.7360	
$\rho = 0.5$	pro	0.0004	0.0320	0.0520	0.0516	0.0098	0.0098
	naive	0.0108	0.0317	0.0654	0.0652	0.0134	0.0132
	lm	0.2352	0.0309	0.8880		0.8538	
$\rho = 0.8$	pro	0.0003	0.0320	0.0524	0.0524	0.0118	0.0116
	naive	0.0171	0.0320	0.0858	0.0858	0.0234	0.0232
	lm	0.3772	0.0293	0.9886		0.9834	

pro, naive, lm and $\overline{s.e.}(\hat{\beta})$ have the same definitions as those in Table 4.1

Table 4.3: Performance comparison under $H_0 : \beta = 0$ for $K = 50$

		$\hat{\beta}$	$\overline{s.e.}(\hat{\beta})$	T test	LRT	T test	LRT
α				0.05	0.05	0.01	0.01
$\rho = 0.3$	pro	-0.0005	0.0322	0.0432	0.0434	0.0090	0.0096
	naive	0.0084	0.0317	0.0548	0.0544	0.0108	0.0108
	lm	0.1397	0.0313	0.7936		0.7298	
$\rho = 0.5$	pro	0.0005	0.0322	0.0490	0.0488	0.0086	0.0086
	naive	0.0150	0.0318	0.0774	0.0772	0.0178	0.0172
	lm	0.2335	0.0307	0.9008		0.8630	
$\rho = 0.8$	pro	-0.0005	0.0321	0.0484	0.0478	0.0106	0.0106
	naive	0.0231	0.0321	0.1148	0.1150	0.0336	0.0336
	lm	0.3746	0.0292	0.9924		0.9854	

pro, naive, lm and $\overline{s.e.}(\hat{\beta})$ have the same definitions as those in Table 4.1

Table 4.4: Power comparison for block size $K = 10$

		$\beta = 0.05$			$\beta = 0.10$			$\beta = -0.05$			$\beta = -0.10$		
		pro	naive	lm	pro	naive	lm	pro	naive	lm	pro	naive	lm
$\rho = 0.3$	$\alpha = 0.05$	0.360	0.389	0.836	0.858	0.880	0.879	0.358	0.323	0.776	0.895	0.882	0.757
	$\alpha = 0.01$	0.165	0.180	0.784	0.711	0.737	0.838	0.161	0.142	0.708	0.728	0.693	0.684
$\rho = 0.5$	$\alpha = 0.05$	0.363	0.405	0.904	0.880	0.908	0.937	0.356	0.306	0.828	0.872	0.849	0.788
	$\alpha = 0.01$	0.167	0.195	0.873	0.700	0.745	0.916	0.159	0.133	0.771	0.710	0.670	0.729
$\rho = 0.8$	$\alpha = 0.05$	0.365	0.433	0.986	0.879	0.905	0.994	0.355	0.284	0.947	0.887	0.847	0.901
	$\alpha = 0.01$	0.165	0.208	0.980	0.688	0.757	0.991	0.158	0.125	0.929	0.724	0.650	0.871

pro, naive, and lm have the same definitions as those in Table 4.1

Table 4.5: Power comparison for block size $K = 30$

		$\beta = 0.05$			$\beta = 0.10$			$\beta = -0.05$			$\beta = -0.10$		
		pro	naive	lm	pro	naive	lm	pro	naive	lm	pro	naive	lm
$\rho = 0.3$	$\alpha = 0.05$	0.347	0.419	0.837	0.859	0.904	0.894	0.347	0.278	0.748	0.873	0.837	0.730
	$\alpha = 0.01$	0.157	0.227	0.790	0.680	0.757	0.855	0.158	0.117	0.671	0.704	0.651	0.645
$\rho = 0.5$	$\alpha = 0.05$	0.350	0.469	0.930	0.882	0.933	0.964	0.345	0.237	0.840	0.884	0.811	0.773
	$\alpha = 0.01$	0.157	0.260	0.903	0.703	0.814	0.949	0.155	0.096	0.790	0.735	0.617	0.705
$\rho = 0.8$	$\alpha = 0.05$	0.343	0.535	0.997	0.878	0.950	0.999	0.385	0.208	0.974	0.876	0.743	0.929
	$\alpha = 0.01$	0.149	0.317	0.995	0.711	0.851	0.999	0.171	0.074	0.959	0.711	0.506	0.906

