STATISTICAL METHODS FOR DETECTING GENETIC RISK FACTORS OF A DISEASE WITH APPLICATIONS TO GENOME-WIDE ASSOCIATION STUDIES

A THESIS SUBMITTED TO THE UNIVERSITY OF KENT FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN THE SUBJECT OF STATISTICS

BY
FADHAA ALI

March 30, 2015

Acknowledgment

I would like to thank ministry of higher education and scientific research of Iraq (MOHESR) for funding this project. I would also like to thank Prof. Dankmar Böhning- University of Southampton and Dr. Fabrizio Leisen- University of Kent for accepting assessing my thesis and giving me a great feedback to improve it's presentation.

I would like to thank many people in the School of Mathematics, Statistics and Actuarial sciences (SMSAS) at the University of Kent. Firstly, I would like to express my deepest gratefulness to my supervisor Prof. Jian Zhang for all the guidance that I received from him regarding my progress and for the positive feedback that I used to get from him. His advices took a significant part in enhancing my knowledge about this exciting research area and introduced me to all the drawbacks of my research that I might come across in the first place. I would also to thank my second advisor Dr. Xue Wang for reviewing my progress at some point and giving me great advices toward completion of my thesis.

During my Ph.D. study, I took valuable advantages from the magnificent Statistics Group in SMSAS and experienced the scientific atmosphere, which were reflected by the great seminar series on very wide range of statistical research areas. I would also like to express my warm thanks for all the Ph.D. students that I met within my time for their organising the wonderful students seminars that gave me a great opportunity to practice my presentation skills in several occasions.

Last but not the least, I would thank all the school administration officers for their unlimited support by providing me a great study environment. A special thank to the hard working lady whose name Claire Carter for her unlimited help since I started my Ph.D. in the School.

Lastly, I also feel very grateful to my family and all my dearest friends for all their supports and their caring about me in my very stressful time of my study.

Abstract

This thesis aims to develop various statistical methods for analysing the data derived from genome wide association studies (GWAS). The GWAS often involves genotyping individual human genetic variation, using high-throughput genome-wide single nucleotide polymorphism (SNP) arrays, in thousands of individuals and testing for association between those variants and a given disease under the assumption of common disease/common variant. Although GWAS have identified many potential genetic factors in the genome that affect the risks to complex diseases, there is still much of the genetic heritability that remains unexplained. The power of detecting new genetic risk variants can be improved by considering multiple genetic variants simultaneously with novel statistical methods. Improving the analysis of the GWAS data has received much attention from statisticians and other scientific researchers over the past decade.

There are several challenges arising in analysing the GWAS data. First, determining the risk SNPs might be difficult due to non-random correlation between SNPs that can inflate type I and II errors in statistical inference. When a group of SNPs are considered together in the context of haplotypes/genotypes, the distribution of the haplotypes/genotypes is sparse, which makes it difficult to detect risk haplotypes/genotypes in terms of disease penetrance.

In this work, we proposed four new methods to identify risk haplotypes/genotypes based on their frequency differences between cases and controls. To evaluate the performances of our methods, we simulated datasets under wide range of scenarios according to both retrospective and prospective designs.

In the first method, we first reconstruct haplotypes by using unphased genotypes, followed by clustering and thresholding the inferred haplotypes into risk and non-risk groups with a two-component binomial-mixture model. In the method, the parameters were estimated by using the modified Expectation-Maximization algorithm, where the maximisation step was replaced the posterior sampling of the component parameters. We also elucidated the relationships between risk and non-risk haplotypes under different modes of inheritance and genotypic relative risk.

In the second method, we fitted a three-component mixture model to genotype data directly, followed by an odds-ratio thresholding.

In the third method, we combined the existing haplotype reconstruction software PHASE and permutation method to infer risk haplotypes.

In the fourth method, we proposed a new way to score the genotypes by clustering and combined it with a logistic regression approach to infer risk haplotypes.

The simulation studies showed that the first three methods outperformed the multiple testing method of (Zhu, 2010) in terms of average specificity and sensitivity (AVSS) in all scenarios considered. The logistic regression methods also outperformed the standard logistic regression method.

We applied our methods to two GWAS datasets on coronary artery disease (CAD) and hypertension (HT), detecting several new risk haplotypes and recovering a number of the existing disease-associated genetic variants in the literature.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Genetic problems

In the past decades, much attention has been paid to complex diseases such as coronary artery disease (CAD) and hypertension (HT), which are potentially caused by both genetic and environmental factors. The genetic factors are often attributed to genetic variants (or polymorphisms), the sites where genetic alleles are varying across individual genomes. It is well-known that although genetic variants can have undetectable marginal effects on the risk of a complex disease, they may have a significant effect as a group due to interactions between variants. Therefore, simultaneously considering disease-associated variants can help us detect risk variants and develop medicines for curing the diseases. Multi-locus genotypes and haplotypes are commonly used to account for the interactions among the variants. Before the completion of Human Genome Project, due to the limitation of biological technology, researchers were only able to focus their studies on a small proportion of regions in genome. After the completion of Human Genome Project, with developments in *single-nucleotide polymorphism (SNP)* genotyping technology, genome-wide association studies (GWAS) have become a feasible and powerful approach to uncover genetic variants with much better resolution by examining hundreds of thousands of SNPs distributed across the whole genome. Most of the existing genome-wide association studies are based on the hypothesis of common disease/common variant (CDCV). Despite the number of genetic variants identified, a large proportion of heritability has not been explained. Rare variants are believed to play an important role in the missing heritability. Studying rare variants and pinpointing the causal alleles accurately provide both opportunities and challenges to modern statistics.

