UNIVERSITY OF CALIFORNIA
RIVERSIDE

# A Randomized Fixed Model Methodology for Genome-Wide Association Studies

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Plant Biology

by

Tiantian Zhu

December 2017

Dissertation Committee:
Dr. Shizhong Xu, Chairperson
Dr. Thomas Girke
Dr. Arthur Jia

The Dissertation of Tiantian Zhu is approved:

---

---

---

Committee Chairperson

University of California, Riverside

Acknowledgements

I wish to express my sincere gratitude to my advisor, Dr. Shizhong Xu, for guiding me through my PhD study. Dr. Xu is so knowledgeable that he always gives me insightful advice in my research. I also thank Dr. Thomas Girke and Dr. Arthur Jia for serving in my dissertation committee and for their valuable suggestions on my dissertation.

I want to thank the Lord for He is my shepherd all the days of my life. I am also thankful to the saints in the Church, especially brother Lin, sister Lin, brother Lii, sister Lii, and all other members for their cherishing and nourishing.

Lastly, I sincerely thank my parents, siblings, and other family members for their unconditional love!

Dedicated to my beloved parents, siblings, and my God

ABSTRACT OF THE DISSERTATION

A Randomized Fixed Model Methodology for Genome-Wide Association Studies

by

Tiantian Zhu

Doctor of Philosophy, Graduate Program in Plant Biology
University of California, Riverside, December 2017
Dr. Shizhong Xu, Chairperson

Genome-wide association studies (GWAS) are statistical tools widely used to identify the associations between genetic variants and a quantitative trait. Through GWAS, the genetic architectures of many complex traits in plants, animals and human have been revealed. A commonly used method in GWAS is the linear mixed model (LMM). This model is called the fixed model (FM) approach when the marker effect is treated as a fixed effect. In contrast to the FM approach, the scanned marker can also be treated as a random effect and such a method is called the random model (RM) approach. The RM approach allows the use of the effective number of tests to perform Bonferroni correction and thus significantly increases the statistical power. However, the RM approach requires estimation of two genetic variance components (the variance of the scanned marker and the polygenic variance) and thus involves high computational cost. The main focus of this dissertation is the development of a new

method named randomized fixed model (RFM) methodology. By this method, we can perform the RM GWAS using results of the FM analysis without involving additional computation.

There are three chapters in this dissertation. The first chapter introduces the main concepts in GWAS, LMM and corrections for multiple hypotheses testing. The second chapter describes the RFM methodology, and demonstrates in both simulated data and real human data that the RFM is as powerful as the RM, with reduced computational complexity. In the third chapter, an outlier detection approach using a mixture model for significance test is described. Compared to Bonferroni correction method, this approach boosts the statistical power with the genome-wide type I error rate still controlled below 0.05. Thus, the outlier detection approach can be an alternative method for Bonferroni correction.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Genome-wide association studies (GWAS)

GWAS (Risch & Merikangas, 1996) are statistical approaches that examine a genome-wide set of single-nucleotide polymorphisms (SNPs) in a large population to identify genetic variations that are associated with a complex trait. Compared to the traditional linkage analysis, GWAS prove to be more powerful and have higher resolution by taking advantage of historical and evolutionary recombination events at the population level (Nordborg & Tavare, 2002; Risch & Merikangas, 1996). In recent years, with the advances of the Human Genome Project (Lander et al., 2001; Venter et al., 2001) and the International Human HapMap project (International HapMap, 2003), novel disease loci that were previously unknown were uncovered to be associated with human complex diseases through GWAS method (Wellcome Trust Case Control, 2007). The same strategy for genetic dissection of complex traits is also being applied to many plant and animal species.

### 1.1.1 Applications of GWAS in plants and human

#### 1.1.1.1 Progress of GWAS in plants

In rice, Huang et al. (2010) sequenced a large collection of 517 diverse rice landraces and constructed a high-density haplotype map containing ~3.6 million SNPs using a highly accurate data-imputation method. Through a GWAS for 14 agronomic traits including morphological characteristics, yield components, grain quality, coloration and physiological features, a total of 80 association signals were detected. Zhao et al. (2011) conducted a GWAS in a collection of 413 diverse *O. sativa* accessions from 82 countries by using 44,100 high-quality SNP variants and 34 agronomic, developmental, and morphological traits. Dozens of common variants were identified to affect the performance of these traits. Revealing the genetic basis of these complex traits provided insights into the improvement of yield, quality and sustainability of rice. In a metabolic GWAS based on ~6.4 million SNPs collected from 529 diverse *Oryza sativa* accessions, Chen et al. (2014) identified hundreds of common variants influencing numerous secondary metabolites with large effects. The GWAS also facilitated the identification and annotation of a total of 166 metabolites by linking the unknown metabolites to related genes. Yano et al. (2016) identified 4 new genes associated with important agronomic traits using GWAS based on whole-genome sequencing and the selection of candidate genes based upon the estimated functional importance of nucleotide polymorphisms.

Wang et al. (2012) carried out a GWAS of head smut resistance in maize using 45,868 SNPs from 144 inbred lines, showing that 18 novel candidate genes were associated with resistance to head smut disease. These candidate genes could be categorized into three groups: resistance genes, disease response genes and genes with possible disease resistance functions. This research provided a basis for cloning the candidate genes to further dissect the complicated mechanism of head smut resistance in maize. In a GWAS using 1.03 million SNP markers characterized in 368 maize inbred lines, Li et al. (2013) studied the genetic basis of maize oil biosynthesis and disclosed 74 loci significantly associated with fatty acid composition and kernel oil concentration. Examined by using coexpression analysis, expression QTL mapping and linkage mapping, more than half of the loci were localized in the mapped QTL intervals, and one third of the candidate genes were implicated in the oil biosynthesis pathway. The 26 loci that were associated with oil concentration explained as high as 83% of the phenotypic variation by using a simple additive effect model, indicating that the additive effect is of great importance in maize oil biosynthesis and accumulation. Mao et al. (2015)discovered that an 82-bp miniature inverted-repeat transposable element (MITE) insertion in the promoter region of a *NAC* gene (*ZmNAC111*) was significantly associated with variation in maize drought tolerance. When heterologously expressed in Arabidopsis, the MITE insertion repressed the expression of *ZmNAC111* via RNA-directed DNA methylation and H3K9 dimethylation. Increased expression of *ZmNAC111* in transgenic maize conferred

drought tolerance. Thus, the insertion of MITE resulted in lower *ZmNAC111* expression and susceptibility to drought.

GWAS have also been applied to other crops. Jia et al. (2013) constructed a haplotype map of foxtail millet by using 0.8 million common SNPs sequenced from 916 diverse varieties. The researcher identified 512 associated SNP loci for 47 agronomic traits in a GWAS. To facilitate gene discovery and marker-assisted breeding in sorghum, Morris et al. (2013) characterized ~265,000 SNPs in 971 sorghum accessions using genotyping-by-sequencing (GBS), and identified several loci for plant height and inflorescence architecture by GWAS.

#### 1.1.1.2 Progress of GWAS in human

Since first exploited to study age-related macular degeneration in 2005 (Klein et al.), GWAS started to gain favorability in research of many kinds of complex human traits. The Wellcome Trust Case Control Consortium (WTCCC, 2007) examined ~2,000 individuals and a shared set of ~3,000 controls, and detected 24 loci associated with six major human diseases with p-values less than $5 \times 10^{-7}$. Soon afterwards, GWAS approaches have been applied to many human complex diseases, with many loci identified for type 1 (Hakonarson et al., 2007; Todd et al., 2007) and type 2 diabetes (A. P. Morris et al., 2012; Zeggini et al., 2008), coronary heart disease (Helgadottir et al., 2007; McPherson et al., 2007; Samani et al., 2007; Schunkert et al., 2011), prostate cancer (Eeles et al., 2008; Gudmundsson et al., 2007; Gudmundsson et al., 2008; Haiman

et al., 2007), breast cancer (Easton et al., 2007; Hunter et al., 2007; Stacey et al., 2007), obesity (Thorleifsson et al., 2009), atrial fibrillation (Gudbjartsson et al., 2007) and schizophrenia (Ripke et al., 2013). National Human Genome Research Institute and European Molecular Biology Laboratory - European Bioinformatics Institute collaboratively developed an online catalog of SNP-trait associations database (www.ebi.ac.uk/gwas) regularly updated from published GWAS for investigating genetic architecture of common diseases (Welter et al., 2014). This catalog includes all eligible GWAS and association studies since the first published GWAS discovery on age-related macular degeneration in 2005 (Klein et al.).

Helgadottir et al. (2007) conducted a GWAS on Icelandic patients with Myocardial Infarction (MI) by testing 305,953 SNPs in a sample of 1607 cases and 6728 controls, and identified that a common sequence variant on chromosome 9p21, located near the tumor suppressor genes *CDKN2A* and *CDKN2B*, was associated with MI with high significance. Based upon a well-powered meta-analysis of 46 GWAS studies, 95 loci were revealed to associate with blood lipid traits such as high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol and triglycerides (TG) in a population of more than 100,000 (Teslovich et al., 2010). These blood lipid traits are strong predictors of heart disease. Schunkert et al. (2011) performed a meta-analysis of 14 GWAS of coronary artery disease (CAD) and identified 13 new loci associated with CAD with a p-value of less than $5 \times 10^{-8}$. They also confirmed 10 out of 12 previously reported loci associated with CAD.

The WTCCC (2007) uncovered 7 loci associated with type 1 diabetes (T1D) and 3 loci associated with type 2 diabetes (T2D) at p-value < $5 \times 10^{-7}$ by using SNP dataset undertaken in the British population. Todd et al. (2007) validated 4 of the 7 T1D associated loci detected by WTCCC (2007) and found robust associations of 4 more novel chromosome regions with T1D. Huang et al. (J. Huang et al., 2012) described that imputation based upon data from 1000 Genomes Project revealed novel association signals attributed to the very dense marker map and the great number of haplotypes. They observed two diabetes associated variants that were undetected in the original WTCCC (2007) analysis, but were reported by other later studies. One locus which is within the *IL2RA* gene is associated with T1D. The other locus associated with T2D is adjacent to the *CDKN2B* gene. Besides, they also identified two novel loci that were not previously reported. One is SNP rs11209026 in exon 9 of *IL23R* associated with Crohn's disease, and the other SNP rs1265564 is in the *CUX2* gene for association with T1D.

Human height is a heritable quantitative trait influenced by multiple loci. Weedon et al. (2007) examined a genome-wide association data from a total of 4,921 individuals, finding that a common variant of *HMGA2* oncogene rs1042725 was associated with adult and childhood height in the general population. Furthermore, this variant could explain approximately 0.3% of population height variation (about 0.4 cm increase in adult height per C allele). Hao et al. (2013) performed a GWAS in a Han Chinese population of 6,534, and identified three novel associated loci for human height. This study also confirmed the two loci *CS* (rs3816804) and *CYP19A1* (rs3751599) that were

previously reported in European populations, and detected 35 SNPs reported by previous study as well. Berndt et al. (2013) identified 4 new loci (*IGFBP4*, *H6PD*, *RSRC1*, *PPP2R2A*) affecting height by analyzing 263,407 European individuals.

### 1.1.2 Linkage disequilibrium

The basis of the comparative high-resolution of association mapping is the structure of linkage disequilibrium (LD) across the genome. LD describes the degree of non-random association between alleles at different loci along the genome in a population. Loci are in LD when the observed frequency of co-occurrence for two alleles is different from the frequency expected if the two loci are independent. The structure of LD could be affected by many genetic and non-genetic factors, such as selection, drift, recombination, mating pattern and admixture (the mixture of two or more genetically distinct populations). In a population of a fixed size under random mating/crossing, accumulated random recombination events will break apart contiguous chromosomal segments. Eventually, a pair of loci on a chromosome in the population will move from linkage disequilibrium to linkage equilibrium.

The rate of LD decay is influenced by many factors, including the number of founding chromosomes in the population, the number of generations through which the population has passed, the population size and the reproduction mode. Generally, the LD decays slower in self-pollinated crops, such as wheat and rice, than it does in cross-pollinated crops, such as maize (Flint-Garcia, Thornsberry, & Buckler, 2003). In a genome

with strong LD, only a small number of SNPs are required to adequately cover the whole genome, but the mapping resolution will be low. In a genome where LD decays in a short distance, a great number of SNPs are needed, but the mapping resolution will be high.

There are several metrics proposed to measure LD (Devlin & Risch, 1995). The two most commonly used measures are *D'* (Lewontin, 1964) and $r^2$ (Hill & Robertson, 1968). Both are ultimately related to *D*, which is the deviation of the observed haplotype frequency from the expected frequency in the equilibrium state (Lewontin & Kojima, 1960).

$$D = P_{AB} - P_A P_B \tag{1.1}$$

$P_{AB}$ is the frequency of haplotype AB; $P_A$ and $P_B$ represent the frequency of the allele A and B, respectively.

$$D' = \frac{|D|}{D_{\max}} \tag{1.2}$$

where $D_{\max} = \min\left(P_A P_b, P_a P_B\right)$, if $D > 0$; $D_{\max} = \min\left(P_A P_B, P_a P_b\right)$, if $D < 0$.

$$r^2 = \frac{D^2}{P_A P_a P_B P_b} \tag{1.3}$$

*D'* is the absolute ratio of *D* over the maximum value that *D* could take given the allele frequencies and is scaled between 0 and 1. A *D'* value of 0 denotes no LD, and *D'* = 1 indicates complete LD, implying no recombination between the two loci. $r^2$ is a

statistical measure of correlation between the two alleles of the two loci. It also has a range between 0 and 1. $r^2$ = 1 indicates that given the genotype of one locus, one can directly predict the genotype of another locus, thus only one of the two loci needs to be genotyped to capture the allelic variation.

The set of SNPs selected based on LD patterns to capture the allelic variation of nearby SNPs are called tag SNPs. In an association study, due to the presence of LD, the detected association between a SNP and a quantitative trait could be a direct association or an indirect association. If the functional SNP itself is genotyped in the study and identified to be associated with the trait, then it's a direct association. If the functional SNP is not typed, but a tag SNP which is in high LD with the causal SNP is typed and identified to associate with the trait, then it's an indirect association. Thus, a statistically significant association signal in a GWAS does not necessarily mean that the associated SNP is the causal variant and additional studies are required to further locate the causal SNP.

### 1.1.3 Population structure and relative kinship

Population structure refers to the presence of allele frequency differences among subpopulations in a population, due to systematic ancestry differences, e.g., various geographical origins in plants, different breeds in animals, and Asian, African and European subpopulations in human. When subpopulations differ both in allele frequencies and in phenotype prevalence, this will lead to spurious associations in a

GWAS. Because some SNPs may be specific in a subpopulation for a historical reason, but not necessarily associated with any traits. If a method does not take population structure into account, these subpopulation-specific SNPs will be detected as loci associated with the trait of interest.

Recently, several methods have been proposed to diagnose and correct for population structure. Structured association (SA, Pritchard, Stephens, & Donnelly, 2000) and principal components analysis (PCA, Price et al., 2006) are the two most common methods. SA uses the STRUCTURE software package to infer population structure (Q) by using a set of random markers and then classify individuals into different subpopulation clusters. One limitation of this method is its intensive computational cost on large data sets. Besides, the definition of subpopulations is typically very subjective, and the assignments of individuals to clusters are quite sensitive to the number of clusters defined. PCA can be implemented by the EIGENSTRAT method. It summarizes genetic variation observed from all markers into a smaller number of underlying variables called principal components. These principal components reduce the data to a small number of dimensions, while accounting for as much variation in the data as possible. Usually the first few principal components will be incorporated as covariates into a GWAS model to adjust for the effects caused by population structure. Compared to SA, PCA does not need to deduce the number of subpopulations, and is also more computationally tractable on a genome-wide scale.

Apart from population structure, relatedness among individuals can also result in false positives. Yu et al. (2006) developed a unified mixed-model method for association mapping to simultaneously account for multiple levels of relatedness as detected by random genetic markers. In this method, marker-based population structure Q and relative kinship (K, defines the degree of genetic covariance among individuals) were simultaneously introduced into a mixed-model framework for marker-trait associations. This method is commonly known as the Q + K model. In general, the Q + K model results in a better performance than the Q model or the K model alone (Yu et al., 2006).

### 1.1.4 Missing heritability

GWAS have identified a great number of genetic variants associated with complex traits, shedding light on the genetic architecture of these traits. However, most variants identified so far accounted for only a trivial fraction of the phenotypic variance in the population. The remaining unexplained heritability is called the missing heritability (Manolio et al., 2009). Many reasons for this missing heritability have been proposed, including rare variants, gene-gene interactions (epistasis) and gene-environment (G × E) interactions.

The power to detect a variant is a function of the allele frequency. Causal variants with a low allele frequency can hardly have adequate effect on the population as a whole, therefore they are difficult to detect (Myles et al., 2009). Even if a rare allele has a strong effect on the phenotype, it might be difficult to detect by population mapping

because it is less well represented in SNP databases and the tag SNPs are usually designed to tag common variants (with frequencies > 5%). Unfortunately, in most species, a large proportion of alleles are rare, according to a population genetics theory. For instance, in rice, there are about 44% of the polymorphisms are rare variants. Family-based mapping can be used to detect such rare functional alleles (Laird & Lange, 2006). Because by creating crosses, the allele frequencies in the progeny can be artificially inflated to provide increased mapping power. De et al. (2013) included both common and rare variants in their analysis by introducing suitable weighting schemes to downweight the more common variants and upweight rarer variants.