pro, naive, and lm have the same definitions as those in Table 4.1

Table 4.6: Power comparison for block size $K = 50$

		$\beta = 0.05$			$\beta = 0.10$			$\beta = -0.05$			$\beta = -0.10$		
		pro	naive	lm	pro	naive	lm	pro	naive	lm	pro	naive	lm
$\rho = 0.3$	$\alpha = 0.05$	0.322	0.441	0.850	0.869	0.926	0.903	0.362	0.260	0.736	0.867	0.826	0.712
	$\alpha = 0.01$	0.142	0.225	0.799	0.716	0.811	0.866	0.146	0.078	0.659	0.704	0.629	0.625
$\rho = 0.5$	$\alpha = 0.05$	0.323	0.513	0.943	0.870	0.935	0.965	0.362	0.194	0.844	0.877	0.784	0.778
	$\alpha = 0.01$	0.145	0.278	0.919	0.703	0.841	0.953	0.145	0.054	0.788	0.723	0.556	0.716
$\rho = 0.8$	$\alpha = 0.05$	0.322	0.609	0.999	0.875	0.967	0.999	0.362	0.119	0.980	0.870	0.648	0.950
	$\alpha = 0.01$	0.144	0.371	0.997	0.711	0.898	0.999	0.142	0.030	0.969	0.692	0.420	0.932

pro, naive, and lm have the same definitions as those in Table 4.1

Table 4.7: Type I error rates from permutation test at $\alpha = 0.05$

	K = 10	K = 30	K=50
$\rho = 0.3$	0.046	0.047	0.043
$\rho = 0.5$	0.047	0.055	0.045
$\rho = 0.8$	0.045	0.044	0.044

Table 4.8: Type I error rates from permutation test at $\alpha = 0.01$

	K = 10	K = 30	K=50
$\rho = 0.3$	0.010	0.006	0.011
$\rho = 0.5$	0.011	0.010	0.010
$\rho = 0.8$	0.012	0.008	0.010

Figure 4.1: QQ plot between true LRT VS permuted LRT for $K = 30$ and $\rho = 0.5$