1.2 Statistical challenges

One of early GWAS projects is the *Welcome Trust Case Control Consortium* (WTCCC) study on seven complex diseases including CAD and hypertension. In the project,

more than 5×10^5 SNPs of more than 14000 unrelated cases (diseased individuals) and 3000 unrelated shared controls (non-diseased individuals) were employed to find disease-associated variants for each disease. Significant progress has been made recently on analysing the WTCCC data (e.g., Kang et al., 2008; Zhu et al, 2010). Various statistical screenings based on odd ratio, logistic regressions, χ^2 and Fisher's exact tests have been conducted to identify the disease-associated SNPs of several complex diseases (Weir, 2005; Zhu et al., 2010; Burton et al., 2007).

Despite of this progress, analysing SNP data still faces a number of challenges related to the problems of missing data, low minor allele frequency, long distance correlation between SNPs, multiple test adjustments, and population substructures, data quality control, and among others. For example, in fitting a logistic regression model to genotypes, a large number of degrees of freedom will be involved, which may cause model over fitting. In contrast, fitting a logistic model to the corresponding hyplotypes is better because haplotypes have a lower dimensionality than genotypes. Unfortunately, as the haplotypes cannot be observed directly, they are required to infer from the unphased genotype data (Stephen, 2001). There are several software which can be used to reconstruct haplotypes from the unphased genotypes such as PHASE (Stephen, 2001) and the Expectation-Maximization (EM) (Excoffier and Slatkin, 1995). The disadvantage of using inferred haplotypes is uncertainty associated with the above haplotype reconstructing process. In fact, the uncertainty may result in underestimating the variation in the data, inflating the type I error.

A closely related issue is the sparsity of genotype/haplotype distributions and high-dimensionality of genotypes/haplotypes, where the counts are often concentrated on a few ones out of a large number of genotypes/haplotyeps. To address the issue, researchers have proposed variety of clustering-based methods. In these methods, haplotypes/genotypes are divided into several subgroups or clusters based on their association with a disease. We assume that the haplotypes/genotypes within the same subgroup have the same risk probability of random effects (Templeton et al., 1987; Molitor et al., 2003; Zhu et al., 2010; Morris, 2005; 2006). Such methods are usually implemented via two or more stages: In the first stage, haplotypes or genotypes are grouped, while in the second stage the risk haplotypes are detected by using various test statistics, such as Z-tests and odd ratio tests. An alternative way is to fit a logistic regression model to clustered haplotypes rather than SNPs (Huang et al., 2011).

Modern statistics faces many challenges in analysing the GWAS data, which are summarised as follows.

1. Finding the causal SNPs of a disease can be difficult as these SNPs may be highly correlated with each other. This means even if we find that a SNP is associated with a disease, the risk may come from other SNPs nearby.

2. Considering multiple-SNP regions can give rise to the problem of high dimensionality of genotypes/haplotypes. Yet, inference on risk variants may be difficult and inaccurate due to rarity of some haplotypes/genotypes.

3. Haplotype-structures are unknown in the real data as we observed SNP data in terms of genotypes only. Therefore, inferring their structures by using statistical methodology such as PHASE may result in reconstructing uncertainty as pointed out before.

4. Non-diseased individuals (controls sample) in the real data come from different sub-populations, which will increase the false positive rate. Moreover, some of the SNP genotypes are of very low frequencies. As a result, significant differences between genotypes counts derived from different sub-populations.

5. Mode of inheritance can result in an increase of false discovery rate. For example, a dominant mode can inflate type I error and recessive model can inflate type II error when the genotype relative risk (GRR) or the sample size is small.

1.3 Contributions of the thesis

This thesis aims to address the above challenges by considering a group of SNPs simultaneously. I will find the evidences about their associations with a disease of interest on the basis of both haplotypes and genotypes.

The contributions of this thesis are as follows: (1) I develop prospective mixture models for clustering haplotypes and genotypes and for identifying risk haplotypes. (2) I propose a new logistic regression model for genotypes. (3) I put forward a permutation-based approach for identifying risk haplotypes. A large simulation studies have been conducted for the above methods and models under both prospective and retrospective settings. The proposed methods and models have been applied to the WTCCC data on CAD and hypertension, identifying a few more disease-associated haplotypes than in the literature.

1.4 Arrangement of the thesis

In Chapter 2, some genetic and statistical background are introduced and a literature review on the topic is conducted. In Chapter 3, a novel two-component haplotype mixture model is proposed for clustering haplotypes and is applied to the simulated and real data sets on CAD and hypertension (HT). In Chapter 4, a novel three-component genotype mixture model is developed for detecting disease-associated haplotypes with applications to the simulated and real data sets. Two papers based on Chapters 3 and 4 have been submitted to two journals. In Chapter 5, a new permutation-based method is introduced and illustrated by applications to the simulated and real data sets. In Chapter 6, a new logistic regression model is proposed and evaluated by its applications to the simulated and real data sets. In Chapter 7, a conclusion is made for this thesis. In particular, a brief discussion on the quality of the results and on the advantages and disadvantages of the proposed methods are presented. The potential future work is also pointed out.

2. BACKGROUND AND LITERATURE REVIEW

Population association studies are a key tool in determining genetic variants which affect the susceptibility to a complex disease. These variants can be produced by genetic drift, natural selection, mutation and recombination. The difference of an allele at a variant site in population groups reflects its associations with certain human trait. The higher the difference is, the more likely the genetic variant is associated with the trait.

Here, we focus on SNPs and their relationships to a disease and develop some statistical methods for this purpose. I start with an introduction to the background, followed by a literature review on the existing methods.

2.1 Single-Nucleotide Polymorphism (SNP)

Since a long time ago, geneticists have used phenotypes, protein sequencing, electrophoresis and microsatellites in order to detect the genetic differences across individuals in terms of the deoxyribonucleic acid (DNA) sequences. After new biotechnologies such as Microarrary Gene Chip being invented, detecting single-base differences has become possible in experiments under various biological conditions or different phenotypes (Kwok, 2003). Since then, the term of single-base differences in DNA amongst individuals has been known as single-nucleotide polymorphism (SNP) (and pronoun snip), see Figure (1.2). Each gene (or segment) in an entire DNA sequences often contains multiple SNPs.