Zuk et al. (2012) thought that a substantial proportion of missing heritability could be due to the genetic interactions among loci. There has been notable development of methodology and software to detect epistasis among loci in the past few years (Wei, Hemani, & Haley, 2014). However, efficient testing of epistasis in a GWAS is still a challenging problem owing to the large number of possible pairs of interaction to be considered. For instance, with a total collection of 1 million SNPs, there are going to be approximately 500 billion possible interactions. The large number of interactions causes a severe penalty to multiple hypothesis testing. With 1 million SNPs, to achieve 5% genome-wide significance, a Bonferroni correction would require a p-value of less than $10^{-13}$ for a single test. Besides, the computational burden further limits the complexity of any genome-wide epistasis testing model that one may consider. In general, people use the two-step strategy for a genome-wide search for epistasis. In the first step, use an

approximate but fast pair-wise interaction tests for an initial genome-wide screening. In the second step, apply the full regression models to test for significance of the most promising interactions from the initial screening (Lewinger et al., 2013).

Complex traits can also be influenced by the environment. Therefore, the missing heritability could be partially due to the G × E interactions, defined as the joint effect of genetic and environmental factors which cannot be explained by their marginal effects (Thomas, 2010). The major challenges to a successful G × E study are exposure assessment, sample size and heterogeneity. Many statistical analysis approaches have been developed to study G × E interactions, into which the various ways of genetic effects can be modulated by environmental exposures and the number of levels of environmental exposures will be taken account.

## 1.2 Linear mixed model (LMM)

LMM, firstly proposed by Yu et al. (2006), is a widely used method in GWAS. When the *k*th marker (*k* = 1, …, m) is scanned, the model is described as

$$y = X\beta + Z_k\gamma_k + \xi + \varepsilon \tag{1.4}$$

where $y$ is an n × 1 vector of phenotypic values for a quantitative trait, $X$ is a design matrix of covariates, $\beta$ is a vector of effects for the covariates to reduce inference from other non-genetic factors, $Z_k$ is an n × 1 vector of genotype indicator variable of the *k*th marker, $\gamma_k$ is the marker effect, $\xi \sim N(0, K\phi^2)$ captures the polygenic effect with a

multivariate normal distribution where *K* is the covariance structure (kinship matrix) derived from genome-wide markers, $\phi^2$ is the polygenic variance, $\varepsilon \sim N(0, I\sigma^2)$ is a vector of residual errors normally distributed with a common error variance $\sigma^2$. If $\gamma_k$ is treated as a fixed effect, the expectation of $y$ is

$$E(y) = X\beta + Z_k\gamma_k \tag{1.5}$$

and the variance is

$$\text{var}(y) = k\phi^2 + I\sigma^2 = \Sigma \tag{1.6}$$

Parameters can be estimated using the restricted maximum likelihood (REML) and the log-restricted likelihood function is

$$L(\theta) = -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}(Y - X\beta)^T\Sigma^{-1}(Y - X\beta) - \frac{1}{2}\ln\left|X^T\Sigma^{-1}X\right| \tag{1.7}$$

where $\theta = \left\{\beta, \gamma_k, \phi^2, \sigma^2\right\}$. The Wald test statistic for $H_0 : \gamma_k = 0$ is

$$W_k = \frac{\hat{\gamma}_k^2}{\text{var}(\hat{\gamma}_k)} \tag{1.8}$$

$W_k$ follows a Chi-square distribution with one degree of freedom under the null hypothesis. The *p*-value for the *k*th marker is

$$p_k = 1 - \Pr\left(\chi_1^2 \le W_k\right) \tag{1.9}$$

However, the original mixed model is very computationally intensive because it involves a large number of matrix multiplications and inverses. Kang et al. (2008) proposed a new method called efficient mixed-model association (EMMA), substantially improving the computational efficiency and reliability of the results by implementing eigen-decomposition. They first decomposed the *K* matrix to reduce the computational cost from cubic to linear complexity, and then estimated the polygenic and residual variance components for each marker by treating the variance ratio $\lambda = \sigma_{\xi}^2 / \sigma^2$ as the parameter. Eigen decomposition of matrix *K* is performed as

$$K = UDU^T \tag{1.10}$$

where $D$ is a diagonal matrix of the eigenvalues and $U$ is an n × n matrix of eigenvectors. The eigenvector matrix is orthogonal in the sense that $UU^T = I$. Let

$$H = K\lambda + I = UDU^T + I = U(D\lambda + I)U^T \tag{1.11}$$

The log determinant of matrix $H$ is

$$\ln|H| = \ln|D\lambda + I| = \sum_{j=1}^{n} \ln(\delta_j \lambda + 1) \tag{1.12}$$

There are various quadratic forms involved in the likelihood function in the form of $a^T H^{-1} b$, for instance, $X^T H^{-1} X$, $X^T H^{-1} y$ and $y^T H^{-1} y$. By employing eigenvalue decomposition, we can rewrite the quadratic form by

$$a^T H^{-1} b = a^T U (D\lambda + I)^{-1} U^T b = a^{*T} (D\lambda + I)^{-1} b^* = \sum_{j=1}^{n} a_j^{*T} b_j^* (\delta_j \lambda + 1)^{-1} \qquad (1.13)$$

where $a^* = U^T a$ and $b^* = U^T b$.

On the basis of EMMA, Zhou and Stephens (2012) proposed an efficient exact method, genome-wide efficient mixed-model association (GEMMA), which is approximately n (the sample size) times faster than EMMA, making exact analysis feasible for GWAS with a large population. Lippert et al. (2011) presented an efficient method, factored spectrally transformed linear mixed models (FaST-LMM), that requires only one singular value decomposition of the *K* matrix to test all SNPs. They also suggested using a relatedness matrix calculated from only a few thousand SNPs to reduce computing time. Several approximate methods have also been proposed to save the computational cost, for example, EMMA expedited (EMMAX, Kang et al., 2010) and population parameters previously determined (P3D, Zhang et al., 2010). These two approaches assumed that the variance parameters are the same across all SNPs and simply used the pre-estimated variance components under the null model for each tested marker, thus avoiding estimating variance components repeatedly and removing the expensive cubic computation per SNP.

## 1.3 Corrections for multiple hypotheses testing

If one hypothesis test is performed at the 5% level, then the null hypothesis is rejected if the p-value falls below 0.05. This means that 5% of the time, the null hypothesis is

rejected incorrectly and we detect a false positive (also known as the type I error). This probability is relative to a single hypothesis test; however, in the case of GWAS where thousands to millions of tests are conducted simultaneously and each one with its own probability of false positive, the probability of detecting one or more false positives over the entire analysis (family-wise error rate, FWER) is given by $1 - (1-\alpha)^m$ (m is the number of tests performed), which will be larger than 0.05. Consider a case where there are 20 hypotheses tests and the significance level of each test is 0.05. The FWER will be $1 - (1-0.05)^{20} \approx 0.64$. With the number of tests increase, the rate of type I error keeps going up. Therefore, in order to retain a prescribed FWER $\alpha$ in an analysis involving multiple tests, the type I error rate for each test must be more stringent than $\alpha$.

### 1.3.1 Bonferroni correction

Bonferroni correction is one of the most commonly used methods to correct for multiple comparisons, and is the most conservative method. The Bonferroni correction adjusts the $\alpha$ value from 0.05 to 0.05/m where m is the number of tests performed. This is based on Boole's inequality that if each of the m tests is conducted to have a type I error rate of $\alpha/m$, the FWER will not exceed $\alpha$. This correction assumes that all tests are independent of one another, which is usually not the case due to the LD among markers. Thus, the Bonferroni correction could be extremely conservative, leading to an increased rate of type II error (false negative).

Several strategies have been proposed to maintain the overall rate of false positives without excessively inflating the rate of false negatives. One technique is to use the number of markers on each chromosome as their own denominator, so that every chromosome has its own significance criteria based on the number of markers on that chromosome. Another strategy is to evaluate the effective number of independent tests, and to use the effective number rather than the total number of markers as the denominator (M. X. Li et al., 2012). Due to the presence of LD in the genome, there are an effective number of independent genomic regions, which represent an effective number of independent tests.

### 1.3.2 Sequentially rejective test

Holm (1979) developed a simple sequentially rejective multiple test procedure, which was also based on the Boole inequality, and was therefore also called the sequentially rejective Bonferroni test. It uses a stepwise algorithm in simultaneous inference by adjusting the rejection criteria of each of the individual hypotheses. Let $H_1 \ldots H_m$ be a family of hypotheses and $P_1 \ldots P_m$ the corresponding p-values. This method starts by ordering the p-values from lowest to highest as $P_{(1)} \ldots P_{(m)}$ and let the associated null hypotheses be $H_{(1)} \ldots H_{(m)}$. For a given significance level $\alpha$, compare $P_{(1)}$ with $\alpha/m$. If $P_{(1)}$ is smaller than $\alpha/m$, we reject $H_{(1)}$, and compare $P_{(2)}$ with $\alpha/(m-1)$, and so forth. Until we reach the kth test where $P_{(k)} > \alpha/(m-k+1)$, then we reject all previous k-1 hypotheses and

do not reject $H_{(k)}$ … $H_{(m)}$. This correction ensures that the FWER will not exceed α and is uniformly more powerful than the classical Bonferroni correction.

### 1.3.3 False discovery rate (FDR)

The FDR is defined as the proportion of false positives among all significant detections. It was formally described by Benjamini and Hochberg (1995) as a less conservative approach than the FWER in flagging possibly noteworthy observations. The Benjamini-Hochberg procedure (BH procedure) controls the FDR at level α. First, we sort the p-values in ascending order and denote them by $P_{(1)}$ … $P_{(m)}$. For a given significance level α, find the largest k such that $P_{(k)} \leq (k/m)\alpha$. Then reject the null hypothesis for all $H_{(i)}$, for i = 1,2, …, k. This procedure which controls the FDR is generally more powerful than the FWER controlling methods. In addition, the advantage increases with the number of tested hypotheses and the number of non-true null hypotheses.

### 1.3.4 Permutation test

Permutation test, which has evolved from the works of Fisher and Pitman in the 1930s, is a type of widely applicable non-parametric test. It is a straightforward approach to generate the empirical distribution of the test statistic under the null hypothesis by random reassigning the labels on the data points. In a GWAS, we randomly shuffle the genotypes of individuals in the dataset while leaving the phenotypes unchanged, effectively breaking the association between genotypes and phenotypes of that dataset.

Perform a GWAS on the permutated data, calculate the test statistic for each marker, and record the largest test statistic from it. Then repeat this process a predetermined number of times n, and we will get n maximum test statistics. The collection of the best statistics which show the greatest associations caused by chance can be used to construct a null distribution. We then compare the test statistics in the original dataset with the null distribution to obtain an estimate of statistical significance. For a significance level of $\alpha$, the 100(1-$\alpha$) percentile is the empirical critical value. Any test statistic that is greater than the 100(1-$\alpha$) percentile is significant, and the type I error rate is controlled to be $\alpha$ or less (Churchill & Doerge, 1994).

A crucial question of permutation test is how large n should be. This depends on the significance level $\alpha$. It is recommended that at least 1,000 shuffles be made at $\alpha = 0.05$, and as many as 5,000 shuffles be used at $\alpha = 0.01$. This method provides robust and powerful significance tests that is easy to apply in practice, although it is more computationally expensive than other statistical tests. Several packages have been developed to conduct permutation test for GWAS, including PLINK (Purcell et al., 2007), PRESTO (Browning, 2008), and PERMORY (Pahl & Schafer, 2010).

# Chapter 2

# Randomized Fixed Model

## 2.1 Introduction

The state of the art method for genome-wide association studies (GWAS) is the mixed model methodology implemented using the exact method in which the polygenic variance is re-estimated for each marker scanned (GEMMA). This model treats the scanned marker effect as a fixed effect. Therefore, we call it the fixed model (FM) approach. In contrast to the FM approach, the scanned marker can be treated as a random effect and such a method is called the random model (RM) approach. The empirical Bayes (EB) method of genome-wide association studies (GWAS) developed by Wang et al. (2016) is a RM approach in the sense that the effect of a scanned marker is considered as a random effect. The estimation and test of a random marker effect are selectively shrunken toward zero, leading to reduced background noise in the test statistic profile (Manhattan) plot of GWAS. To compensate the reduced test statistic, Wang et al. (2016) adopted the effective number of tests to perform Bonferroni correction and showed that the modified Bonferroni correction has significantly increased the statistical power and, in the meantime, maintained the genome-wide type I error below the controlled 0.05 level. The RM approach requires simultaneous use of

eigen-decomposition and Woodbury matrix identity to improve the computational speed so that the method is practically applicable to data sizes commonly seen in GWAS.

The RM approach requires estimation of two genetic variance components (the variance of the scanned marker and the polygenic variance) and thus involves high computational cost. In this study, we developed a new method to perform the RM GWAS using results of the FM analysis (marker effects treated as fixed effects) without involving additional computation. We call the new method randomized fixed model (RFM) methodology. With this new method, the modified Bonferroni correction (Q. Wang et al., 2016) can still be used to control false positive rates more precisely to boost the statistical power.

## 2.2 Methods

### 2.2.1 GWAS using the FM

Let $y$ be an $n \times 1$ vector of a quantitative trait to be studied. Let $k = 1, ..., m$ indexes markers where *m* is the total number of markers. When the *k*th marker is scanned, the FM is described as

$$y = X\beta + Z_k\gamma_k + \xi + \varepsilon \quad (2.1)$$

where $X\beta$ represents some systematic covariates placed in the model to control the residual error, $Z_k$ is a vector of numeric codes for the genotypes of the *k*th marker, $\gamma_k$ is the marker effect, $\xi \sim N(0, K\phi^2)$ is a polygenic effect with a multivariate normal distribution with a covariance structure *K* inferred from genome-wide markers (called the kinship matrix), $\phi^2$ is the polygenic variance, $\varepsilon \sim N(0, I\sigma^2)$ is a vector of residual errors of normally distributed with a common error variance $\sigma^2$. Under the FM approach, $\gamma_k$ is a fixed effect (parameter) estimated and tested using the conventional mixed model methodology. Although the model is a mixed model, we call it FM because the marker effect of interest is treated as a fixed effect. This is in contrast to the RM where $\gamma_k$ is treated as a random effect with a normal distribution.

Parameters are estimated using the restricted maximum likelihood (REML) as given by Zhou and Stephens (2012). Let $\tilde{\gamma}_k$ and $\tilde{V}_k = \mathrm{var}(\tilde{\gamma}_k)$ be the estimated marker effect and the variance from the FM analysis. The Wald test for $H_0 : \gamma_k = 0$ is

$$\tilde{W}_k = \frac{\tilde{\gamma}_k^2}{\tilde{V}_k} \tag{2.2}$$

Assuming that the Wald test statistic follows a Chi-square distribution with one degree of freedom, the *p*-value for the *k*th marker is calculated using

$$p_k = 1 - \Pr(\chi_1^2 \leq \tilde{W}_k) \tag{2.3}$$

### 2.2.2 Randomization of the FM

For the same linear model given in equation (1), if $\gamma_k \sim N(0, \phi_k^2)$ is assumed, the model becomes a RM. Approximate and exact methods for REML estimates of variance components are available (Wang et al. 2015). Here, we do not estimate $\phi_k^2$ along with the other variance components; rather, we use the results of the FM analysis to estimate $\phi_k^2$ and then derive the posterior mean of $\gamma_k$. This posterior mean is called the best linear unbiased prediction (BLUP). Significance test is then performed for the BLUP of $\gamma_k$. The new method of estimation for $\phi_k^2$ is based on the assumption that $\tilde{\gamma}_k$ and $\tilde{V}_k$ estimated from the FM are sufficient statistics (all information from the data has been captured by $\tilde{\gamma}_k$ and $\tilde{V}_k$). This assumption is true when the polygene and the residual error are normally distributed. We want to derive a randomized counterpart of $\tilde{\gamma}_k$ from these sufficient statistics.