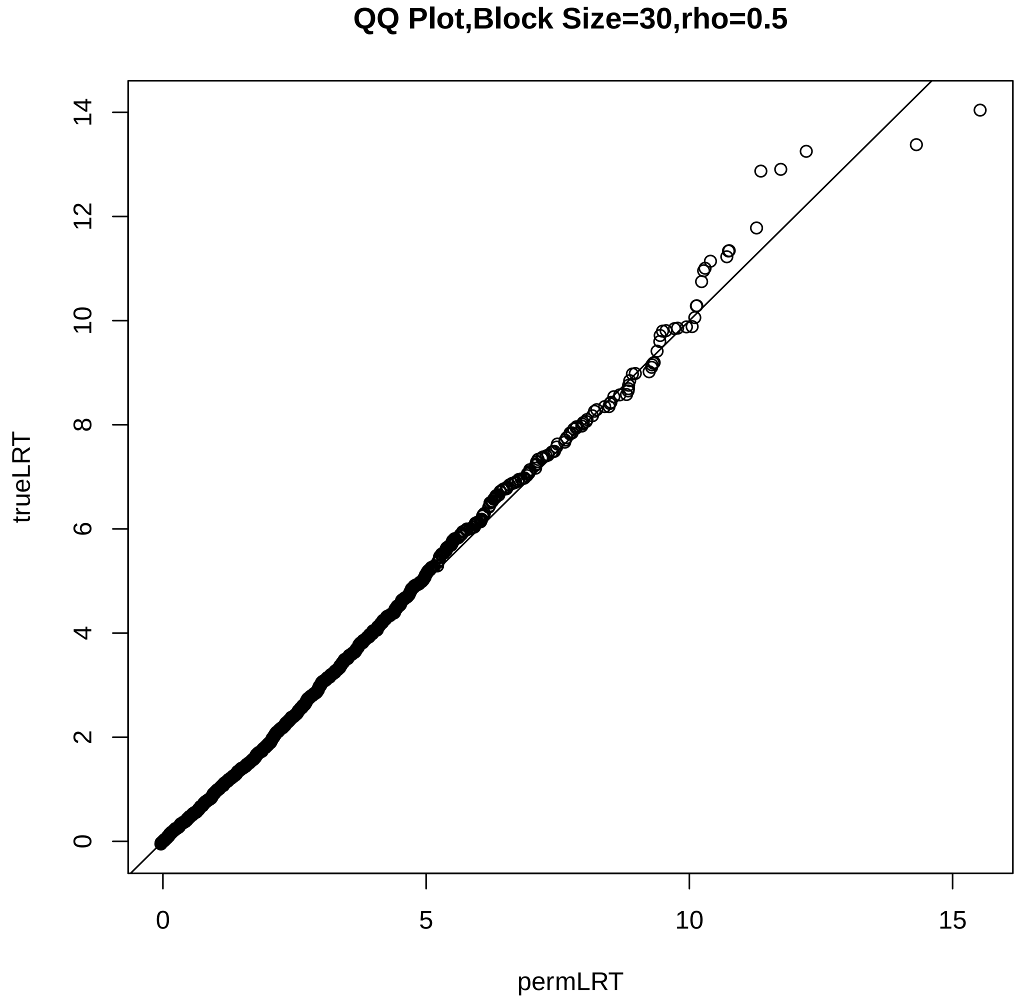
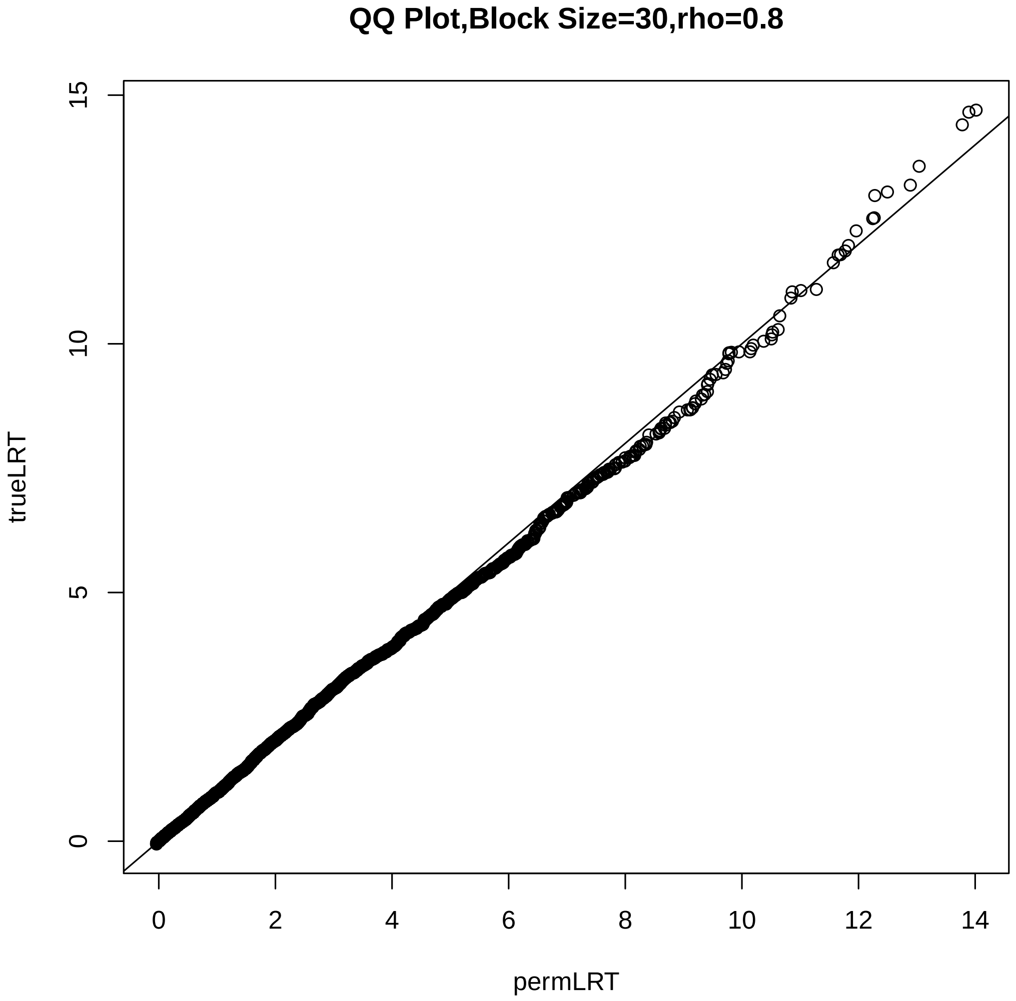