As probable as it may seem, it has been proved beyond doubt that variations in genes may contribute to certain diseases. This can be seen from two aspects of genetic variants. Firstly, any disorder in DNA sequence can result in differences in gene regulations which in turn result in some diseases. Secondly, provided that some of chromosomes' genes are coded for specific proteins, any variations in their SNP alleles may cause differences in their functions and their expressions (Balding et al., 2007).

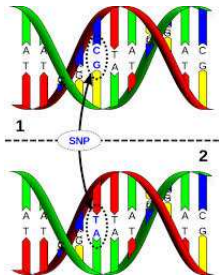


Fig. 2.1: The image from *http://www.dnabaser.com/articles/SNP/SNP-single-nucleotide-polymorphism.html* shows a segment of diploid DNA with one SNP

Tab. 2.1: Different sequences of two DNA segments of five individuals at the same positions on a chromosome pair. The segments comprise three SNPs at three loci coloured by different colours. Each SNP involves two alleles which are vary across individuals. The sequence of the three alleles at the same segment is called haplotype. Each pair of haplotypes is called genotype.

Individuals	chromosomes	DNA segment with 3 loci	haplotypes	genotypes
Individual 1	chromosome 1	...ACCTGTAATCGGGTCA...	TAT	genotype 1
	chromosome 2	...ACCGGTATTCGGGCCA...	GTC	
Individual 2	chromosome 1	...ACCTGTAATCGGGTCA...	TAT	genotype 2
	chromosome 2	...ACCTGTATTCGGGTCA...	TTT	
Individual 3	chromosome 1	...ACCGGTATTCGGGCCA...	GTC	genotype 3
	chromosome 2	...ACCGGTATTCGGGCCA...	GTC	
Individual 4	chromosome 1	...ACCTGTAATCGGGTCA...	TAT	genotype 4
	chromosome 2	...ACCTGTAATCGGGTCA...	TAT	
Individual 5	chromosome 1	...ACCTGTAATCGGGTCA...	TAT	genotype 5
	chromosome 2	...ACCGGTATTCGGGTCA...	GTT	

The position of a SNP on a chromosome is so called locus (plural loci). Most of which are based on pair of alleles within a diploid chromosome. Each pair is so called genotype, whereas each allele is so called haplotype. The extension of the notion leads to multi-locus studies by which more than one locus is conducted for a study. A commonly way for detecting SNPs (or haplotypes) underlying a particular disease is to consider groups of individuals under different conditions or traits (Weir, 1996).

In the following, I will introduce some basic ideas on population genetics.

2.1.1 Genotype/haplotype frequencies and their estimation

Suppose that we have a sample of genotype observations on a binary locus with alleles C and T for N individuals in a sample with counts n_1, n_2 and population frequencies p_1, p_2, respectively. The possible genotypes for this locus will be CC, CT and TT. Let the genotype counts denoted by N_1, N_2, N_3 with population frequencies q_1, q_2, q_3. The counts of these genotypes are random variables following multinomial distribution (Weir, 1996).

$$P_r(N_1, N_2, N_3) = \frac{N!}{N_1!N_2!N_3!} q_1^{N_1} q_2^{N_2} q_3^{N_3},$$

whereas the distribution of alleles C and T is binomial, which can be written as

$$P_r(n_1, n_2) = \frac{(2N)!}{n_1!n_2!} p_1^{n_1} p_2^{n_2}.$$

We can then derive the relationship of genotypes and haplotypes counts(or frequencies) as follows:

$$n_1 = 2N_1 + N_2 \text{ and } n_2 = 2N_3 + N_2, \tag{2.1}$$

$$N = N_1 + N_2 + N_3,\ n_1 + n_2 = 2N.$$

Similarly,

$$p_1 = q_1 + \frac{1}{2}q_2,\ p_2 = q_3 + \frac{1}{2}q_2, \tag{2.2}$$

where

$$q_1 + q_2 + q_3 = p_1 + p_2 = 1.$$

However, in many cases, the population frequencies of genotypes as well as their alleles are unknown, so we may use their sample counts to estimate them by

$$\hat{p}_1 = \frac{n_1}{2N} \text{ and } \hat{q}_1 = \frac{N_1}{N}. \tag{2.3}$$

The above model and estimation can be easily extended to the case of multiple-locus genotypes/haplotypes, where a multinomial model is required.

Many other methods have been introduced to estimate population frequencies such as Maximum Likelihood Estimation (Weir, 1996; Excoffier and Slatkin, 1995; McLachlan and Peel, 2003). We will use some of them in the upcoming chapters.

2.1.2 Hardy-Weinberg Equilibrium

The relationship between the frequencies of the genotypes and their haplotypes is important as far as association with diseases is concerned. Therefore, Hardy (1877-1947) and Weinberg (1862-1937) independently formulated what is now called Hardy-Weinberg model in 1908. The mathematical relationship between the frequencies of the single-locus genotypes and haplotypes can be described by:

$$q_1 = p_1^2; q_2 = 2p_1p_2; q_3 = p_2^2,$$

for a given SNP on a chromosome with two alleles C and T. Hardy-Weinberg Equilibrium (HWE) can be applicable only when some assumptions hold in the population (Hartl and Clark, 1997). These assumptions include that (1) the individuals under study should be diploid; (2) no overlapping exists amongst generations; (3) mutation is not important; (4) alleles under study are not affected by natural selection; (5) migration is trivial; (6) the size of population is large; (7) individuals are mated randomly; and (8) SNPs are biallelic (involve two alleles only). If any of these assumptions is not true, we then say the population under Hardy-Weinberg disequilibrium, by which the alleles of a particular locus or the haplotypes of multi-locus regions are not under random mating (Weir, 1996). The mathematical relationship between genotypes and alleles frequencies is given by