Let $\gamma_k$ be the true but unknown effect. Before we collect data and obtain the sufficient statistics, we assign a prior distribution to the effect, $\gamma_k \sim N(0, \phi_k^2)$. Assume that $\tilde{\gamma}_k$ is an unbiased estimate of the true effect, we can write a simple RM for $\tilde{\gamma}_k$

$$\begin{aligned} \tilde{\gamma}_k &= \gamma_k + \nu_k \\ \gamma_k &\sim N(0, \phi_k^2) \\ \nu_k &\sim N(0, \tilde{V}_k) \end{aligned} \tag{2.4}$$

where $\nu_k$ is an error term with a normal distribution of mean zero (because $\tilde{\gamma}_k$ is an unbiased estimate of $\gamma_k$) and known variance $\tilde{V}_k$. This model allows us to estimate $\phi_k^2$ and then predict $\gamma_k$. The log likelihood function of the RM is

$$L(\phi_k^2) = -\frac{1}{2}\ln(\phi_k^2 + \tilde{V}_k) - \frac{1}{2}\frac{\tilde{\gamma}_k^2}{\phi_k^2 + \tilde{V}_k} \tag{2.5}$$

The derivative of the likelihood function with respect to $\phi_k^2$ is

$$\frac{\partial L(\phi_k^2)}{\partial \phi_k^2} = -\frac{1}{2}\frac{1}{(\phi_k^2 + \tilde{V}_k)} + \frac{1}{2}\frac{\tilde{\gamma}_k^2}{(\phi_k^2 + \tilde{V}_k)^2} \tag{2.6}$$

Setting the derivative to zero and solving for $\phi_k^2$ yields

$$\tilde{\phi}_k^2 = \tilde{\gamma}_k^2 - \tilde{V}_k \tag{2.7}$$

When $\tilde{\gamma}_k^2 < \hat{V}_k$, we set $\tilde{\phi}_k^2 = 0$ because a variance cannot be negative. The BLUP of $\gamma_k$ is

$$\hat{\gamma}_k = \left(\frac{1}{\tilde{\phi}_k^2} + \frac{1}{\tilde{V}_k}\right)^{-1}\left(\frac{0}{\tilde{\phi}_k^2} + \frac{\tilde{\gamma}_k}{\tilde{V}_k}\right) = \left(\frac{1}{\tilde{\phi}_k^2} + \frac{1}{\tilde{V}_k}\right)^{-1}\frac{\tilde{\gamma}_k}{\tilde{V}_k} = d_k\tilde{\gamma}_k \tag{2.8}$$

where

$$d_k = 1 - \frac{\tilde{V}_k}{\tilde{\gamma}_k^2} \tag{2.9}$$

Derivation of the above equation was obtained by substituting $\tilde{\phi}_k^2$ by $\tilde{\gamma}_k^2 - \tilde{V}_k$ (see Supplemental Material Note S1). The posterior variance of $\hat{\gamma}_k$ is

$$\text{var}(\hat{\gamma}_k) = \hat{V}_k = \left( \frac{1}{\tilde{\phi}_k^2} + \frac{1}{\tilde{V}_k} \right)^{-1} = d_k \tilde{V}_k \tag{2.10}$$

The Wald test statistic under the RM framework is

$$\hat{W}_k = \frac{\hat{\gamma}_k^2}{\hat{V}_k} = \left( \frac{1}{\tilde{\phi}_k^2} + \frac{1}{\tilde{V}_k} \right)^{-1} \frac{\tilde{\gamma}_k^2}{\tilde{V}_k^2} = d_k \tilde{W}_k \tag{2.11}$$

The degree of confidence (Mackay, 1992) of marker *k* is defined as

$$d_k = 1 - \frac{\hat{V}_k}{\tilde{\phi}_k^2} = 1 - \frac{\tilde{V}_k}{\tilde{\gamma}_k^2} \tag{2.12}$$

Note that

$$d_k = 1 - \frac{\tilde{V}_k}{\tilde{\gamma}_k^2} = \frac{\tilde{W}_k - 1}{\tilde{W}_k} \tag{2.13}$$

Further manipulation of $\hat{W}_k$ leads to

$$\hat{W}_k = d_k \tilde{W}_k = \tilde{W}_k - 1 \tag{2.14}$$

When $\tilde{W}_k < 1$, we set $\hat{W}_k = 0$. Since majority of the markers are not associated with the trait and they should have $\tilde{W}_k < 1$ and thus $\hat{W}_k = 0$, which explains why the RM approach will reduce the background noise to zero in the Manhattan plot of the test statistics.

The score test under the FM for $H_0 : \gamma_k = 0$ is equivalent to the Wald test $\tilde{W}_k$, as proved in Note S2 in the Supplemental Material. The advantage of the score test is that nuisance parameters are only estimated under the null model and the parameter of interest $\gamma_k$ is not estimated. The randomized score test remains the same as the randomized Wald test. The degree of confidence $d_k$ is calculated using $d_k = (\tilde{W}_k - 1)/\tilde{W}_k$.

### 2.2.3 Likelihood ratio test for the variance

Alternatively, we can test the null hypothesis $H_0 : \phi_k^2 = 0$ using the likelihood ratio test. The likelihood value evaluated at $\phi_k^2 = \tilde{\phi}_k^2 = \tilde{\gamma}_k^2 - \tilde{V}_k$ (MLE of the parameter) is

$$L(\tilde{\phi}_k^2) = -\frac{1}{2}\ln(\tilde{\phi}_k^2 + \tilde{V}_k) - \frac{1}{2}\frac{\tilde{\gamma}_k^2}{\tilde{\phi}_k^2 + \tilde{V}_k} = -\frac{1}{2}\ln(\tilde{\gamma}_k^2) - \frac{1}{2}\frac{\tilde{\gamma}_k^2}{\tilde{\gamma}_k^2} = -\frac{1}{2}\ln(\tilde{\gamma}_k^2) - \frac{1}{2} \qquad (2.15)$$

Under the null model, the likelihood value is

$$L(0) = -\frac{1}{2}\ln(0 + \tilde{V}_k) - \frac{1}{2}\frac{\tilde{\gamma}_k^2}{0 + \tilde{V}_k} = -\frac{1}{2}\ln(\tilde{V}_k) - \frac{1}{2}\frac{\tilde{\gamma}_k^2}{\tilde{V}_k} \qquad (2.16)$$

The likelihood ratio test statistic is

$$\begin{aligned} \Lambda_k &= -2\left[L(0) - L(\tilde{\phi}_k^2)\right] = \ln(\tilde{V}_k) + \frac{\tilde{\gamma}_k^2}{\tilde{V}_k} - \ln(\tilde{\gamma}_k^2) - 1 \\ &= \ln\left[\frac{\tilde{V}_k}{\tilde{\gamma}_k^2}\right] + \frac{\tilde{\gamma}_k^2}{\tilde{V}_k} - 1 = \tilde{W}_k - \ln(\tilde{W}_k) - 1 \end{aligned} \qquad (2.17)$$

The degree of shrinkage of the likelihood ratio test is even stronger than the Wald test.

### 2.2.4 Effective number of tests

Let *m* be the number of markers scanned in the entire genome. The effective number of tests is defined as

$$m_e = 1 + \sum_{k=1}^{m} d_k \tag{2.18}$$

where the value 1 added to make sure that the effective number of test is at least 1. Since $0 \le d_k \le 1$, we have $m_e \le m$. The Bonferroni correction is performed with $m_e$ instead of *m* so that the nominal *p*-value criterion should be calculated using $0.05 / m_e$ to control a genome-wide type I error at 0.05.

### 2.2.5 Theoretical consideration about the degree of confidence

Recall that the degree of confidence per marker is

$$d = \frac{\tilde{\gamma}^2 - \tilde{V}}{\tilde{\gamma}^2} \tag{2.19}$$

which can be further manipulated as

$$d = \frac{\tilde{\gamma}^2 - \tilde{V}}{\tilde{\gamma}^2} = \frac{\tilde{\gamma}^2 / \tilde{V} - \tilde{V} / \tilde{V}}{\tilde{\gamma}^2 / \tilde{V}} = \frac{\tilde{\gamma}^2 / \tilde{V} - 1}{\tilde{\gamma}^2 / \tilde{V}} = \frac{z^2 - 1}{z^2} \tag{2.20}$$

where $z = \tilde{\gamma} / \sqrt{\tilde{V}}$ is a standardize normal variable under the null model. The expectation of *d* is

$$E(d) = \int_{-\infty}^{-1} \frac{z^2 - 1}{z^2} \phi(z) dz + \int_{1}^{\infty} \frac{z^2 - 1}{z^2} \phi(z) dz = 2 \times 0.0753398 = 0.1506796 \qquad (2.21)$$

This means that, on average, the effective number of tests is 15% of the actual number of markers. Therefore, using the randomized FM approach, we may approximately use

$$m_e \approx 0.15m \qquad (2.22)$$

to calculate the effective number of tests. This approximate effective number of tests only serves as a guide line. The best strategy is to use data to calculate the effective number of tests.

## 2.3 Results

### 2.3.1 Demonstration in simulated data

The purpose of the simulation experiment is to compare the result of the RM approach with the RFM approach in terms of the estimated marker variances and the BLUP estimates of marker effects. In addition, we demonstrated that this method can be applied to QTL mapping in line cross experiments. Therefore, we simulated an $F_2$ population with a sample size 1000. We simulated a single large chromosome with 2400 cM in length. We placed 961 markers evenly distributed along the genome with 2.5 cM

distance per marker interval. We simulated 20 QTL with sizes and locations depicted in Figure 2.1. The mean of the simulated trait was 10 and the residual error variance was 10. No polygene was simulated but when a marker is scanned, the QTL not overlapping with the scanned marker will go to the polygene and be captured by the polygene in the model.

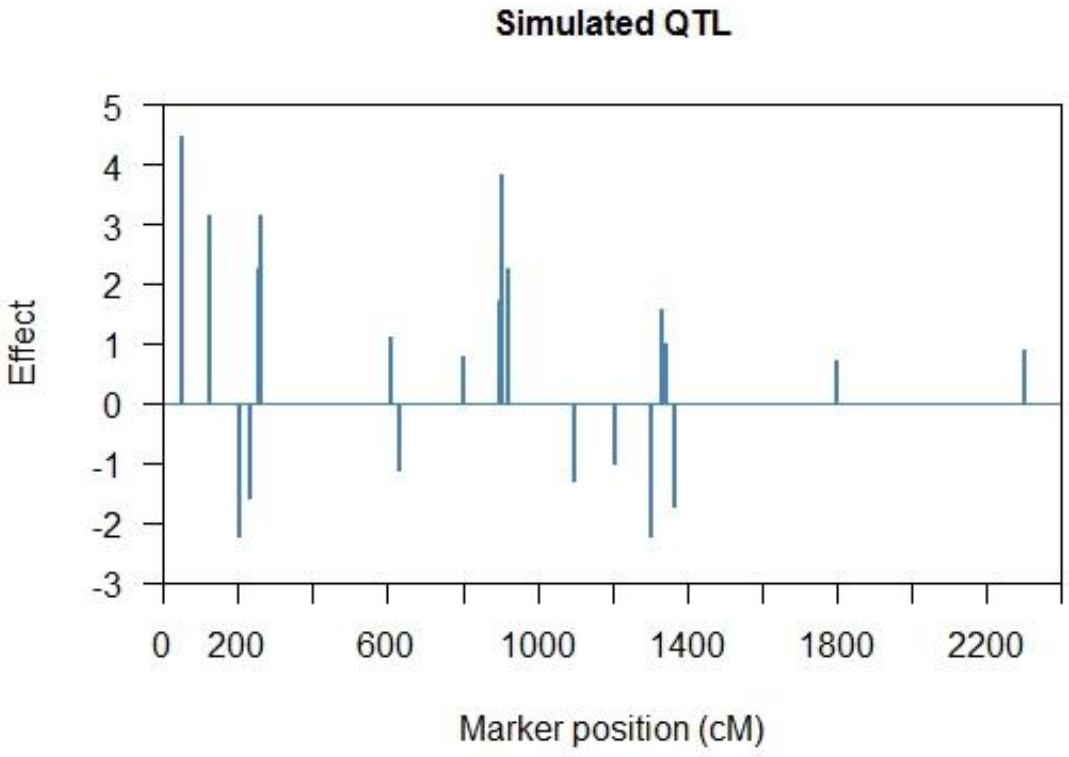


Figure 2.1 Effects of 20 QTL in the simulated data.

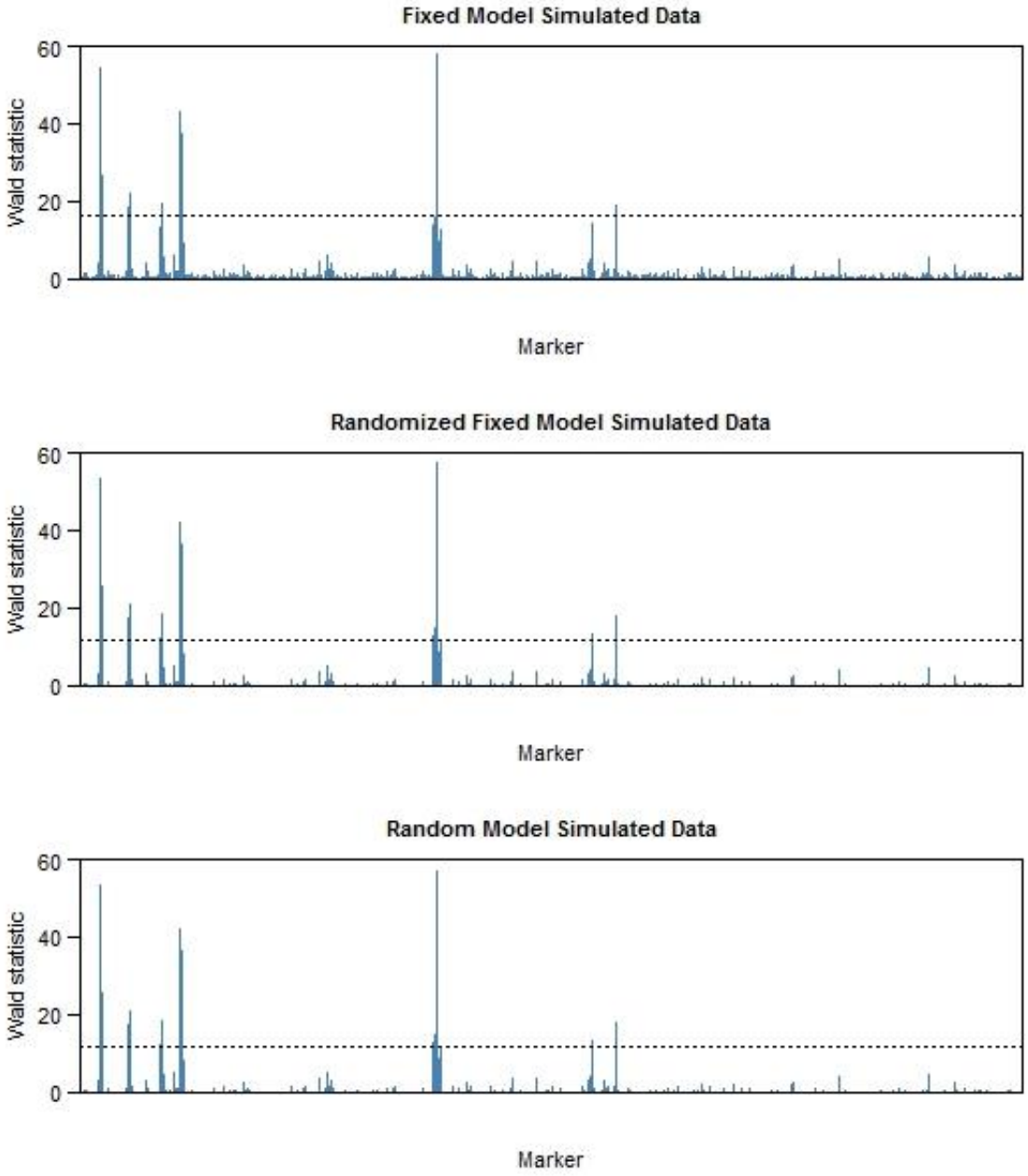


Figure 2.2 Wald test statistic profiles of GWAS on the simulated data by using three different models: FM, RFM, and RM. The horizontal dashed lines represent the critical values of each analysis at the significance level of 0.05.