Figure 4.2: QQ plot between true LRT VS permuted LRT for $K = 30$ and $\rho = 0.8$



4.3.2 Application to CF data

The cystic fibrosis genome wide association study of Wright et al. [2011] established the evidence of significant variants near EHF and APIP located at chr11p13 as modifier loci of quantitative lung function trait in cystic fibrosis (CF). In addition, HLA class II genes were investigated and the results suggested they were associated with the phenotype. Besides GWAS genotyped SNP data, gene expression of 12 HLA genes is available for 754 independent CF subjects which allows us to model the association of the phenotype with both genomic type data. We regress the phenotype, gene expression and SNPs within each gene on 19 covariates, including sex and genotype PCs to obtain their residualized values respectively for further analyses. The quantitative trait is assumed to depend on both gene expression and the corresponding SNPs within that gene, as described in model (4.1). We investigate whether the effect of gene expression of each HLA gene is associated with the quantitative trait, conditional on the effect of multiple SNPs within the gene. The proposed method and naive method are both applied to analyze the true data. The results are summarized in Tables 4.9 and 4.10 and they only show slight differences between those two methods, which is due to the fact that the correlation between the gene expression and SNP-set is small. In this situation, the naive method and the proposed method are expected to perform similarly.

Table 4.9: Application of the proposed method to HLA data

Gene Name	Block Size	Point Estimate	p-value
$HLA-B$	80	0.2606	0.4760
$HLA-DMA$	15	0.2265	0.1023
$HLA-DMB$	9	-0.0414	0.7261
$HLA-DOA$	44	0.1085	0.4108
$HLA-DOB$	32	0.0873	0.3967
$HLA-DPA1$	13	0.2007	0.3151
$HLA-DPB1$	11	0.0408	0.5260
$HLA-DPB2$	25	0.0722	0.4230
$HLA-DRA$	29	0.2276	0.1939
$HLA-DRB1$	96	0.0146	0.7049
$HLA-E$	23	0.4505	0.0030
$HLA-F$	196	0.3165	0.0152

Table 4.10: Application of the naive method to HLA data

Gene Name	Block Size	Point Estimate	p-value
$HLA-B$	80	0.2607	0.5300
$HLA-DMA$	15	0.2662	0.0582
$HLA-DMB$	9	-0.0413	0.7269
$HLA-DOA$	44	0.1204	0.3787
$HLA-DOB$	32	0.0876	0.3944
$HLA-DPA1$	13	0.2555	0.1628
$HLA-DPB1$	11	0.1012	0.3395
$HLA-DPB2$	25	0.0722	0.4049
$HLA-DRA$	29	0.2561	0.1375
$HLA-DRB1$	96	0.0150	0.5534
$HLA-E$	23	0.4507	0.0030
$HLA-F$	196	0.3165	0.0097

4.4 Discussion

In this chapter, we integrate gene expression information as a fixed covariate into SNP-set analyses, where genetic effects of individuals are treated as random effects. Our research interest is to make a valid statistical inference on the effect of gene expression on the complex trait, conditional on a given SNP-set. When a traditional mixed effects model is proposed, the independence assumption between the fixed effect of a gene expression and the random effects from multiple SNPs is likely to be violated. Our extensive simulations with different block sizes and various correlation strengths have shown that ignoring the correlation structure would lead to a biased parameter estimate for the gene expression effect. Moreover, the type I error rates are highly inflated if this correlation is strong or the number of SNPs is relatively large in the given set. We propose an improved mixed effects model where the correlation between gene expression and SNP-set is incorporated to relax the independence assumption. The proposed method has well controlled type I error and makes a valid statistical inference and parameter estimation. The power advantage of the proposed method depends on the direction of the effect of gene expression, but the power gain for the naive method is largely due to the inflated type I errors. To test the joint effects of the SNP-set and gene expression, a permutation procedure is performed to approximate the null distribution of LRT. The results of type I errors and QQ plots under various scenarios illustrate that this approximation works well.