$$q_1 = p_1^2 + D_C; q_2 = 2p_1p_2 - 2D_C; q_3 = p_2^2 + D_C,$$

which implies

$$D_C = q_1 - p_1^2,$$

where D_C is called disequilibrium coefficient. The MLE of D_C is

$$\hat{D}_C = \hat{q}_1 - \hat{p}_1^2.$$

The expected value of this estimation can be calculated by the formulas

$$E(\hat{D}_C) = D_C + \frac{1}{2N}[p_1(1-p_1) + D_C],$$

and

$$Var(\hat{D}_C) = \frac{1}{N}[p_1^2(1-p_1)^2 + (1-2p_1)^2 D_C - D_C^2].$$

The latter formula can be simplified by using Fisher variance approximation

$$\hat{D}_C \sim N(E(\hat{D}_C), Var(\hat{D}_C)),$$

provided the sample size is relatively large. Hence, under the null hypothesis H_0 : $D_C = 0$, the following statistics

$$z = \frac{\hat{D}_C - E(\hat{D}_C)}{\sqrt{Var(\hat{D}_C)}}$$

is distributed as standard normal. We can use this as a test statistic to find out whether a population is under HWE or not. This is equivalent to testing the null hypothesis $H_0 : D_C = 0$. Additionally, there are two more tests which can be used for the same purpose, namely, Fisher exact test and likelihood ratio test. The HWE can be extended to the case of multiple-locus genotypes/haplotypes in terms of random mating.

HWE is an important assumption in testing the association between the frequencies of the haplotypes and the phenotypes under the null hypothesis of no association between them. If the association is found, the proportion of genotypes (or haplotypes) in cases will then differ significantly from the ones in controls. Furihata, Ito and Kamatani (2006) found a method, which is so called likelihood-based algorithm PENHAPLO, to test the association when the HWE assumption was invalid in cases.

2.2 Mode of inheritance

Mode of inheritance refers to the way that genetic variants affect the probability of being diseased. It can be determined by a function that is so called penetrance by which the conditional probability of being affected, given a specific genotype. There are several modes which can be defined according to the relationships between haplotype risks and genotype risks. For simplicity, we consider a single-locus with two alleles (or haplotyopes). Let allele D be the risk allele and N be the non-risk allele. The possible genotypes will be DD, ND and NN. We can then define the penetrance functions as follows:

$$f_0 = P(\text{ affected } |NN), f_1 = P(\text{ affected } |ND) \text{ and } f_2 = P(\text{ affected } |DD)$$

Let λ denote the *genotypic relative risk (GRR)*, so that

$$\lambda = \frac{P(\text{ affected } |DD)}{P(\text{ affected } |NN)}$$

Hence, modes of inheritance can be defined as follows (Hartl and Clark, 1997):

1. Dominant model: The risk of getting disease will increase in amount of λ, given the individual's genotype is type ND or DD.

$$f_2 = f_1 = \lambda f_0.$$

2. Recessive model: The risk of getting disease will increase in amount of λ, given the individual's genotype is only type DD.

$$f_2 = \lambda f_0, f_1 = f_0.$$

3. Multiplicative model: The risk of getting disease will increase in amount of λ, given the individual's genotype is type DD and in amount of $\sqrt{\lambda}$, given the individual's genotype is type ND.

$$f_2 = \lambda f_0, f_1 = \sqrt{\lambda} f_0$$

The notation can be easily extended to the case of multiple-locus genotypes/haplotypes (Hartl and Danial, 1997).

2.3 Maximum likelihood method

The maximum likelihood method is the most common method for estimating parameters in a parametric model. Let us consider the locus we mentioned in Section 2.1.1. The likelihood function can be written as

$$L(p_c) = \frac{(2N)!}{n_1! n_2!} {p_1}^{n_1} (1 - p_1)^{(n_2)},$$

where

$$2N = n_1 + n_2 \text{ and } p_1 + p_2 = 1.$$

The log-likelihood function can be written as

$$\log L(p_1) = c + n_1 \log p_1 + n_2 \log(1 - p_1),$$

where c is a constant.

On equating

$$\frac{\partial \log L(p_1)}{\partial p_1}$$

to zero and solving the equation, we have

$$\frac{\partial(p_1)}{\partial p_1} = \frac{n_1}{p_1} - \frac{n_2}{1-p_1} = 0$$

$$\Rightarrow \hat{p}_1 = \frac{n_1}{2N}.$$

Similarly, we have

$$\hat{p}_2 = \frac{n_2}{2N}.$$

2.4 Finite mixture model

Let $n_1, n_2, ..., n_J$ denote a random sample of size N, where $n_j, 1 \leq j \leq J$ is a d-dimensional random vector with probability density function $f(n_j; \theta)$ on $\Re^d$. Suppose that n_j can be classified into k subgroups based on the similarity and the dissimilarity of these observations. This can easily be conducted by using the mixture model given by

$$f(n_j; \theta) = \sum_{i=1}^{k} \pi_i f(n_j; p_i),$$

where $\theta = (\pi_1, ..., \pi_{k-1}; p_1, ..., p_k)^T$, and k is the number of components in the model, $0 \leq \pi_i \leq 1$ is the mixed weights and $\sum_{i=1}^{k} \pi_i = 1$.

The likelihood of θ given data $\mathbf{n}$ can be calculated by

$$L(\theta|\mathbf{n}) = \prod_{j=1}^{J} \sum_{i=1}^{k} \pi_i f(n_j; p_i), \tag{2.4}$$

The log likelihood can be written as

$$l(\theta|\mathbf{n}) = \sum_{j=1}^{J} \log \sum_{i=1}^{k} \pi_i f(n_j; p_i), \tag{2.5}$$

The above likelihood in the equation 2.5 is called the incomplete likelihood. To

formulate the complete one, we need to define group membership indicators. For simplicity, let us assume that k=2 when we are interested in finding the risk and non-risk haplotypes. The haplotypes can be denoted by $\{H_j, 1 \leq j \leq J\}$, and $\theta = (\pi, p_r, p_{\bar{r}}^T)$, where $p_r, p_{\bar{r}}$ refer to risk and non-risk haplotype group, respectively. With that, group membership indicators can be defined by I_{jr} and $I_{j\bar{r}}$,

$$I_{jr} = \begin{cases} 1, & H_j \text{ in the risk group} \\ 0, & \text{otherwise} \end{cases}, \quad I_{j\bar{r}} = \begin{cases} 1, & H_j \text{ in the non-risk group} \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq j \leq J$. Set $\mathbf{I} = \{(I_{jr}, I_{j\bar{r}})^T : 1 \leq j \leq J\}$.