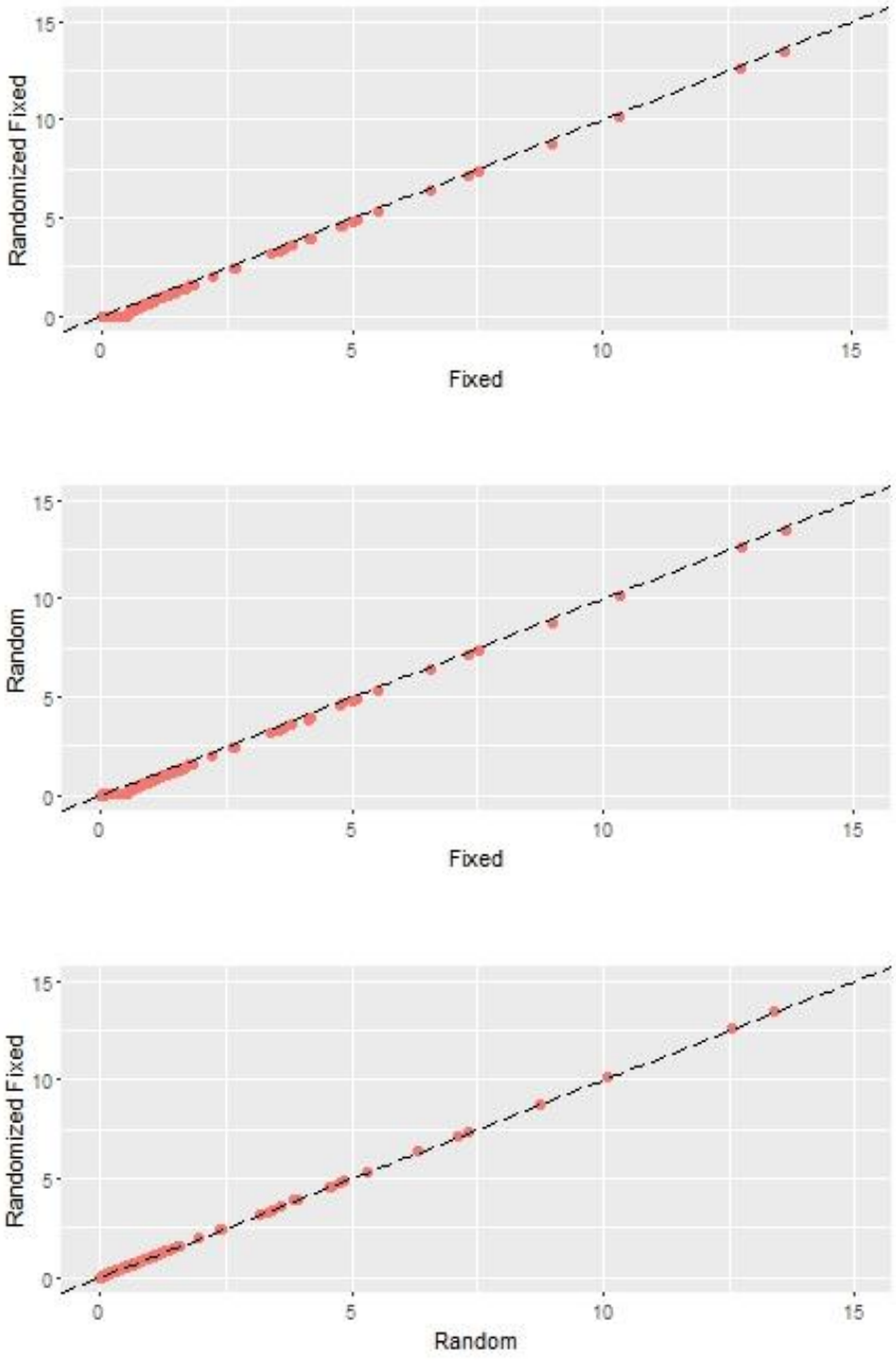


Figure 2.3 Pairwise comparisons of the -$\log_{10}$(p) among the three models: FM, RFM and RM. The GWAS was performed using the simulated data. The x axes and y axes represent the -log10(p) of the markers under the corresponding models.

Table 2.1 Number of detected markers and type I error rate under three models when performing GWAS on the simulated data.

| Method | Number of detected SNPs | Type I error rate |
|---|---|---|
| FM | 12 | 0 |
| RFM | 17 | 0 |
| RM | 17 | 0 |

Figure 2.2 showed GWAS result using three models in the simulated data. Compared to the FM analysis, markers in both RFM and RM analysis had a shrunken Wald statistic, leading to a reduced background noise. In spite of the shrinkage of the test statistic, the RFM and RM both had an increased power in detecting significant markers. This is due to the advantage that RFM and RM allow the use of an effective number of tests for Bonferroni correction and still maintain the genome-wide type I error rate below the controlled 0.05 level. From the figure, we can see that even though the Wald statistics were lowered down in RFM and RM methods, the criteria values were raised up, resulting in an increasing statistical power. The criteria of Wald statistic which follows a Chi-square distribution with one degree of freedom, was calculated from the critical p-value after Bonferroni correction. The critical Wald values in FM, RFM and RM were 16.37, 11.81, and 11.84, respectively. There were 12 markers detected as significant in the FM approach, 17 markers in the RFM approach and exactly the same 17 markers in the RM approach. All the 12 markers identified in FM were also detected in RFM and RM. There was no type I error, since all markers detected in the three methods were either true simulated QTL or neighboring QTL which is 2.5 cM apart from the true QTL.

More importantly, this figure supported that our RFM method produced almost the same result as the RM method with regards to the Wald statistic, the criteria level, and the detected SNPs. Therefore, the RM analysis could be realized by performing a RFM analysis since RFM is more computational efficient.

Figure 2.3 showed pairwise comparisons of the $-\log_{10}(p)$ among the three models. Both the x axes and y axes represented the $-\log_{10}(p)$ of the markers under the corresponding models. Negative $\log_{10}(p)$ is positive correlated with the Wald statistic. It was shown that FM had higher $-\log_{10}(p)$ values than RFM and RM, and RFM had almost identical $-\log_{10}(p)$ as RM, supporting that we can use the efficient RFM to replace RM.

### 2.3.2 Demonstration in Framingham heart study data

Since originating in 1948, the Framingham Heart Study (FHS), has been committed to identifying risk factors that contribute to cardiovascular disease (CVD), under the direction of the National Heart, Lung and Blood Institute (NHLBI). They have followed CVD development over a long period of time in a large group of three generations of participants who had not yet developed overt symptoms of CVD. The FHS first recruited an original cohort consisted of 5,209 individuals (2,336 men and 2,873 women) between the ages of 30 and 62 with no history of heart attack or stroke at the time of first examination, in the town of Framingham starting in 1948. Since then, participants of the original cohort have continued to return to the study every two years for a detailed physical examination, laboratory tests, and medical history. In 1971, the FHS added an

offspring cohort where adult children of the original participants and the spouses of these adult children were enrolled. A third-generation cohort was founded in 2002, consisting of children of the offspring cohort and grandchildren of the original cohort participants. Over the years, study on the FHS data had led to the identification of the major CVD risk factors including blood pressure, blood cholesterol, blood triglycerides, HDL (high density lipoprotein), diabetes, obesity, smoking, physical inactivity, as well as age, gender, and psychosocial issues. As to today, the study remains a world-class center for leading-edge heart, brain, bone, and sleep research.

In our study, the FHS data were downloaded from the dbGAP databases (phg000005.v5). There are 6,161 subjects genotyped at ~500,000 SNP markers and examined with 21 clinical phenotypes including HDL, total cholesterol and triglycerides, as well as a few phenotypes that can be used as covariates such as IDtype (generation), sex, age and body mass. In section 2.3.2, our purpose is to demonstrate that our RFM model generates almost the same GWAS result as that of the RM. Therefore, we only use markers on chromosome 8 and triglycerides phenotype to perform GWAS. In the next section (section 2.3.3), we will show the GWAS result using the genome-wide markers of all 22 chromosomes with three phenotypes: triglycerides, total cholesterol and HDL.

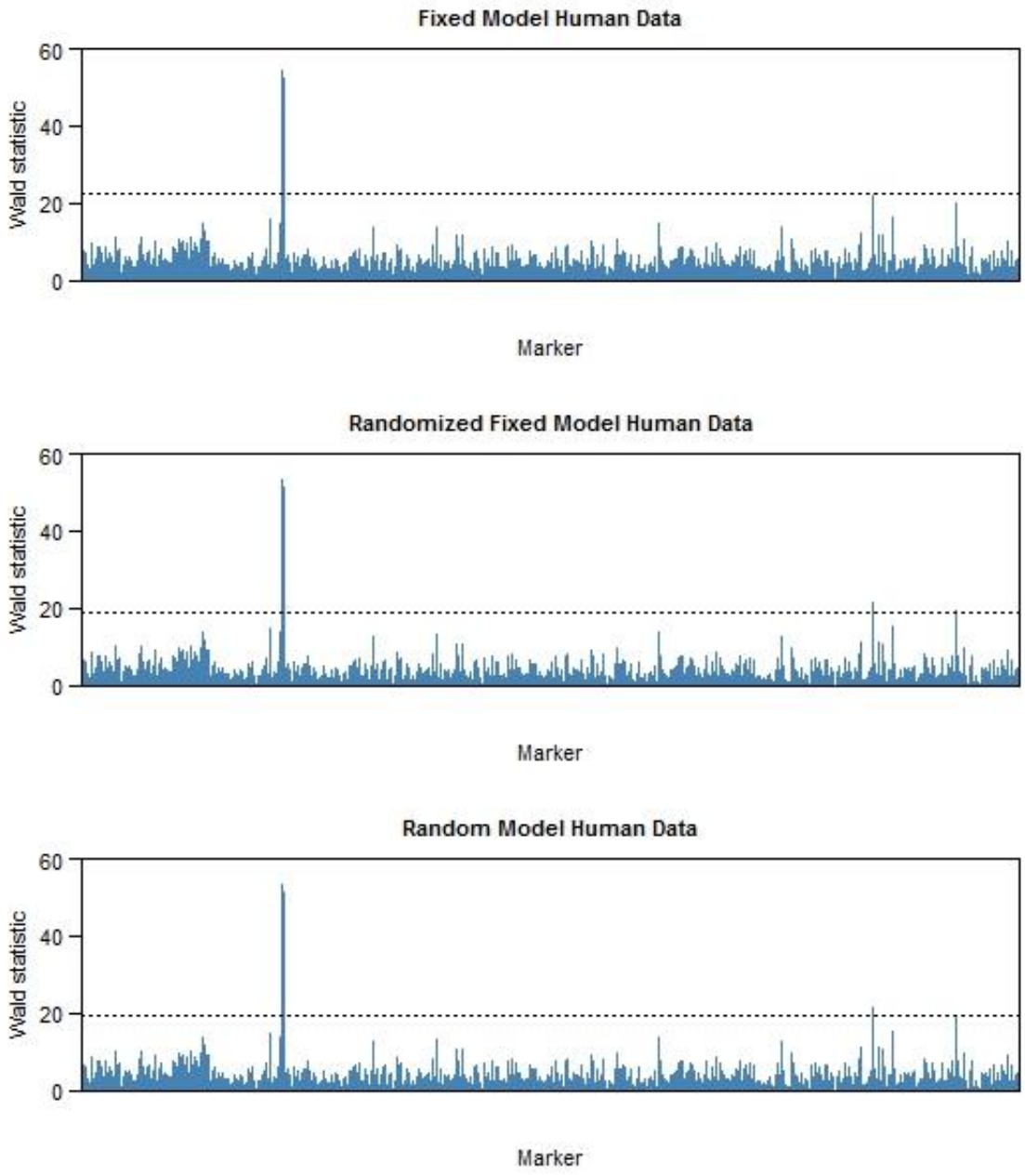


Figure 2.4 Wald test statistic profiles of GWAS on the FHS data by using three different models: FM, RFM, and RM. The GWAS was performed using SNPs on chromosome 8 and trait triglycerides. The horizontal dashed lines represent the critical values of each analysis at the significance level of 0.05.

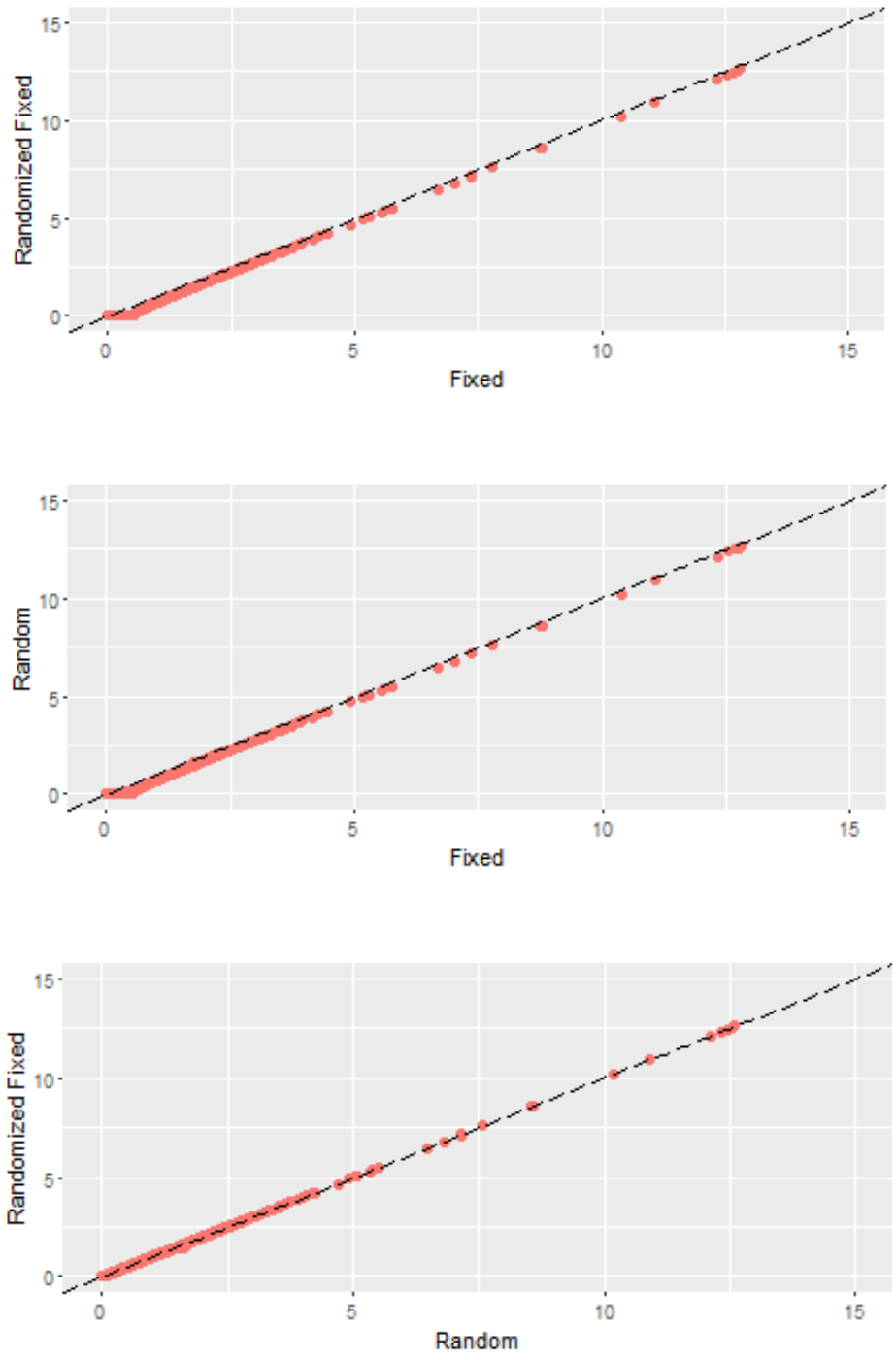


Figure 2.5 Pairwise comparisons of the -$\log_{10}$(p) among the three models: FM, RFM and RM. The GWAS was performed using the FHS data. The x axes and y axes represent the -log10(p) of the markers under the corresponding models.

Table 2.2 Critical p-values and numbers of detected SNPs under three models when performing GWAS on the FHS data

| Method | Critical p-value | Number of detected SNPs |
|---|---|---|
| FM | 2.09E-06 | 16 |
| RFM | 1.33E-05 | 21 |
| RM | 1.07E-05 | 21 |

Similar to the results in the simulation study, GWAS using the RFM had a reduced background noise compared to the FM, due to the shrinkage of the Wald statistic, as shown in figure 2.4. The effective number in the RFM was 3772, thus the critical p-value after Bonferroni correction was 1.33e-5 and the critical Wald statistic was 18.97, which was less than the critical Wald value (22.51) in the FM analysis. There were 16 markers identified using FM, and additional 5 more markers were detected using RFM. Furthermore, RFM and RM produced very similar result by comparing the middle panel and the bottom panel in figure 2.4. GWAS using RM also identified 21 associated markers with triglycerides. Besides, the 21 markers detected in RM are exactly the same as those detected in RFM. Application in both simulated data and real data demonstrated that the computational expensive RM can be substituted by our RFM approach.