BIBLIOGRAPHY

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.

Baz, K., Emin, E. M., Yazici, A. C., and et al. (2008). Association between tumor necrosis factor-alpha gene promoter polymorphism at position -308 and acne in Turkish patients. *Archives for Dermatolological Research*, 300(7):371–376.

Beecham, G. W., Martin, E. R., Li, Y. J., and et al. (2009). Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. *American Journal of Human Genetics*, 84(1):35–43.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

Benjamini, Y. and Yekutieli, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics*, 171(2):783–790.

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., and et al. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, 76(2):306–317.

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., and et al. (2012). *OpenMx 1.2 User Guide.*

Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics*, 3(11):872–882.

Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S. A., Tiret, L., and Richardson, S. (2011). Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189(4):1449–1459.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822.

Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R., and Dwyer, T. (2005). Regression models for twin studies: a critical review. *International Journal of Epidemiology*, 34(5):1089–1099.

Chipman, K. C. and Singh, A. K. (2011). Bayesian detection of expression quantitative trait loci hot spots. *BMC Bioinformatics*, 12(7):doi: 10.1186/1471–2105–12–7.

Chou, Y., Lepore, N., Chiang, M. C., Avedissian, C., Barysheva, M., and et al. (2009). Mapping genetic influences on ventricular structure in twins. *NeuroImage*, 44(4):1312–1323.

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

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Crainiceanu, C. M. and Ruppert, D. (2003). Proofs of theorems for the paper "likelihood ratio tests in linear mixed models with one variance component". *Technical Report TR1389. Department of Statistics, Cornell University, Ithaca.*

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185.

Crainiceanu, C. M., Ruppert, D., and Vogelsang, T. J. (2003). Some properties of likelihood ratio tests in linear mixed models. *Available at www. orie. cornell. edu/ ~davidr/ papers.*

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(1):247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1):33–43.

Dermitzakis, E. T. (2008). From gene expression to disease risk. *Nature Genetics*, 40(5):492–493.

Diao, G. and Vidyashankar, A. N. (2013). Assessing genome-wide statistical significance for large p small n problems. *Genetics*, 194(3):781–783.

Doerge, R. W. and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1):285–294.

Dudbridge, F. (2006). A note on permutation tests in multi-stage association scans. *American Journal of Human Genetics*, 78(6):1094–1095.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Fourth edition. Addison Wesley Longman, Harlow, Essex, UK.

Feng, R., Zhou, G., Zhang, M., and Zhang, H. (2009). Analysis of twin data using SAS. *Biometrics*, 65(2):584–589.

Garcia-Closas, M., Couch, F. J., Lindstrom, S., Michailidou, K., Schmidt, M. K., and et al. (2013). Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nature Genetics*, 45(4):392–398.

Ghazalpour, A., Doss, S., Kang, H., and et al. (2008). High-resolution mapping of gene expression using association in an outbred mouse stock. *PLoS Genetics*, 4(8):e1000149.

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, 24(8):408–415.

Han, B., Kang, H. M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, 5(4):e1000456.

Hanson, R. L., Muller, Y. L., Kobes, S., and et al. (2014). A genome-wide association study in American Indians implicates DNER as a susceptibility locus for type 2 diabetes. *Diabetes*, 63(1):369–376.

Hernandez, D. G., Nalls, M. A., Moore, M., and et al. (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of Disease*, 47(1):20–28.

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231.

Hoggart, C. J., Whittaker, J. C., DeIorio, M., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7):e1000130.

Hsu, Y. H., Zillikens, M. C., Wilson, S. G., Farber, C. R., Demissie, S., and et al. (2010). An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits. *PLoS Genetics*, 6(6):e1000977.