To this end, the log-likelihood in the equation 2.5 can be written as

$$l(\theta|\mathbf{n}, \mathbf{I}) = \sum_{j=1}^{J} \Big\{ I_{jr} \log(\pi f((n_{0j}, n_{1j})^T|p_r)) + I_{j\bar{r}} \log((1-\pi) f((n_{0j}, n_{1j})^T|p_{\bar{r}})) \Big\}, \quad (2.6)$$

where $n_j = n_{0j} + n_{1j}$.

Calculating the maximum likelihood estimation of θ can be difficult analytically. Therefore, some of iterative methods can be employed to calculate the ML estimators numerically. A common way is by employing Newton raphson method or EM algorithm.

2.4.1 Newton-Raphson algorithm

Given the incomplete likelihood in the equation 2.5, Newton-raphson method can be used to find the ML of θ. To illustrate the basic idea behind it, let $S(\mathbf{n}; \theta)$ denote the score and $M(\theta; \mathbf{n})$ denote (Fisher) information matrix of the log-likelihood in 2.5, where $\theta = (\pi, p_r, p_{\bar{r}})$. We then have

$$S(\mathbf{n}; \theta) = (\frac{\partial l}{\partial \pi}, \frac{\partial l}{\partial p_r}, \frac{\partial l}{\partial p_{\bar{r}}})^T,$$

and

$$M(\theta; \mathbf{n}) = \begin{pmatrix} -\frac{\partial^2 l}{\partial \pi^2} & -\frac{\partial^2 l}{\partial \pi \partial p_r} & -\frac{\partial^2 l}{\partial \pi p_{\bar{r}}} \\ -\frac{\partial^2 l}{\partial \pi \partial p_r} & -\frac{\partial^2 l}{\partial p_{\bar{r}}^2} & -\frac{\partial^2 l}{\partial p_r \partial p_{\bar{r}}} \\ -\frac{\partial^2 l}{\partial \pi \partial p_{\bar{r}}} & -\frac{\partial^2 l}{\partial p_r \partial p_{\bar{r}}} & -\frac{\partial^2 l}{\partial p_{\bar{r}}^2} \end{pmatrix}.$$

By Taylor's theorem, we can expand the derivative of the log-likelihood around $\theta^{(t)}$. This gives

$$S(\mathbf{n}; \theta) \approx S(\mathbf{n}; \theta^{(t)}) - (\theta - \theta^{(t)}) M(\theta^{(t)}; \mathbf{n}).$$

Solving the above equation gives

$$\theta^{(t+1)} = \theta^t + M^{-1}(\theta^{(t)}; \mathbf{n}) S(\mathbf{n}; \theta^{(t)}).$$

We repeat these steps until getting convergence to certain values (McLachlan and Krishnan, 2008).

2.4.2 EM algorithm

EM algorithm is an iterative method to find ML of θ by maximising the log-likelihood in (2.6). This algorithm can be performed in two steps: expectation step and maximization step (McLachlan and Peel, 2000a).

Given the current value $\theta^{(t)} = (p_r^{(t)}, p_{\bar{r}}^{(t)}, \pi^{(t)})^T$ and the data $\mathbf{n}$, we first calculate the current log-likelihood $l(\theta^{(t)}|\mathbf{n})$. Then, in the E-step, we calculate the expectation of the complete-data log-likelihood with respect to $\mathbf{I}$,

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E[l(\theta|\mathbf{n}, \mathbf{I})|\mathbf{n}, \theta^{(t)}] \\ &= \sum_{j=1}^{J} (\tau_{jr}^{(t)} \log(\pi) + \tau_{j\bar{r}}^{(t)} \log(1-\pi)) \\ &\quad + \sum_{j=1}^{J} (\tau_{jr}^{(t)} \log(f((n_{0j}, n_{1j})^T|p_r)) + \tau_{j\bar{r}}^{(t)} \log(f((n_{0j}, n_{1j})^T|p_{\bar{r}}))), \end{aligned}$$

where

$$\begin{aligned} \tau_{jr}^{(t)} &= P(I_{jr} = 1|(n_{0j}, n_{1j})^T, \theta^{(t)}) = \frac{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_r^{(t)})}{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_r^{(t)}) + (1-\pi^{(t)}) f((n_{0j}, n_{1j})^T|p_{\bar{r}}^{(t)})}, \\ \tau_{j\bar{r}}^{(t)} &= P(I_{j\bar{r}} = 1|(n_{0j}, n_{1j})^T, \theta^{(t)}) = \frac{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_{\bar{r}}^{(t)})}{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_r^{(t)}) + (1-\pi^{(t)}) f((n_{0j}, n_{1j})^T|p_{\bar{r}}^{(t)})}. \end{aligned}$$

In the M-step, we update $\theta^{(t)}$ by solving the partial derivatives equations

$$\frac{\partial Q}{\partial \pi} = 0, \quad \frac{\partial Q}{\partial p_r} = 0, \quad \frac{\partial Q}{\partial p_{\bar{r}}} = 0.$$

We obtain

$$\pi^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{jr}^{(t)}}{J}, \quad p_r^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{jr}^{(t)} n_{1j}}{\sum_{j=1}^{J} \tau_{jr}^{(t)} (n_{1j} + n_{0j})}, \quad p_{\bar{r}}^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{j\bar{r}}^{(t)} n_{1j}}{\sum_{j=1}^{J} \tau_{j\bar{r}}^{(t)} (n_{1j} + n_{0j})}.$$

An important use of this algorithm is to classify the data $\mathbf{n}$ into risk and non-

risk groups. Based on $\tau_{jr}^{(t+1)}$ and $\tau_{j\bar{r}}^{(t+1)}$, the estimated risk and non-risk haplotype clusters can be defined by

$$S_r^{(t+1)} = \{H_j : \tau_{jr}^{(t+1)} > \tau_{j\bar{r}}^{(t+1)}\}, \quad S_{\bar{r}}^{(t+1)} = \{H_j : \tau_{jr}^{(t+1)} \leq \tau_{j\bar{r}}^{(t+1)}\}.$$

We will show how this algorithm works practically in our approach as we use it to estimate parameters of interest.