### 2.3.3 Application of RFM in Framingham heart study data

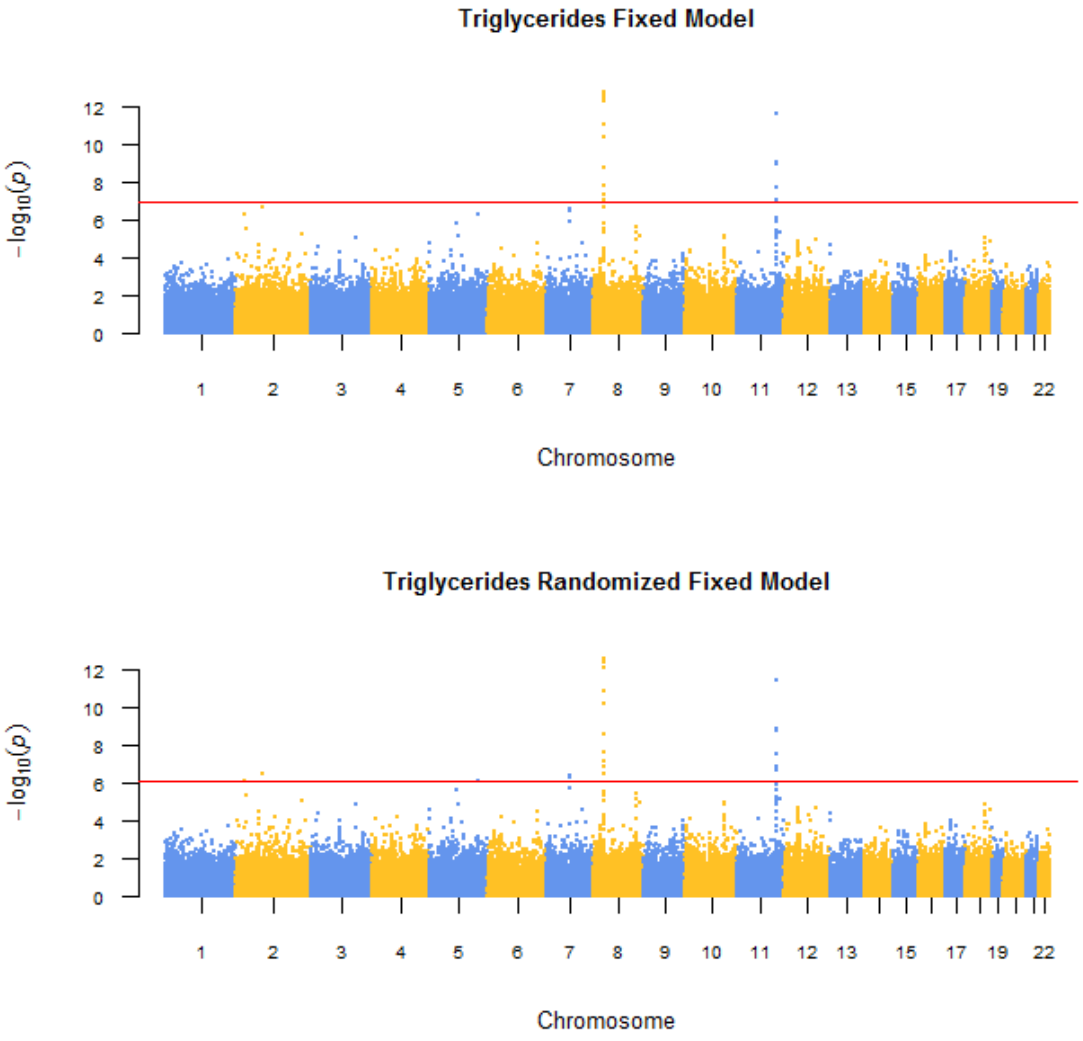


Figure 2.6 Manhattan plot of GWAS for triglycerides using FM and RFM approaches. The x axes represent the order of chromosomes and y axes represent statistic -log10(p). The horizontal red lines represent the critical values of each analysis at the significance level of 0.05.

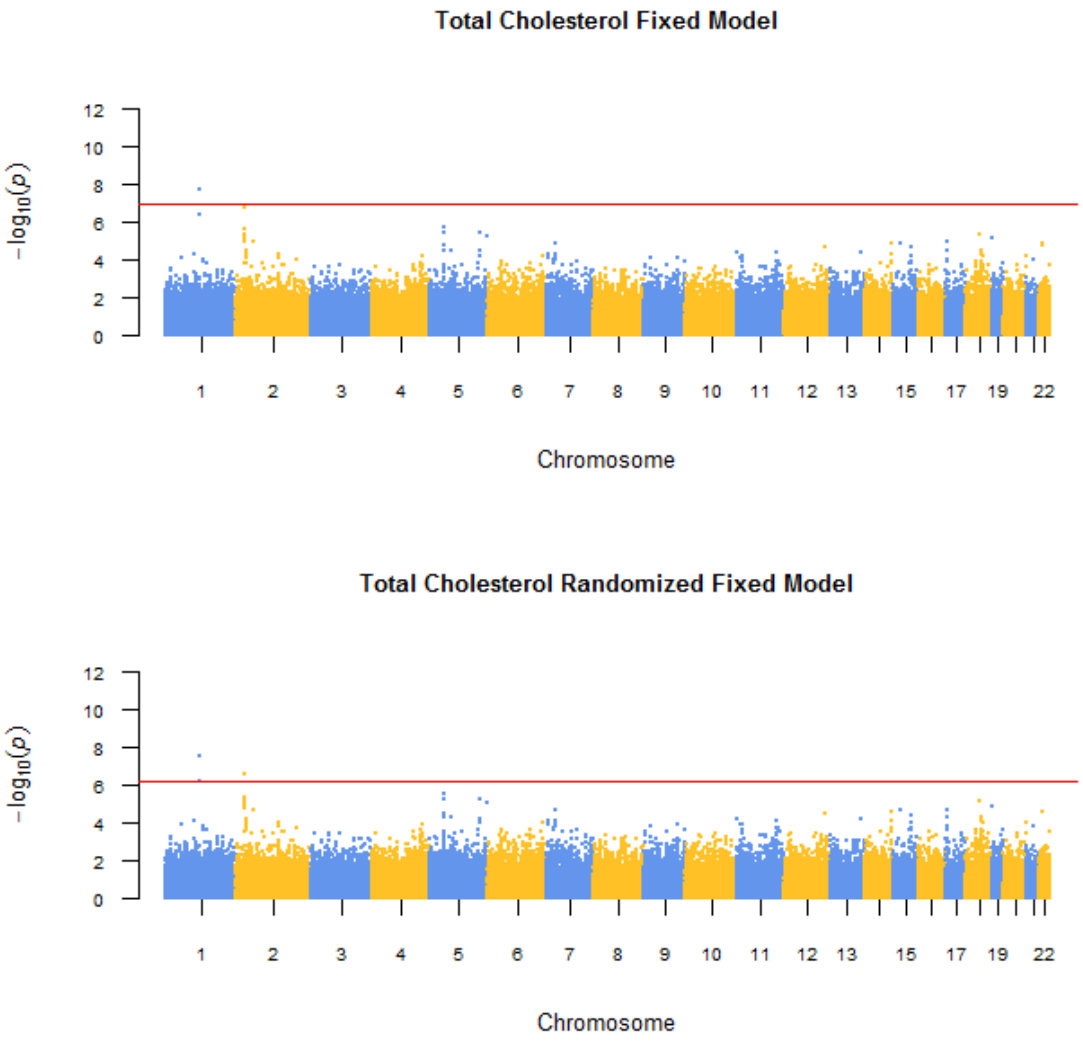


Figure 2.7 Manhattan plot of GWAS for total cholesterol using FM and RFM approaches. The x axes represent the order of chromosomes and y axes represent statistic -$\log_{10}$(p). The horizontal red lines represent the critical values of each analysis at the significance level of 0.05.

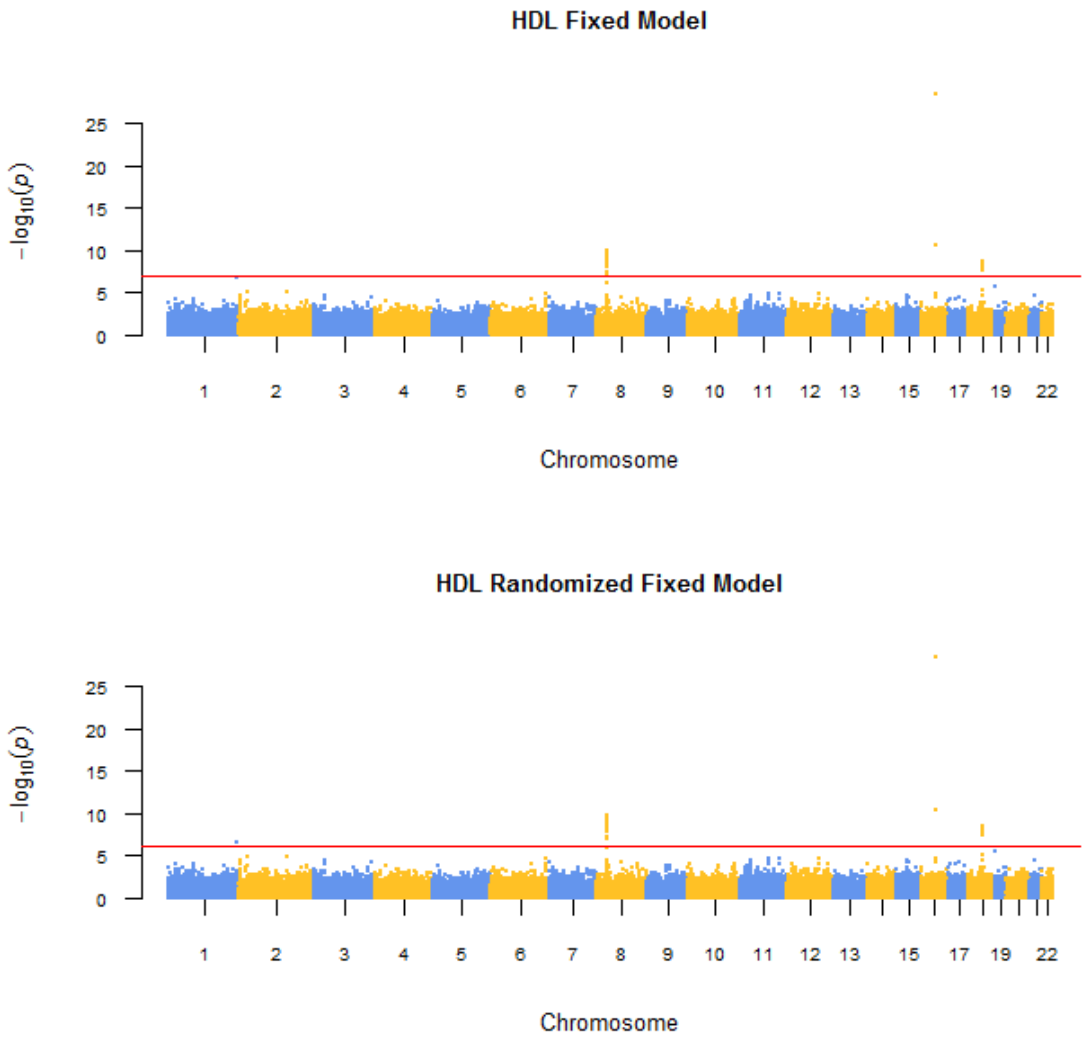


Figure 2.8 Manhattan plot of GWAS for HDL using FM and RFM approaches. The x axes represent the order of chromosomes and y axes represent statistic -$\log_{10}$(p). The horizontal red lines represent the critical values of each analysis at the significance level of 0.05.

Table 2.3 Effective number of each chromosome in three analyzed phenotypes when performing GWAS using RFM approach

| Chromosome | Total analyzed markers | Effective number | | |
|---|---|---|---|---|
| | | Triglycerides | Total cholesterol | HDL |
| 1 | 34,186 | 4,883 | 5,155 | 5,415 |
| 2 | 35,768 | 5,488 | 5,518 | 5,216 |
| 3 | 29,396 | 4,423 | 4,284 | 4,267 |
| 4 | 27,796 | 4,208 | 4,291 | 4,149 |
| 5 | 28,276 | 4,425 | 4,242 | 4,102 |
| 6 | 28,149 | 4,384 | 4,319 | 4,247 |
| 7 | 22,886 | 3,308 | 3,478 | 3,377 |
| 8 | 23,974 | 3,772 | 3,513 | 3,690 |
| 9 | 20,165 | 2,973 | 2,994 | 3,153 |
| 10 | 24,940 | 3,750 | 3,574 | 3,643 |
| 11 | 22,759 | 3,569 | 3,605 | 3,677 |
| 12 | 21,764 | 3,266 | 3,209 | 3,432 |
| 13 | 16,659 | 2,423 | 2,401 | 2,544 |
| 14 | 13,640 | 1,990 | 2,076 | 2,151 |
| 15 | 12,284 | 1,817 | 1,868 | 1,964 |
| 16 | 13,111 | 2,007 | 1,969 | 2,075 |
| 17 | 9,820 | 1,584 | 1,482 | 1,484 |
| 18 | 12,768 | 1,982 | 2,048 | 1,921 |
| 19 | 5,582 | 821 | 871 | 898 |
| 20 | 10,860 | 1,507 | 1,709 | 1,620 |
| 21 | 6,201 | 882 | 1,013 | 972 |
| 22 | 5,414 | 786 | 838 | 836 |
| Total | 426,398 | 64,248 | 64,457 | 64,833 |

Table 2.4 Numbers of detected SNPs in three phenotypes under two models

| | Number of detected SNPs | |
|---|---|---|
| | FM | RFM |
| Triglycerides | 19 | 24 |
| Total cholesterol | 1 | 3 |
| HDL | 19 | 20 |

Since it was shown that RFM was as powerful as RM, we performed GWAS using three phenotypes using RFM. First, we used GEMMA to perform GWAS using FM. Then we randomized the marker effect by subtracting one from the Wald statistic of FM, and assigned each marker a degree of confidence $d_k$ according to equation 2.13, with the limitation that $d_k \geq 0$. The total effective number for Bonferroni correction could be derived by summarizing $d_k$ across all markers. Figure 2.6, 2.7, 2.8 showed the Manhattan plots of GWAS under two models (FM and RFM) with three phenotypes: triglycerides, total cholesterol and HDL. For each phenotype, RFM was more statistical powerful than FM since that more markers were identified using RFM approach. Table 2.4 summarized the number of SNPs detected by using the two models for the three phenotypes of interest. In GWAS using RFM, there were 24 markers identified to be associated with triglycerides, 3 markers associated with total cholesterol, and 20 markers associated with HDL. Twelve markers were associated with both triglycerides and HDL, and no marker was associated with all three phenotypes.

## 2.4 Discussion

Among all associated markers in this study, most of them were reported in previous GWAS results. Besides, we detected one novel SNP (rs17005774) associated with triglycerides, and another novel SNP (rs10925994) associated with HDL. The SNP rs17005774 is located in the intron of gene TGFA (transforming growth factor alpha). It's annotated that variants in this gene are associated with body weights, obesity related traits, etc. The SNP rs10925994 is within the intron region of gene CHRM3 (cholinergic receptor, muscarinic 3), and variants of this gene are associated with traits including blood pressure, body fat distribution, body mass index, epilepsy, erythrocytes and fibrinogen. Further experiments need to be carried out to verify the association and dissect the genetic architecture of the two newly identified SNPs with coronary heart disease related traits.

When performing GWAS, the RM approach is very computationally intensive since it requires estimation of two genetic variance components (the variance of the scanned marker and the polygenic variance). However, the advantage of the RM approach is that it allows us to adopt the effective number of tests to perform the Bonferroni correction, therefore boosting the statistical power. In this study, with the objective of taking advantage of the RM approach and avoiding its disadvantage, we proposed a RFM method to perform the RM GWAS using results of the FM analysis without involving additional computation cost. In this new method, the modified Bonferroni correction

can still be used to control false positive rates more precisely to improve the statistical power.

In the simulation study, we performed the association analysis using the FM, RFM and RM, respectively. Five more QTL were identified as significant in the RFM compared to the FM. And the RFM detected the same markers as those detected in the RM approach. In analysis of the real human data, both the RFM and RM approaches identified 5 more SNPs than the FM approach. Figure 2.4 showed that RFM and RM generated very similar plots of the Wald test statistic. These results supported that we can get the RM GWAS results by using the results from the FM approach and save computational cost without losing statistical power. This RFM approach could be widely used in many GWAS research, especially those involve high-density genetic markers and a large number of individuals.

# Chapter 3

# Significance Tests Using an Outlier Detection Approach

## 3.1 Introduction

Bonferroni correction is the most widely used method for multiple hypotheses testing in GWAS. However, this method is usually somehow too conservative, since it assumes that each test is independent, leading to an increase in the type II error rate. FDR controlling (Benjamini & Hochberg, 1995) is another strategy used to correct for multiple comparisons, but it has been observed that the power will decrease with the increase of the dependency among SNPs (Sabatti, Service, & Freimer, 2003).

We assume that the SNPs that have effect on the phenotype account for only a small proportion of the entire set of genome-wide SNPs, thus, these associated SNPs can be treated as outliers. Here we propose an alternative method for multiple testing, outlier detection approach, to detect significant markers using a mixture model. A mixture model refers to a mixture distribution which describes the probability distribution of each observation in the population. To be simple, a mixture distribution is a mixture of two or more distributions. We will show that the mixture model outlier detection

approach can be a powerful alternative method for Bonferroni correction and FDR controlling in GWAS.

## 3.2 Methods

### 3.2.1 Gaussian mixture

Let $y_j$ be the *j*th observation of a target variable in the outlier detection problem. We assume that $y_j$ follows a mixture of two distributions. We assume that the majority of the observations come from one distribution and the outliers come from another distribution. The two distributions may be from the same family of distributions or from different distribution families. For simplicity, let us assume that the two distributions come from the same family of distributions, e.g., normal distribution. The simplest case is the Gaussian mixture where both distributions are normal. There are five parameters in a Gaussian mixture of two components. The first component is $N(\mu_1, \sigma_1^2)$ and the second component is $N(\mu_2, \sigma_2^2)$, each component has two parameters (mean and variance). There is another parameter $\rho$ called the mixing proportion which represents the proportion of observations coming from the first distribution. Therefore, the Gaussian mixture is described by

$$y_j \sim \rho N(\mu_1, \sigma_1^2) + (1-\rho) N(\mu_2, \sigma_2^2) \tag{3.1}$$

The probability density of such a mixture distribution is written as

$$f(y_j \mid \theta) = \rho f_1(y_j \mid \theta_1) + (1-\rho) f_2(y_j \mid \theta_2) \tag{3.2}$$

where

$$f_k(y_j \mid \theta_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[ -\frac{1}{2\sigma_k^2}(y_j - \mu_k)^2 \right] \text{ for } k = 1,2 \tag{3.3}$$

and $\theta_k = \left\{ \mu_k, \sigma_k^2 \right\}$ are parameters of the *k*th components. Parameters are estimated using the maximum likelihood methods. The finite mixture model procedure in SAS is particularly design to perform such analysis. The procedure is called PROC FMM.