Huang, Y. T., Vanderweele, T. J., and Lin, X. (2014). Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Annual of Applied Statistics*, 8(1):352–376.

Jia, P., Wang, L., Meltzer, H. Y., and Zhao, Z. (2010). Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. *Schizophrenia Research*, 122(1-3):38–42.

Jorsekog, K. G. and Sorborn, D. (1986). *Lisrel VI.* Scientific Software, Mooresville, Indiana.

Kang, H. M., Ye, C., and Eskin, E. (2008). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925.

Kendziorski, C. and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome*, 17(6):509–517.

Klein, R. J., Zeiss, C., Chew, E. Y., and et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

Kraft, P. and Cox, D. G. (2008). Study designs for genome-wide association studies. *Advances in Genetics*, 60(6):465–504.

Kuna, S. T., Maislin, G., Pack, F. M., Staley, B., Hachadoorian, R., Coccaro, E. F., and Pack, A. I. (2012). Heritability of performance deficit accumulation during acute sleep deprivation in twins. *Sleep*, 35(9):1223–1233.

Kuo, B. S. (1999). Asymptotics of ML estimator for regression models with a stochastic trend component. *Econometric Theory*, 15(1):24–49.

Laird, N. M. and Lange, C. (2011). *The fundamentals of modern statistical genetics*. Springer, New York.

Lettre, G., Lange, C., and Hirschhorn, J. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, 31(4):358–362.

Lewis, C. (2002). Genetic association studies: design, analysis and interpretation. *Brief Bioinform*, 3(2):146–153.

Lewontin, R. C. (1964). The interaction of selection and linkage.I.general considerations; heterotic models. *Genetics*, 49(1):49–67.

Lewontin, R. C. and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evoluation*, 14(4):458–472.

Li, S., Sanna, S., Maschio, A., and et al. (2007). The GLUT9 gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. *PLoS Genetics*, 3(11):e194.

Lin, D. Y. and Sullivan, P. F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. *The American Journal of Human Genetic*, 85(6):862–872.

Loos, R. J., Lindgren, C. M., Li, S., and et al. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics*, 40(6):768–775.

Mayhew, G., Zou, F., Knowles, M. R., Durie, P. R., and Wright, F. A. (2013). An extended similarity-based snp-set approach for genome-wide association studies. *Submitted to American Journal of Human Genetics*, Available at www.oatd.org.

McCarthy, M. I. and Hirschhorn, J. N. (2008). Genome-wide association studies: potential next steps on a genetic journey. *Human Molecular Genetics*, 17(2):R156–R165.

Michaelson, J. J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, 48(3):265–276.

Min, J. L., Taylor, J. M., Richards, J. B., Watts, T., Pettersson, F. H., and et al. (2011). The use of genome-wide eqtl associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. *PLoS One.*, 6(7):e22070.

Morley, M., Molony, C. M., Weber, T. M., and et al. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747.

Muthen, L. K. and Muthen, B. O. (1998). *Mplus User's Guide.* Muthen & Muthen, Los Angeles, California.

Neale, M. C., Boker, S. M., Xie, G., and Maes, H. H. (1999). *Mx:Statistical Modeling.* Department of Psychiatry, Medical College of Virginia of Virginia Commonwealth University, Richmond, Virginia.

Neale, M. C. and Cardon, L. R. (1992). *Methodology for genetic studies of twins and families.* Kluwer Academic, Dordrecht, the Netherlands.

Neale, M. C., Heath, A. C., Hewitt, J. K., Eaves, L. J., and Fulker, D. W. (1989). Fitting genetic models with LISREL: Hypothesis testing. *Behavior Genetics*, 19(1):37–49.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., and Beazley, C. (2010). Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLoS Genetics*, 6:p.e.1000895.

Nicodemus, K. K., Liu, W., Chase, G. A., Tsai, Y. Y., and Fallin, M. D. (2005). Comparison of type I error for multiple test corrections in large SNP studies using principal components versus haplotype blocking algorithms. *BMC Genetics*, 6(Suppl 1):S78.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*, 6(4):e1000888.