2.5 Multi-locus haplotype inference

A sequence of the alleles at different loci on a chromosome is so called haplotype. As each locus is based on two alleles, the possible number of different haplotypes resulted from k SNPs will be 2^k. Haplotypes-based studies are more important than studying SNPs as the latter do not account for the joint behavior of SNPs very well when they are highly correlated to each other. However, we might find evidence of association for one haplotype or many by conducting simultaneous analyses of multiple SNPs that may jointly provide such evidence.

The functional aspects of protein are identified by a sequence of amino acids, corresponding to DNA variations on a haplotype (Clark, 2004). However, in most (if not all) of these studies, haplotypes are generally unknown. Therefore we need to infer them by using available or known genotype data by using some programs such as PHASE (Stephens et al., 2001) or fastPHASE (Scheet and Stephens, 2006) which implement a bayesian framework to phase estimation.

2.5.1 Haplotype reconstructing

Excoffier and Slatkin(1995) proposed a method to reconstruct haplotypes. Assume that a sample of n diploid individuals are observations from a population. Let $\mathbf{G} = (G_1, G_2, ..., G_n)$ denote the genotypes for these individuals, and $\mathbf{H} = (H_1, H_2, ..., H_n)$ the unknown haplotype pairs that produced $\mathbf{G}$, where $H_j = (H_{1j}, H_{2j})$. Let $h = (h_1, h_2, ..., h_k)$ denote all possible haplotypes that can result in $\mathbf{G}$ and $p = (p_1, p_2, ..., p_k)$ be the unknown population frequencies of h. Thus, we can write the maximum likelihood function as follows

$$L(p|\mathbf{G}) = P(\mathbf{G}|p) = \prod_{i=1}^{n} P(G_i|p).$$

Under HWE,

$$P(G_i|p) = \sum_{j=1}^{m_i} P(H_{i1}^j)P(H_{i2}^j),$$

where $m_i = 2^{t_i-1}$, and t_i is the number of heterozygots in the genotype G_i. To clarify this point, suppose we have a set $\mathbf{G}$ consists of 4 individuals genotypes such as

$$\mathbf{G} = \{(2,1,0,2)^T, (1,1,0,0)^T, (1,0,0,1)^T, (0,2,2,1)^T\},$$

where $0, 1$ refers to homozygots and 2 refers to heterozygots (Zhang et al., 2005). In this genotype, there are two heterozygots in this genotype which means it can be decomposed into 4 possible ways. In addition, there are 8 different haplotypes that can result in G. Therefore, the possible different haplotypes will be $h = \{h_1 = (0,1,0,0), h_2 = (1,1,0,1), h_3 = (0,1,0,1), h_4 = (1,1,0,0), h_5 = (1,0,0,1), h_6 = (0,0,0,1), h_7 = (0,1,1,1), h_8 = (0,0,1,1)\}$, and the 4 possible ways, namely assignments, that we can decompose $\mathbf{G}$ into are as follows

$$H_1 = \{(h_1,h_2),(h_4,h_4),(h_5,h_5),(h_6,h_7)\}$$

$$H_2 = \{(h_1,h_2),(h_4,h_4),(h_5,h_5),(h_3,h_8)\}$$

$$H_3 = \{(h_3,h_4),(h_4,h_4),(h_5,h_5),(h_6,h_7)\}$$

$$H_4 = \{(h_3,h_4),(h_4,h_4),(h_5,h_5),(h_3,h_8)\}$$

We use the EM algorithm to estimate the parameters p_i's. We define an indicator vector $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$, where $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{im_i})$,

$$z_{ij} = \begin{cases} 1, & \text{if haplotype pair } (H_{i1}^j, H_{i2}^j) \text{ consistent with } G_i \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

Hence, the complete likelihood can be written as

$$L(p, \mathbf{Z}, \mathbf{G}) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} [P(H_{i1}^j)P(H_{i2}^j)]^{Z_{ij}},$$

and the log-likelihood can be written as

$$\log L(p, \mathbf{Z}, \mathbf{G}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} Z_{ij} \log[P(H_{i1}^j)P(H_{i2}^j)].$$

As we know that EM algorithm can be done through two steps:

E-Step:

$$Q(\theta, \theta_{(t)}) = E[\log L(p, \mathbf{Z}|\mathbf{G})|p_{(t)}]$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{Z}_{ij}[P(H_{i1}^j)P(H_{i2}^j)].$$

We can calculate $\hat{Z}_{ij}$ by using the conditional expectation,

$$\hat{Z}_{ij} = E[Z_{ij}|p, \mathbf{G}, p^{(t)}] = P(Z_{ij} = 1|p, \mathbf{G}, p^{(t)})$$
$$= \frac{P(G_i|Z_{ij} = 1, p, p^{(t)})P(Z_{ij} = 1|p^{(t)})}{P(G_i|p)}$$
$$= \frac{\frac{1}{m_i}[(P(H_{i1}^j))^{(t)}(P(H_{i2}^j))^{(t)}]}{P(G_i|p)}$$
$$= \frac{\frac{1}{m_i}[(P(H_{i1}^j))^{(t)}(P(H_{i2}^j))^{(t)}]}{\sum_{j=1}^{m_i}(P(H_{i1}^j))^{(t)}(P(H_{i2}^j))^{(t)}}.$$