After the parameters are estimated, we can calculate the posterior probability of cluster for each observation using the Bayes theorem,

$$\pi_j = \frac{\hat{\rho} f_1(y_j \mid \hat{\theta}_1)}{\hat{\rho} f_1(y_j \mid \hat{\theta}_1) + (1-\hat{\rho}) f_2(y_j \mid \hat{\theta}_2)} \tag{3.4}$$

Observation *j* is classified into the second component (the outlier cluster) if $\pi_j < 0.5$. An observation classified into the outlier group is considered to be statistically significant. In fact, it should be interpreted as statistically different from the majority of observations.

Any mixture model analysis is associated with a cluster identifiability problem. Without any constraints, PROC FMM may treat the first cluster as the outlier. To make sure that the second cluster is the outlier group, we often place the following constraint,

$\sigma_2^2 > c\sigma_1^2$, where $c$ is a positive integer arbitrarily assigned by the investigator, say $c = 100$. The larger the *c*, the smaller the outlier cluster.

### 3.2.2 Statistics used as the target variable in outlier detection

(1) Estimated marker effect

In genome-wide association studies, $y_j$ may represent the estimated effect for the *j*th SNP, denoted by $\hat{b}_j$, i.e., $y_j = \hat{b}_j$. In this case, we should incorporate the variance of the estimation $\mathrm{var}(\hat{b}_j)$. PROC FMM allows the use a WEIGHT variable to incorporate such information. The weight variable in this case is defined as,

$$W_j = \frac{1}{\mathrm{var}(\hat{b}_j)} \tag{3.5}$$

Gaussian mixture will be appropriate for this analysis.

(2) t test statistic

The target variable $y_j$ may be the t test statistic, i.e., $y_j = t_j$, where

$$t_j = \frac{\hat{b}_j}{\sqrt{\mathrm{var}(\hat{b}_j)}} \tag{3.6}$$

Gaussian mixture without a weight variable will be appropriate for the *t* variable analysis.

(3) p-value

If the $y_j$ variable is the p-value, $0 < p_j < 1$, there are two options we can try. One is directly model $p_j$ using a Beta mixture distribution, i.e., both components are Beta distributions with different parameters. Alternatively, we may perform probit transformations,

$$y_j = \Phi^{-1}(p_j) \tag{3.7}$$

and then model $y_j$ using the Gaussian mixture.

## 3.3 Results

### 3.3.1 Application in simulated data

We performed the outlier detection approach for significance test on the simulated data first. Statistics t-value, weighted marker effect and the probit transformation of p-value (probit(p)) were selected as the target variables used in the Gaussian mixture model. We also tried to use the raw p-values for a Beta mixture distribution, but the detection result was not satisfactory since that almost 300 out of a total of 426,398 markers were grouped into the outlier cluster. Thus, we used probit(p) rather than the p-value for the mixture model. Figure 3.1 to 3.3 showed the mixture distribution of probit(p), t-values and weighted marker effects in the simulated data. In all three cases, most markers followed the distribution with a small variance and only a small number of markers were

categorized into the distribution with a greater variance. The latter group is the distribution of the outlier. Table 3.1 summarized number of detected SNPs and type I error rate by using the three different statistics as the target variable. Mixture models using t-values and weighted effects were more powerful than model using probit(p). But mixture model using probit(p) had a lower type I error rate. We detected 28 markers as outlier (significant) using probit(p) with a type I error rate of 0.0055, and 38 markers were detected using t-value and weighted effect with type I error rate both at 0.011.

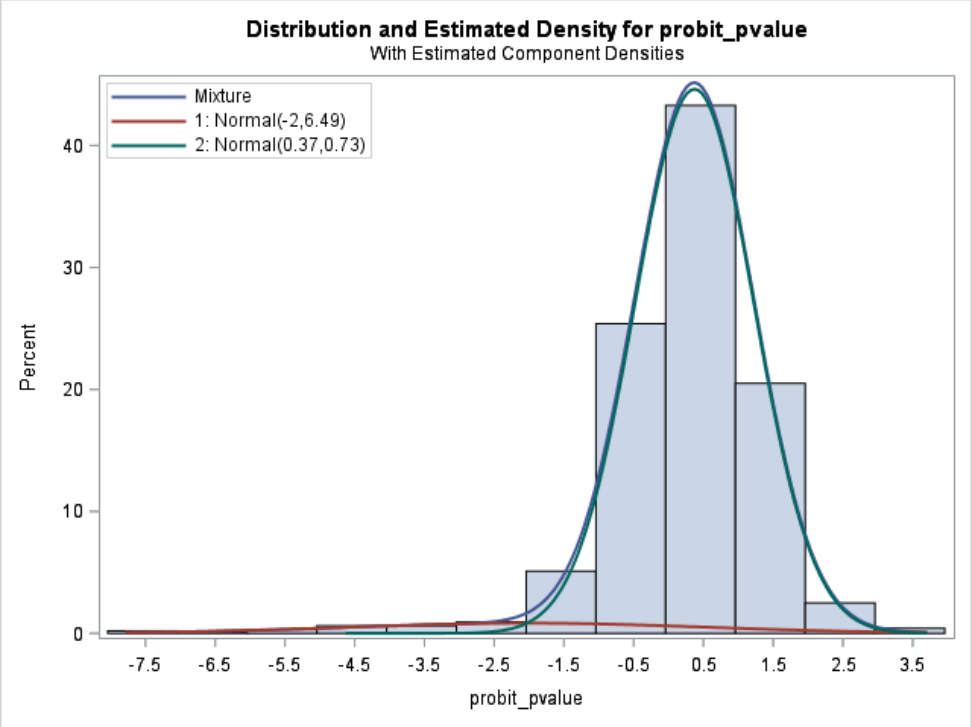


Figure 3. 1 Mixture distribution of the probit transformation of p-values of the simulated data.

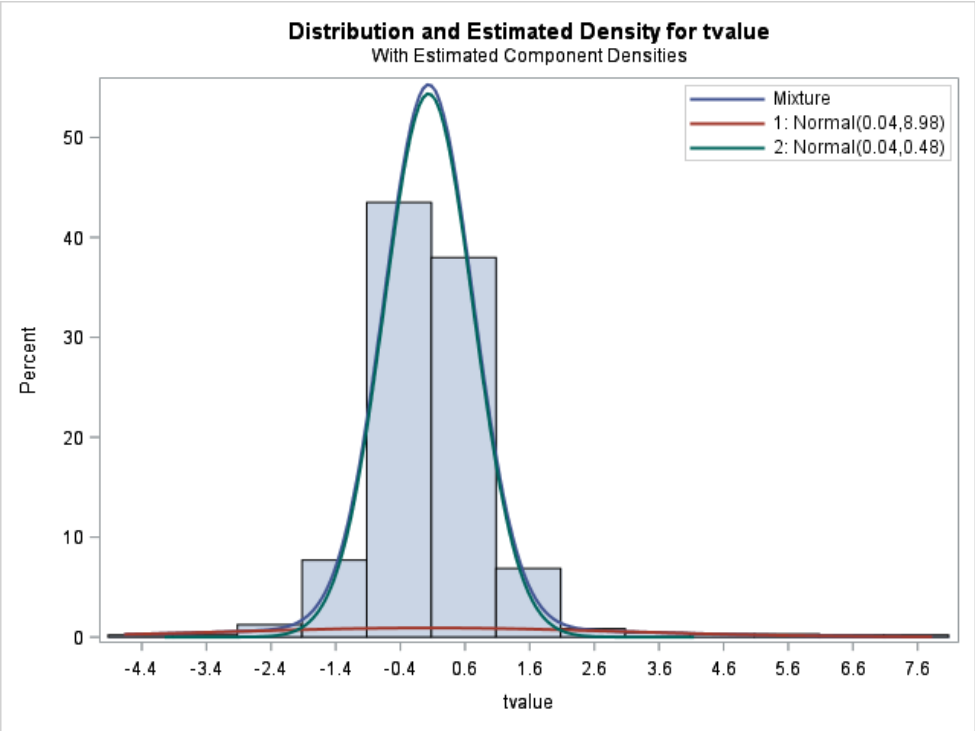


Figure 3. 2 Mixture distribution of the t-values of the simulated data.

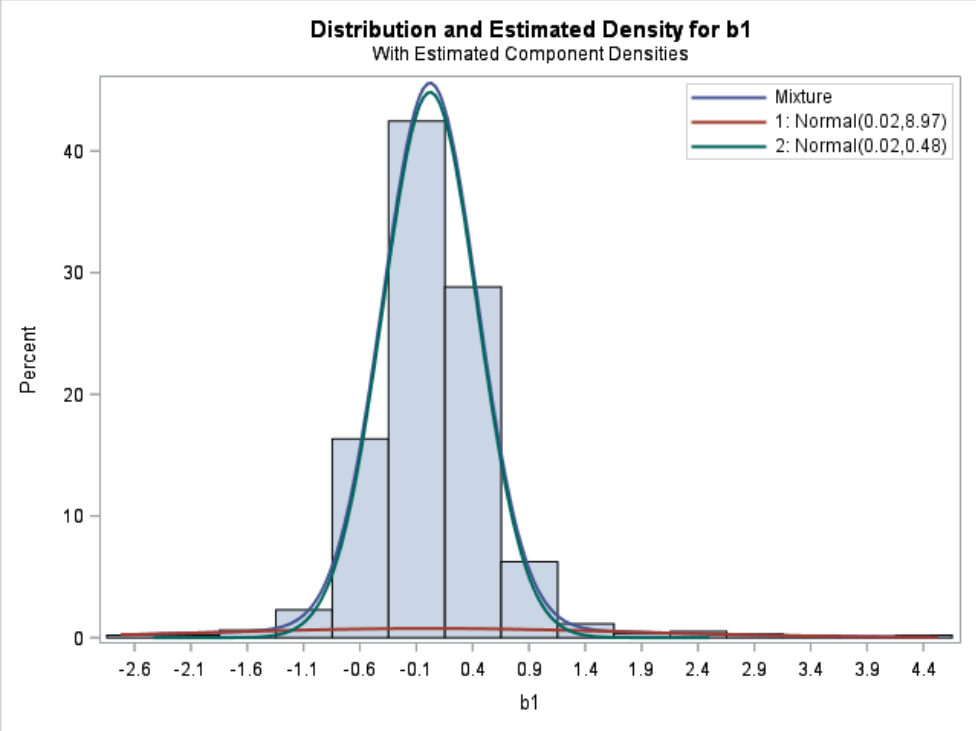


Figure 3. 3 Mixture distribution of the weighted marker effects of the simulated data.

Table 3. 1 Number of detected SNPs and type I error rate by using three different statistics as the target variable in the mixture model for outlier detection

| Variable | Number of detected SNPs | Type I error rate |
|---|---|---|
| probit(p) | 28 | 0.0055 |
| t-value | 38 | 0.011 |
| weighted effect | 38 | 0.011 |

### 3.3.2 Application in Framingham heart study data

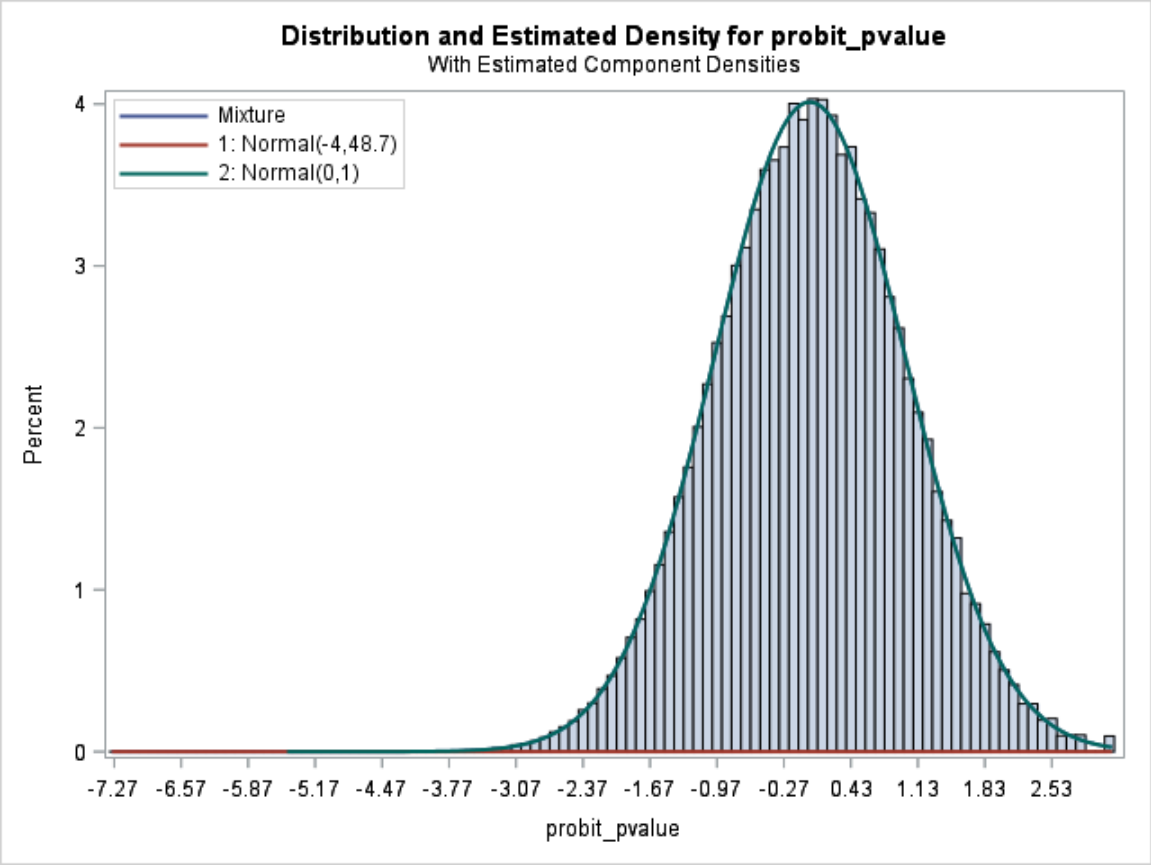


Figure 3. 4 Mixture distribution of the probit transformation of p-values from GWAS for triglycerides using the FHS data.

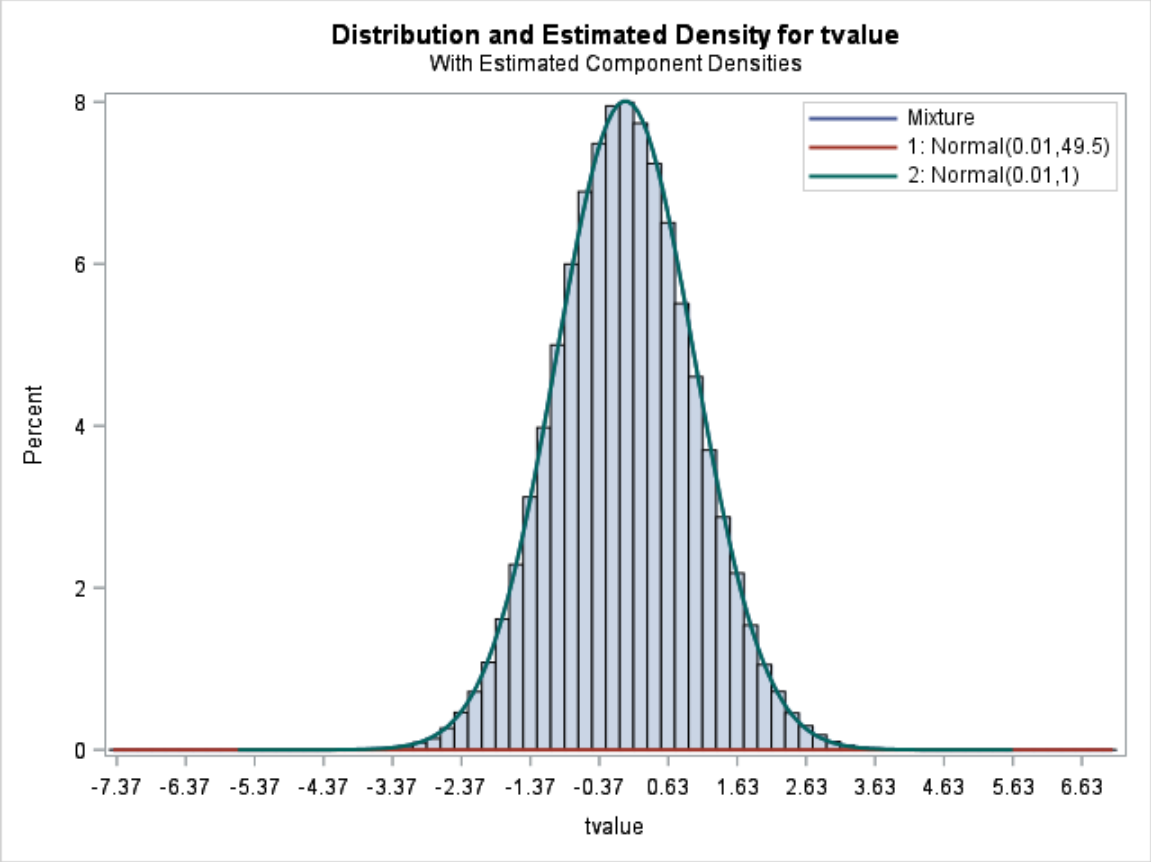


Figure 3. 5 Mixture distribution of the t-values from GWAS for triglycerides using the FHS data.