Park, J. H., Song, Y. M., Sung, J., and et al. (2012). The association between fat and lean mass and bone mineral density: the Healthy Twin Study. *Bone*, 50(4):1006–1011.

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

Pearson, K. (1909). On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) A given intensity is recorded for each grade of A. *Biometrika*, 7(1-2):96–105.

Pearson, K. (1910). On a new method of determining correlation, when one variable is given by alternative and the other by multiple categories. *Biometrika*, 7(3):248–257.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus.* Springer-Verlag, New York.

Pirinen, M., Donnelly, P., and Spencer, C. C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, 44(8):848–851.

Purcell, S., Neale, B., Todd-Brown, K., and et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81(3):559–575.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S., Skrondal, A., and Gjessing, H. K. (2008). Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*, 64(1):280–288.

Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, 10:117–133.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., and et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Human Genetics*, 70(2):425–434.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358.

Sidak, Z. K. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Silventoinen, K., Sammalisto, S., Perola, M., and et al. (2003). Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research*, 6(5):399–408.

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 Computational Biology*, 6(5):e1000770.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):1171–1177.

Tenesa, A., Farrington, S. M., Prendergast, J., and et al. (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature Genetics*, 40(5):631–637.

Tomlinson, I., Webb, E., Carvajal-Carmonaet, L., and et al. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics*, 39(8):984–988.

True, W. R., Rice, J., Eisen, S. A., Heath, A. C., Goldberg, J., Lyons, M. J., and Nowak, J. (1993). A twin study of genetic and environmental contributions to liability for posttraumatic stress symptoms. *Arch Gen Psychiatry*, 50(4):257–264.

Tzeng, J. Y. and Zhang, D. (2007). Haplotype-based association analysis via variance-components score test. *American Journal of Human Genetics*, 81(5):927–938.

Tzeng, J. Y., Zhang, D., Chang, S. M., Thomas, D. C., and Davidian, M. (2009). Gene-trait similarity regression for multimarker-based association analysis. *Biometrics*, 65(3):822–832.

Tzeng, J. Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M. I., Sale, M. M., Worrall, B. B., Hsu, F. C., Thomas, D. C., and Sullivan, P. F. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *American Journal of Human Genetics*, 89(2):277–288.

Vaccarino, V., Brennan, M. L., Miller, A. H., and et al. (2008). Association of major depressive disorder with serum myeloperoxidase and other markers of inflammation: a twin study. *Biological Psychiatry*, 64(6):476–483.

Wang, X., Guo, X., He, M., and Zhang, H. (2011). Statistical inference in mixed models and analysis of twin and family data. *Biometrics*, 67(3):987–995.

Weng, L., Macciardi, F., Subramanian, A., and et el. (2011). SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, 12:99.

Winham, S. J., Cuellar-Barboza, A. B., Oliveros, A., McElroy, S. L., Crow, S., Colby, C., Choi, D.-S., Chauhan, M., Frye, M., and Biernacka, J. M. (2013). Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. *Molecular Psychiatry*, doi:10.1038/mp.2013.159.

Wright, F. A., Shabalin, A. A., and Rusyn, I. (2012). Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics*, 13(3):343–352.

Wright, F. A., Strug, L. J., Doshi, V. K., and et el. (2011). Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. *Nature Genetics*, 43(6):539–546.

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., and et al. (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, 46(5):430–437.

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. *American Journal of Human Genetics*, 86(6):929–942.

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

Xiong, Q., Ancona, N., Hauser, E. R., Mukherjee, S., and Furey, T. S. (2012). Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Research*, 22(2):386–397.

Yang, T. P., Beazley, C., Montgomery, S. B., and et al. (2010). Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, 26(19):2474–2476.

Yu, X. Q., Li, M., Zhang, H., and et al. (2012). A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. *Nature Genetics*, 44(2):178–182.

Zhen, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457–1468.

Zhu, J., Zhang, B., Smith, E. N., and et al. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40(7):854–861.

Zou, F., Fine, J. P., Hu, J., and Lin, D. Y. (2004). An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics*, 168(4):2307–1216.

Zou, W., Aylor, D. L., and Zeng, Z. (2007). eQTL Viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics*, 8(7):doi:10.1186/1471–2105–8–7.