Similarly, we can calculate $[\hat{P}(H_i^j)]^{(t)}$ as follows

$$[\hat{P}(H_i^j)]^{(t)} = E(H_i^j|Z_{ij} = 1, \mathbf{G}, p, p^{(t)})$$
$$= \frac{P(G_i|H_i^j, Z_{ij} = 1, p, p^{(t)})P(H_i^j|Z_{ij} = 1, p^{(t)})}{P(G_i|p)}$$
$$= \frac{P(G_i|H_i^j, Z_{ij} = 1, p, p^{(t)})(P(H_{i1}^j))^{(t)}(P(H_{i2}^j))^{(t)}}{\sum_{j=1}^{m_i}(P(H_{i1}^j))^{(t)(P(H_{i2}^j))^{(t)}}},$$

where $H_i^j = (H_{i1}^j, H_{i2}^j)$.

M-Step: We use gene count method to calculate the population frequencies as follows

$$\hat{p}_\ell^{(t+1)} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i} x_{ij}[\hat{P}(H_i^j)]^{(t)},$$

where $x_{ij} = 0, 1, 2$ depends on how many times the haplotype h_ℓ present in haplotype pair H_i^j, $\ell = 1, 2, ..., k$.

To this end, we can use the calculated frequencies to choose $\hat{\mathbf{H}}$ to maximize $P(\mathbf{H}|\hat{p}, \mathbf{G})$. That is, by choosing the most probable haplotype assignment, given the genotype data. However, within this procedure, it is not clear how best to reconstruct haplotypes. Stephen, Smith and Donnelly (2001) develop a new technique to reconstruct haplotypes by using Gibbs Sampling, a type of MCMC algorithm. This method represents constructing a Markov chain $H^{(0)}, H^{(1)}, H^{(2)}, ..,$ with sta-

tionary distribution $P(\mathbf{H}|\mathbf{G})$, in light of all possible haplotype reconstructions. The steps of their algorithm are as follows:

We assign an initial haplotypes reconstruction $\mathbf{H}^{(0)}$. We then choose an individual ,i, uniformly and at random from all ambiguous individuals. We next sample $H_i^{(t+1)}$ from $P(H_i|\mathbf{G}, H_{-i}^{(t)})$, where H_{-i} is the set of haplotypes excluding individual i. Having done that, we set $H_j^{(t+1)} = H_j^{(t)}$ for $j = 1, 2.., n, j \neq i$. We repeat these steps enough times until we get a convergence. However, the distribution $P(H_i|\mathbf{G}, H_{-i})$ is not even known for most models, so that Stephens, Smith and Donnelly (2001) found out it would be helpful to use the constructing full-conditional distribution for any haplotype pair $H_i = (h_{i1}, h_{i2})$ that consistent with G_i. That is,

$$P(H_i|\mathbf{G}, H_{-i}) \propto P(H_i|H_{-i}) \propto \psi(h_{i1}|H_{-i})\psi(h_{i2}|H_{-i}, h_{i1}),$$

where $\psi(.|\mathbf{H})$ is the conditional distribution of a future sampled haplotype, given a set $\mathbf{H}$ of previously sampled haplotypes. This distribution is also not known generally in many occasions. However, it is known in special case of parent-independent mutation which means that the type of a mutant offspring and the type of parent are independent. To improve and fasten the procedure, we calculate $\psi(h|\mathbf{H})$ as follows

$$\psi(h|\mathbf{H}) = \sum_{\alpha \in E} \sum_{s=0}^{\infty} \frac{r_\alpha}{r} (\frac{\theta}{r+\theta})^s \frac{r}{r+\theta} (P^s)_{\alpha h},$$

where r is the total number of haplotypes in $\mathbf{H}$, r_α is the number of haplotypes of type α in the set $\mathbf{H}$, E refers to the countable set of types of mutation models, θ is the scaled mutation rate, s is sampled from geometric distribution, and P is the(reversible) mutation matrix.

2.5.2 SNP array segmentation

A large number of SNPs can result in sparsely distributed haplotypes with many rare haplotypes, which makes it difficult to detect their associations with the disease. Therefore, some statistical approaches are required to handle the rare haplotypes for complex traits in a population. The existing methods include the two-stage (or multiple Z-testing) method (Zhu et al., 2010), sibpair and odds ratio weighted sum statistics (SPWSS,ORWSS) (Feng et al., 2011), and weighted haplotype and imputation-based test (Li et al., 2010). Detecting the associations of haplotypes with the disease by using conventional statistics such as chi-square test and odds ratio test can suffer from the problem of multiple testing adjustment due to the high

dimensionality of haplotypes. To overcome the above limitation, in this thesis, I first segmentate the SNP array and then reduce the number of tests by use of clustered haplotypes/genotypes, rather than individual haplotypes/genotypes.

2.6 Genome-wide association studies

Genome-wide association studies (GWAS) have played an important role in identifying genetic polymorphisms contributing to complex human diseases. With the rapid pace of developing SNP genotyping technology, a large number of markers (some times greater than 500000) have been considered by many researchers for a large number of individuals to investigate the effects of genetic variants on diseases. The earlier studies have focused on multiple single-locus analyses with an appropriate adjustment for multiple testing effects (Balding et al., 2007).

The problems of high dimensionality, sparsity problems, and linkage disequilibrium can arise from GWAS. In the following sub-section, I will review some existing methods of detecting rare risk variants.

2.6.1 Case-control studies of SNPs with a disease

Studying contributions of SNPs to a disease can be performed by taking two samples of individuals from a population: one with the disease(cases) and the other without the disease (control). For simplicity, assume that the alleles of the suspicious SNP are $\{C, T\}$. But the following tests can be extended to the case of multiple-loci. We use chi-square to test the null hypothesis of no association between SNP and disease. To do so, we represent a contingency table of the observed and the expected genotypes counts for cases and controls (see table (2.2)).