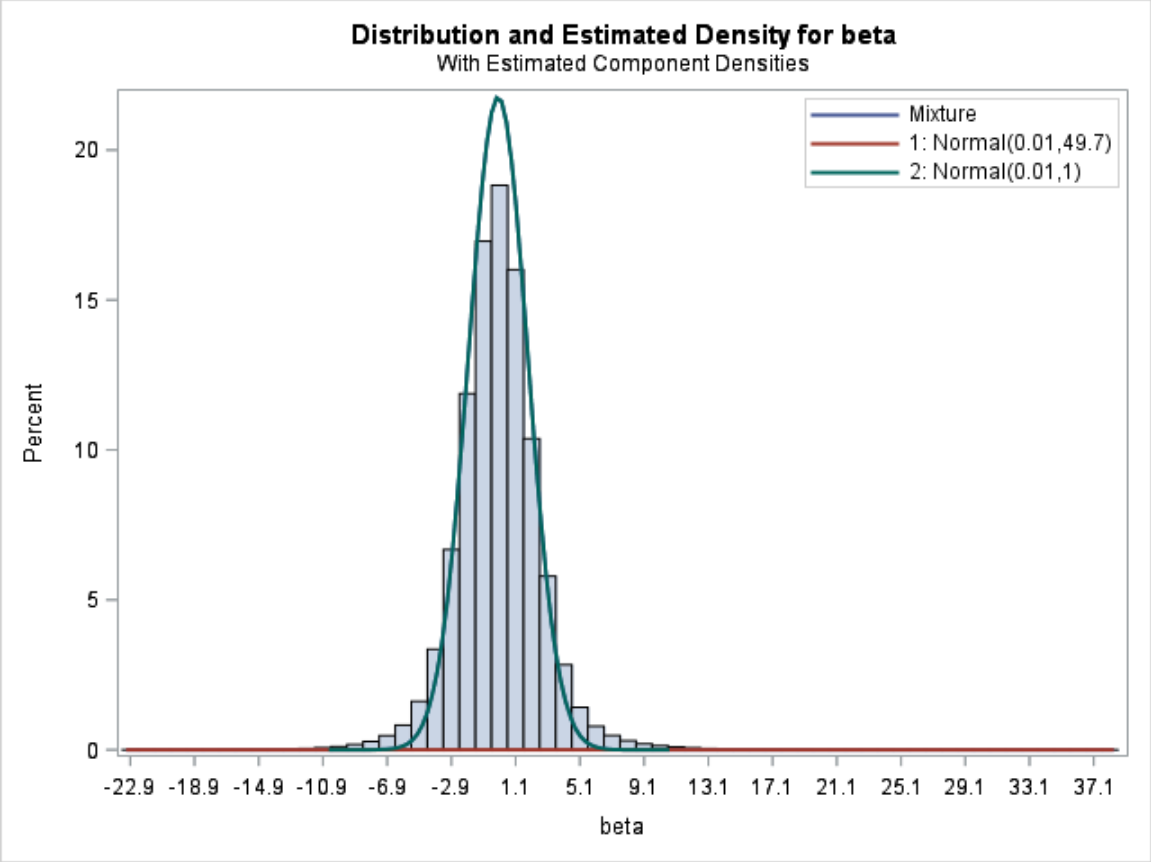


Figure 3. 6 Mixture distribution of the weighted marker effects from GWAS for triglycerides using the FHS data.

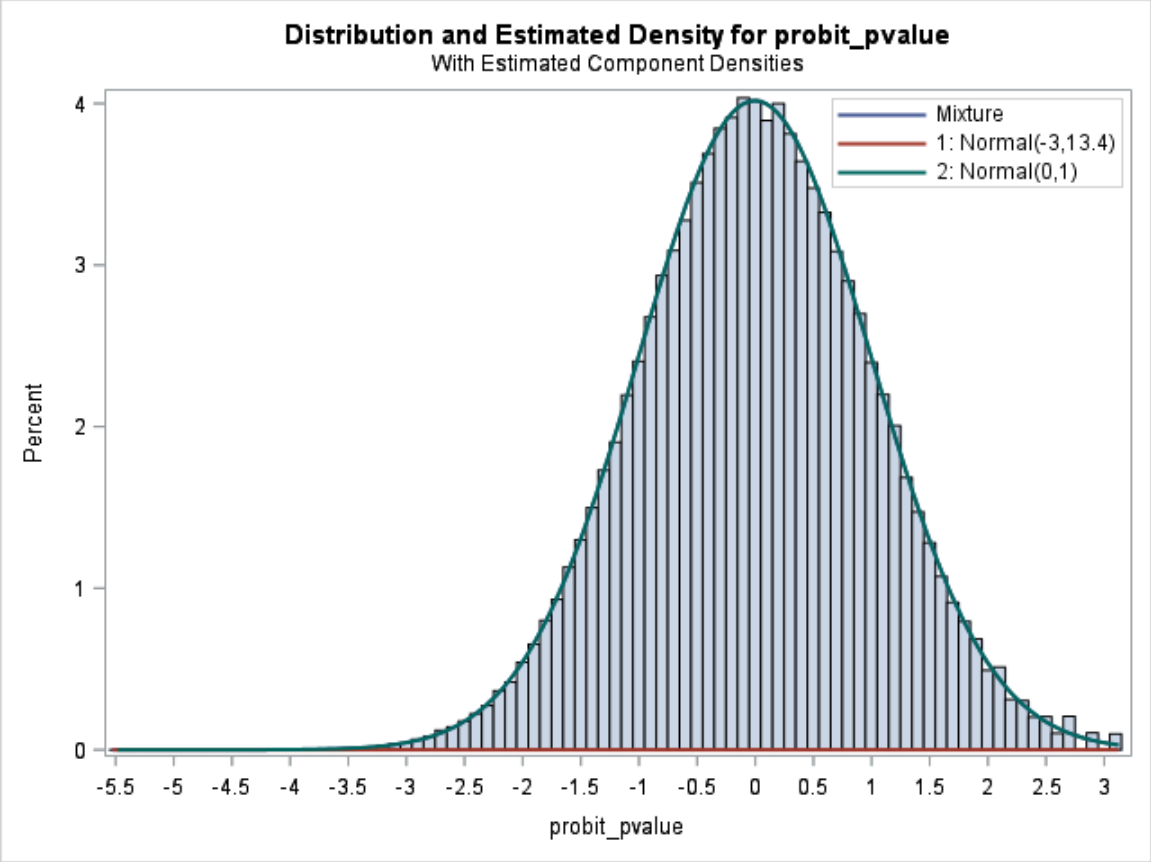


Figure 3. 7 Mixture distribution of the probit transformation of p-values from GWAS for total cholesterol using the FHS data.

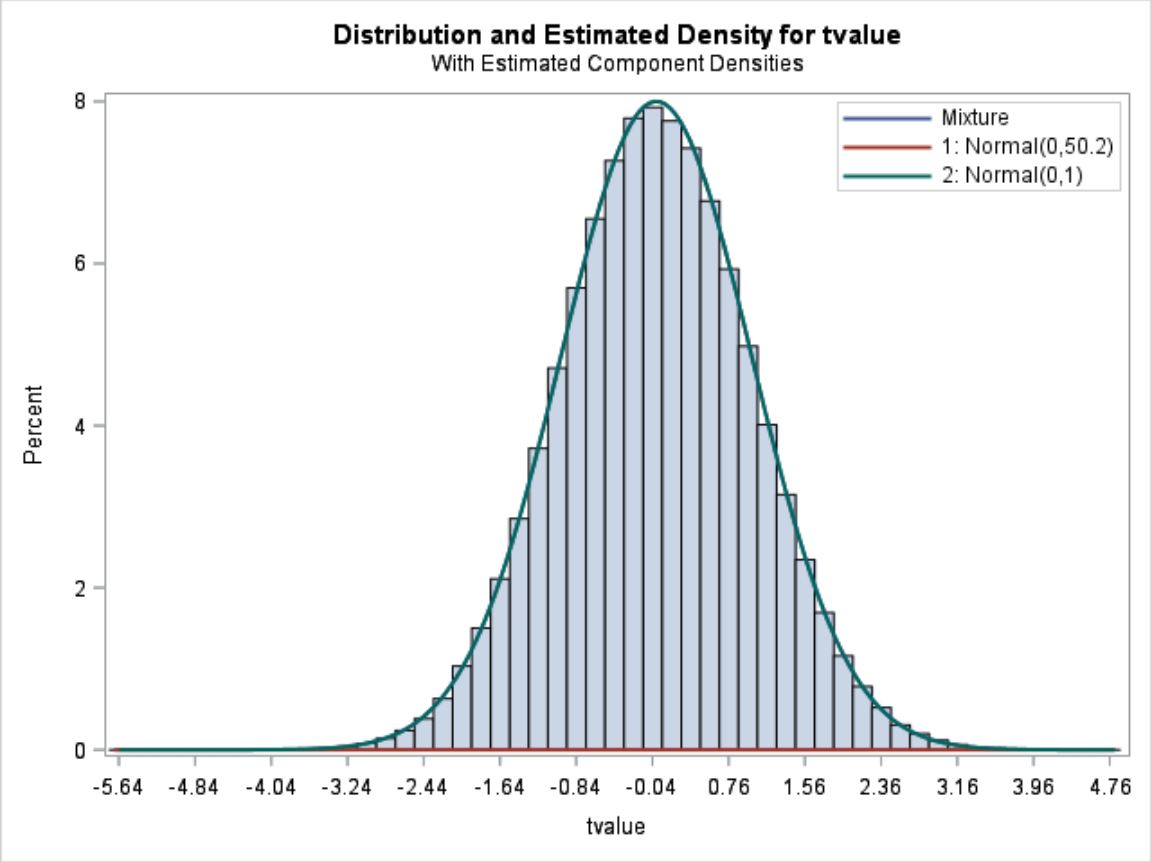


Figure 3. 8 Mixture distribution of the t-values from GWAS for total cholesterol using FHS data.

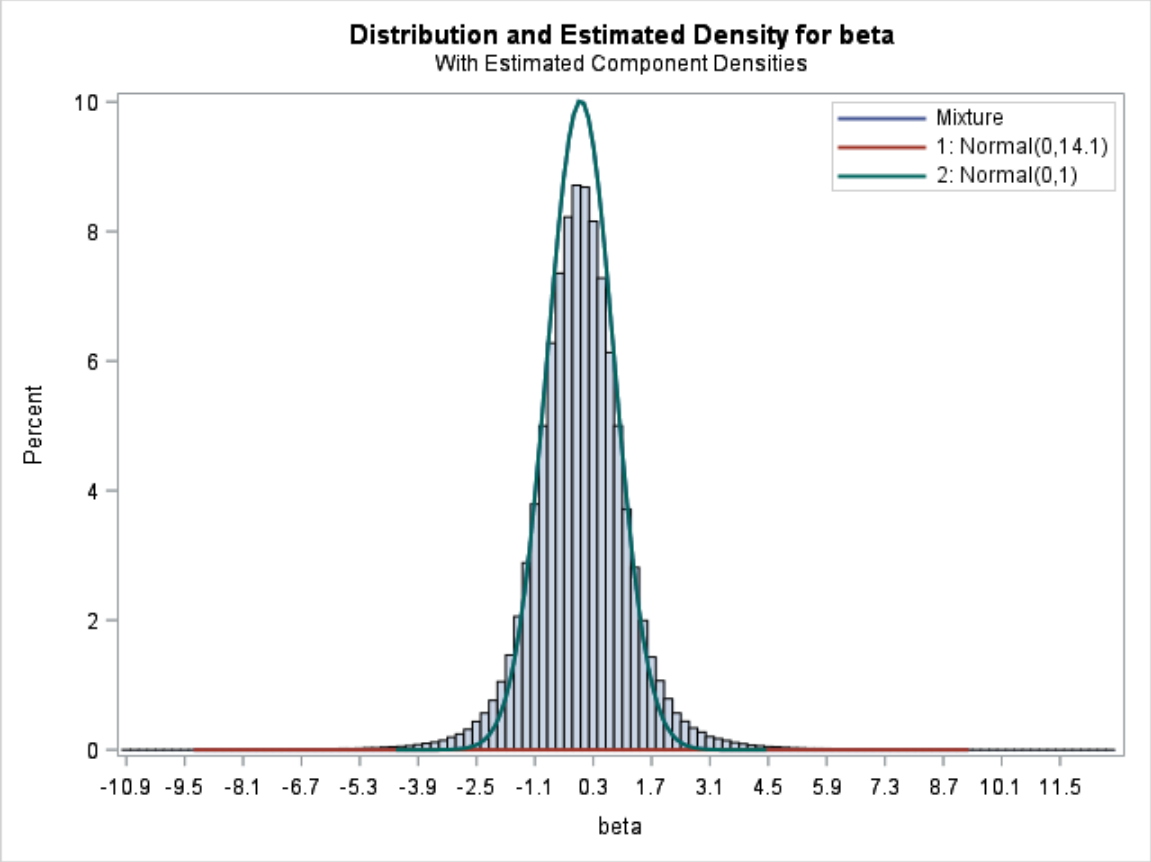


Figure 3. 9 Mixture distribution of the weighted marker effects from GWAS for total cholesterol using the FHS data.

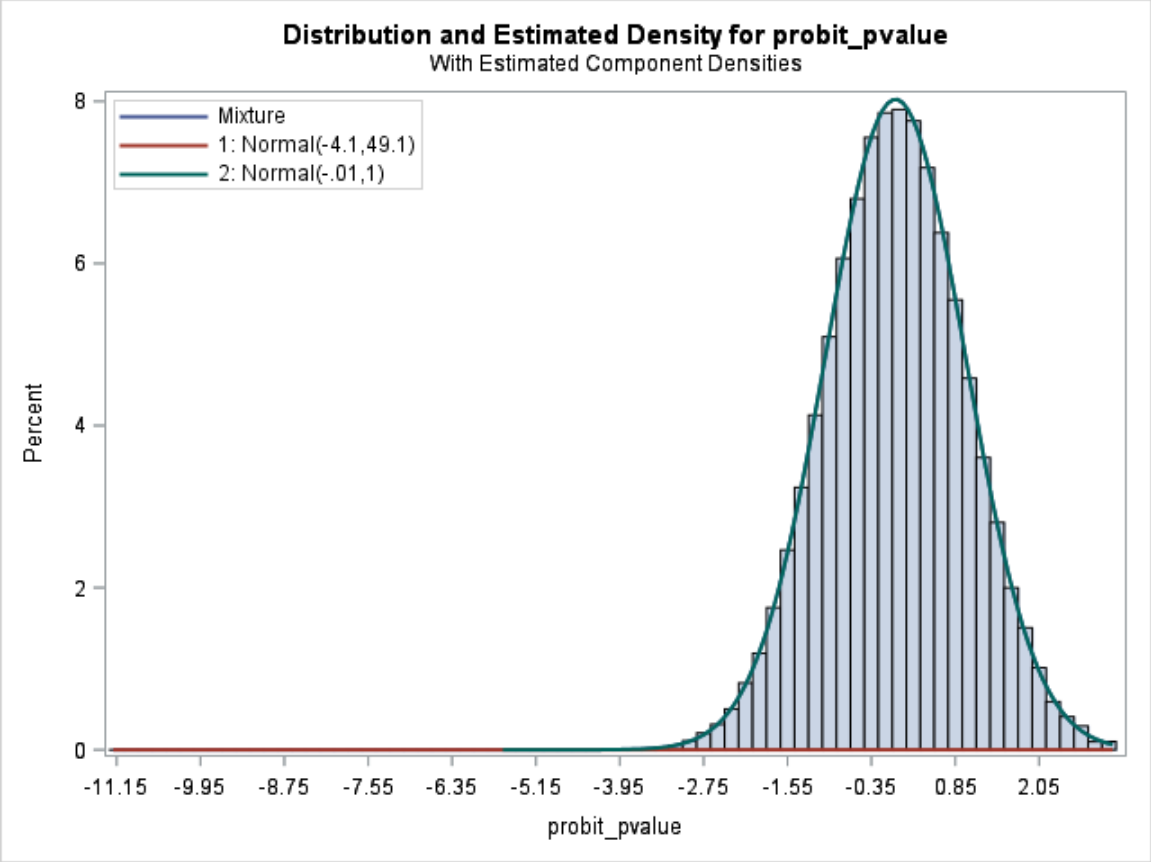


Figure 3. 10 Mixture distribution of the probit transformation of p-values from GWAS for HDL using the FHS data.

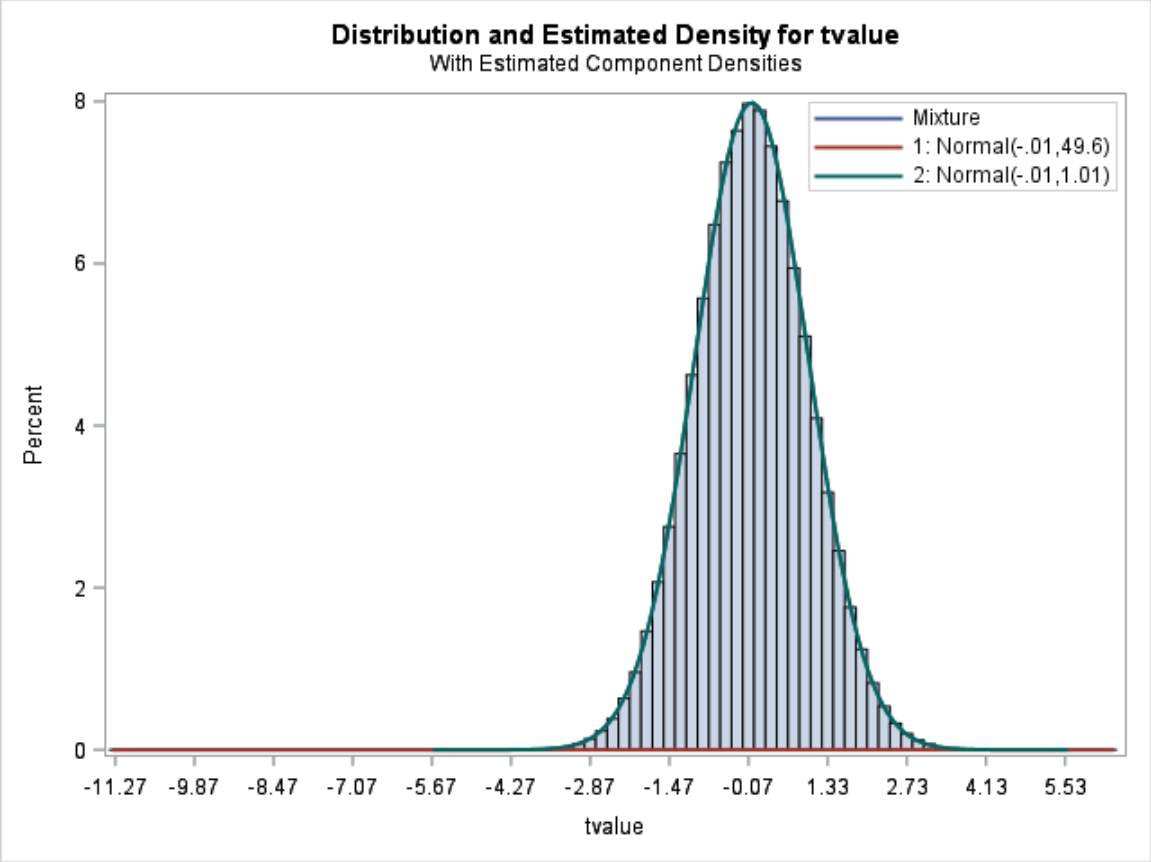


Figure 3. 11 Mixture distribution of the t-values from GWAS for HDL using the FHS data.

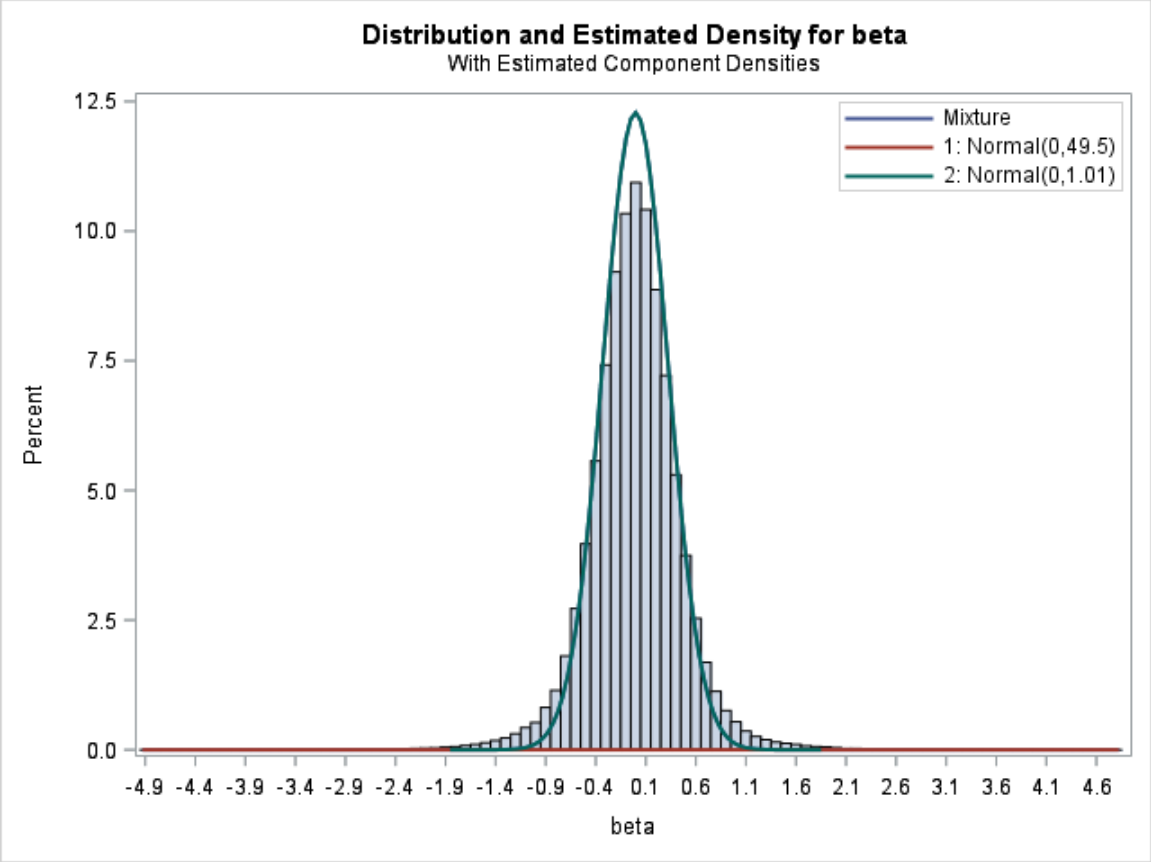


Figure 3. 12 Mixture distribution of the weighted marker effects from GWAS for HDL using the FHS data.

Table 3. 2 Comparison of number of markers detected using the outlier detection approach and Bonferroni correction method

| | Outlier detection | | | Bonferroni correction | |
|---|---|---|---|---|---|
| Phenotypes | probit(p) | t-value | weighted effect | traditional | effective number |
| Triglycerides | 27 | 33 | 33 | 19 | 24 |
| Total cholesterol | 2 | 2 | 3 | 1 | 3 |
| HDL | 22 | 22 | 22 | 19 | 20 |

Figure 3.4 to 3.12 showed the mixture distribution of probit(p), t-values and weighted marker effects with three phenotypes in the FHS data. Similarly, most markers followed the distribution with a small variance and only a small proportion of markers followed the distribution with a greater variance. Table 3.2 summarized number of detected SNPs by using outlier detection and Bonferroni correction. In general, significance test using the outlier detection approach was more powerful than the Bonferroni correction method even with the use of the effective number of tests.

## 3.4 Discussion

In GWAS and any other studies that involve multiple hypotheses testing, the choosing of the correction method is crucial. Bonferroni correction is the method that is used most widely, because it is very conservative and can guarantee a low type I error rate. However, this is always accompanied with the price of decreased statistical power and an increased type II error rate. In chapter 2 of this dissertation, we adopted the concept of effective number of tests in performing Bonferroni correction. This modified Bonferroni correction is significantly less conservative compared to the original Bonferroni correction.

In this study, we used a different strategy to realize significance testing. We did not try to set up a critical p-value for each test; rather, we considered the significantly associated markers as outliers and the whole population followed a mixture distribution. We employed probit(p), t-value and weighted marker effect as the target

variable to perform outlier detection using Gaussian mixture model in both simulated data and the FHS data.

In the simulation study, the outlier detection approach showed enhanced power compared to the result generated from Bonferroni correction either with or without using the effective number of tests, with the type I error rate equal to or less than 0.011. Then we applied this method to the FHS data to detect associated SNPs for triglycerides, total cholesterol and HDL. Similarly, the outlier detection approach detected 11 more significant SNPs than the modified Bonferroni correction method in GWAS of all three phenotypes. Overall, this method can be an alternative method for Bonferroni correction and can be widely applied to not only GWAS, but also many other studies that involve multiple testing.

# Bibliography


Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological, 57*(1), 289-300.

Berndt, S. I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M. F., . . . Ingelsson, E. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics, 45*(5), 501-U569. doi:10.1038/ng.2606

Browning, B. L. (2008). PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *Bmc Bioinformatics, 9*. doi:Artn 30910.1186/1471-2105-9-309

Chen, W., Gao, Y. Q., Xie, W. B., Gong, L., Lu, K., Wang, W. S., . . . Luo, J. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics, 46*(7), 714-721. doi:10.1038/ng.3007

Churchill, G. A., & Doerge, R. W. (1994). Empirical Threshold Values for Quantitative Trait Mapping. *Genetics, 138*(3), 963-971.

De, G., Yip, W. K., Ionita-Laza, I., & Laird, N. (2013). Rare Variant Analysis for Family-Based Design. *Plos One, 8*(1). doi:ARTN e4849510.1371/journal.pone.0048495

Devlin, B., & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics, 29*(2), 311-322. doi:10.1006/geno.1995.9003

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., . . . Ponder, B. A. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature, 447*(7148), 1087-1093. doi:10.1038/nature05887

Eeles, R. A., Kote-Jarai, Z., Giles, G. G., Al Olama, A. A., Guy, M., Jugurnauth, S. K., . . . Collaborators, U. P. S. (2008). Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genetics, 40*(3), 316-321. doi:10.1038/ng.90

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology, 54*, 357-374. doi:10.1146/annurev.arplant.54.031902.134907

Gudbjartsson, D. F., Arnar, D. O., Helgadottir, A., Gretarsdottir, S., Holm, H., Sigurdsson, A., . . . Stefansson, K. (2007). Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature, 448*(7151), 353-357. doi:10.1038/nature06007

Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L. T., Gudbjartsson, D., Helgason, A., . . . Stefansson, K. (2007). Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genetics, 39*(5), 631-637. doi:10.1038/ng1999

Gudmundsson, J., Sulem, P., Rafnar, T., Bergthorsson, J. T., Manolescu, A., Gudbjartsson, D., . . . Stefansson, K. (2008). Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nature Genetics, 40*(3), 281-283. doi:10.1038/ng.89

Haiman, C. A., Patterson, N., Freedman, M. L., Myers, S. R., Pike, M. C., Waliszewska, A., . . . Reich, D. (2007). Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genetics, 39*(5), 638-644. doi:10.1038/ng2015

Hakonarson, H., Grant, S. F. A., Bradfield, J. P., Marchand, L., Kim, C. E., Glessner, J. T., . . . Polychronakos, C. (2007). A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature, 448*(7153), 591-U597. doi:10.1038/nature06010

Hao, Y., Liu, X., Lu, X., Yang, X., Wang, L., Chen, S., . . . Gu, D. (2013). Genome-wide association study in Han Chinese identifies three novel loci for human height. *Hum Genet, 132*(6), 681-689. doi:10.1007/s00439-013-1280-9

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., . . . Stefansson, K. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science, 316*(5830), 1491-1493. doi:10.1126/science.1142842

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet, 38*(6), 226-231. doi:10.1007/BF01245622

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Huang, J., Ellinghaus, D., Franke, A., Howie, B., & Li, Y. (2012). 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet, 20*(7), 801-805. doi:10.1038/ejhg.2012.3

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., . . . Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics, 42*(11), 961-967. doi:10.1038/ng.695

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., . . . Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics, 39*(7), 870-874. doi:10.1038/ng2075

International HapMap, C. (2003). The International HapMap Project. *Nature, 426*(6968), 789-796. doi:10.1038/nature02168

Jia, G. Q., Huang, X. H., Zhi, H., Zhao, Y., Zhao, Q., Li, W. J., . . . Han, B. (2013). A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (Setaria italica). *Nature Genetics, 45*(8), 957-U167. doi:10.1038/ng.2673

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., . . . Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics, 42*(4), 348-U110. doi:10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics, 178*(3), 1709-1723. doi:10.1534/genetics.107.080101

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., . . . Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*(5720), 385-389. doi:10.1126/science.1109557

Laird, N. M., & Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics, 7*(5), 385-394. doi:10.1038/nrg1839

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and

analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Lewinger, J. P., Morrison, J. L., Thomas, D. C., Murcray, C. E., Conti, D. V., Li, D. L., & Gauderman, W. J. (2013). Efficient Two-Step Testing of Gene-Gene Interactions in Genome-Wide Association Studies. *Genetic Epidemiology, 37*(5), 440-451. doi:10.1002/gepi.21720

Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics, 49*(1), 49-67.

Lewontin, R. C., & Kojima, K. (1960). The Evolutionary Dynamics of Complex Polymorphisms. *Evolution, 14*(4), 458-472. doi:Doi 10.2307/2405995

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., . . . Yan, J. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nature Genetics, 45*(1), 43-50. doi:10.1038/ng.2484

Li, M. X., Yeung, J. M. Y., Cherny, S. S., & Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics, 131*(5), 747-756. doi:10.1007/s00439-011-1118-2

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods, 8*(10), 833-U894. doi:10.1038/Nmeth.1681

Mackay, D. J. C. (1992). Bayesian Interpolation. *Neural Computation, 4*(3), 415-447. doi:DOI 10.1162/neco.1992.4.3.415

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi:10.1038/nature08494

Mao, H. D., Wang, H. W., Liu, S. X., Li, Z., Yang, X. H., Yan, J. B., . . . Qin, F. (2015). A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. *Nature Communications, 6*. doi:ARTN 832610.1038/ncomms9326

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., . . . Cohen, J. C. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science, 316*(5830), 1488-1491. doi:10.1126/science.1142447

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., . . . Replication, D. G. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics, 44*(9), 981-+. doi:10.1038/ng.2383

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., . . . Kresovich, S. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America, 110*(2), 453-458. doi:10.1073/pnas.1215985110

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z. W., Costich, D. E., & Buckler, E. S. (2009). Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. *Plant Cell, 21*(8), 2194-2202. doi:10.1105/tpc.109.068437

Nordborg, M., & Tavare, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics, 18*(2), 83-90. doi:Doi 10.1016/S0168-9525(02)02557-X

Pahl, R., & Schafer, H. (2010). PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics, 26*(17), 2093-2100. doi:10.1093/bioinformatics/btq399

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904-909. doi:10.1038/ng1847

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*(3), 559-575. doi:10.1086/519795

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S., . . . Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics, 45*(10), 1150-1159. doi:10.1038/ng.2742

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273*(5281), 1516-1517.

Sabatti, C., Service, S., & Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics, 164*(2), 829-833.

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., . . . the Cardiogenics, C. (2007). Genomewide association analysis of coronary artery disease. *N Engl J Med, 357*(5), 443-453. doi:10.1056/NEJMoa072366

Schunkert, H., Konig, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., . . . Samani, N. J. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics, 43*(4), 333-338. doi:10.1038/ng.784

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., . . . Stefansson, K. (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nature Genetics, 39*(7), 865-869. doi:10.1038/ng2064

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., . . . Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature, 466*(7307), 707-713. doi:10.1038/nature09270

Thomas, D. (2010). Gene-environment-wide association studies: emerging approaches. *Nature Reviews Genetics, 11*(4), 259-272. doi:10.1038/nrg2764

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., . . . Stefansson, K. (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics, 41*(1), 18-24. doi:10.1038/ng.274

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., . . . Consor, W. T. C. C. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics, 39*(7), 857-864. doi:10.1038/ng2068

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. doi:10.1126/science.1058040

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y., & Zheng, Y. (2012). Genome-wide association study (GWAS) of resistance to head smut in maize. *Plant Sci, 196*, 125-131. doi:10.1016/j.plantsci.2012.08.004

Wang, Q., Wei, J., Pan, Y., & Xu, S. (2016). An efficient empirical Bayes method for genomewide association studies. *Journal of Animal Breeding and Genetics, 133*(4), 253-263. doi:10.1111/jbg.12191

Weedon, M. N., Lettre, G., Freathy, R. M., Lindgren, C. M., Voight, B. F., Perry, J. R. B., . . . Con, W. T. C. C. (2007). A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nature Genetics, 39*(10), 1245-1250. doi:10.1038/ng2121

Wei, W. H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics, 15*(11), 722-733. doi:10.1038/nrg3747

Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), 661-678. doi:10.1038/nature05911

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006. doi:10.1093/nar/gkt1229

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., . . . Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics, 48*(8), 927-+. doi:10.1038/ng.3596

Yu, J. M., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., . . . Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics, 38*(2), 203-208. doi:10.1038/ng1702

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., . . . Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics, 40*(5), 638-645. doi:10.1038/ng.120

Zhang, Z. W., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., . . . Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics, 42*(4), 355-U118. doi:10.1038/ng.546

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., . . . McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature Communications, 2*. doi:ARTN 46710.1038/ncomms1467

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics, 44*(7), 821-U136. doi:10.1038/ng.2310

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America, 109*(4), 1193-1198. doi:10.1073/pnas.1119675109