Tab. 2.2: Contingency table of genotypes counts for Cases and Controls. In this table 1, 0 refer to genotypes counts in cases and controls respectively.

	Case			Control		
Genotype	CC	CT	TT	CC	CT	TT
Observed counts	n^1_{CC}	n^1_{CT}	n^1_{TT}	n^0_{CC}	n^0_{CT}	n^0_{TT}
Expected counts	$n^1[\tilde{p}_C^2]^1$	$2n^1[\tilde{p}_C(1-\tilde{p}_C)]^1$	$n^1[(1-\tilde{p}_C)^2]^1$	$n^0[\tilde{p}_C^2]^0$	$2n^0[\tilde{p}_C(1-\tilde{p}_C)]^0$	$n^0[(1-\tilde{p}_C)^2]^0$
Observed-Expected	$n^1\hat{D}^1_C$	$-2n^1\hat{D}^1_C$	$n^1\hat{D}^1_C$	$n^0\hat{D}^0_C$	$-2n^0\hat{D}^0_C$	$n^0\hat{D}^0_C$

. The chi-square test can be calculated by

$$\chi^2 = \sum_{cases} \sum_{genotypes} \frac{(Observed - Expected)^2}{Expected}.$$

We compare the value of the observed test with the tabled one derived from Chi-square distribution of 2 degrees of freedom. However, if there is any of the classes in the table with count less than 5, we will then get non-significant value for the test statistic despite the fact that the SNP could be associated with disease. For this reason, we need to use a continuity correction of 0.5 in the numerator of chi-square to overcome such problem (Yates, 1934).

$$\chi^2 = \sum_{cases} \sum_{genotypes} \frac{(|Observed - Expected| - 0.5)^2}{Expected}$$

In many cases, some diseases can result from more than one SNP. Therefore, some statistical methods are needed to detect the association between SNPs and disease such as multi loci(or haplotype) methods. We will mention some of these methods in the coming sections.

In addition, we can use Fisher's exact test to detect the association, most commonly, when we have one of the classes has count less than five. We can fit the contingency table of the genotypic counts in the cases and the controls as shown in Table 2.3.

Tab. 2.3: The contingency table of genotypic counts of a locus with two alleles C and T in a case-control sample.

Genotype	CC	CT	TT	Total
Case	n^1_{CC}	n^1_{CT}	n^1_{TT}	n_{case}
Control	n^0_{CC}	n^0_{CT}	n^0_{TT}	$n_{control}$
Total	n_{CC}	n_{CT}	n_{TT}	N

We calculate the p as follows

$$p = \frac{n_{case}!\ n_{control}!\ n_{CC}!\ n_{CT}!\ n_{TT}!}{N!\ n^0_{CC}!\ n^0_{CT}!\ n^0_{TT}!\ n^1_{CC}!\ n^1_{CT}!\ n^1_{TT}!}. \tag{2.8}$$

We then need to form all the other samples that follow the prospective and retrospective distribution of the observed data in Table 2.3. We calculate $\{p_i\}$ of all these samples by using 2.8. The fisher's exact test equal to $\sum_{p_i \leq p} p_i$.

The significance of this test is determined by pre-defined significant level.

2.7 Haplotypes clustering

A challenging issue with haplotype-based analyses is the lack of enough phase information to reconstruct haplotypes as many haplotypes may be consistent with unphased genotype data. Rarity is another drawback that we need to cope with when conducting haplotype-based inference. Several methods in the literature have been proposed based on classifying haplotypes according to some similarities into several subgroups and assume the haplotypes of each subgroup have the same effect on disease prevalence in the sample (Templeton et al., 1987; Molitor et al., 2003; Morris, 2005; 2006). Zhu et al. (2010) have proposed a multiple testing method to co-classify the haplotypes in selected subsample based on the difference between P(haplotype|cases) and P(haplotype|controls) in the first stage of his method.

Fitting a logistic regression model to the clustered haplotypes can be more efficient rather than SNPs in measuring their association with a disease (Huang et al., 2011), as the haplotypes can be more informative than SNPs in terms of underlying biological relationships with the disease. However, the rarity and high dimensionality are also challenges in the logistic regression-based analyses (Igo et al., 2009) as they can result in a high degree of freedom that can undermine the estimation of the parameters. In the following two subsections, we will describe briefly the method of Zhu et al. (2010) and the method of the standard logistic regression.

2.7.1 Detecting disease-associated haplotypes

Multiple testing method

Zhu et al. (2010) proposed a method to detect the association between disease and unrelated cases as well as affected sibpairs. This method can be done through two stages. The former stage represents co-classifying the rare risk haplotypes in unrelated cases as well as affected sibpairs. The latter stage represents using Fisher's exact test to find out whether the co-classifying haplotypes are associated with particular disease or not.

Assume that we examine the association of haplotypes with a disease. We first let $H = \{H_1, H_2, ..., H_n\}$ be a set of risk haplotypes with corresponding haplotypes frequencies $p_1, p_2, ..., p_n$ in affected cases and $p_1^0, p_2^0, ..., p_n^0$ in controls, and let H_{n+1} be the rest of the non-risk haplotypes with the total frequency p_{n+1} in cases and p_{n+1}^0 in controls, respectively. We define the cumulative risk haplotype frequency

$p = \sum_{i=1}^{n} p_i$. Let f_2, f_1, f_0 be the three penetrances and defined as follows

$$f_2 = P_r(\text{affected}|H_iH_j),$$

$$f_1 = P_r(\text{affected}|H_iH_{n+1}),$$

and

$$f_0 = P_r(\text{affected}|H_{n+1}H_{n+1}),$$

where $i, j = 1, 2, ..., n$.

We can then calculate the frequency of a rare risk haplotype H_i in cases as follows:

$$h_i = P_r(H_i|\text{affected}) = P_r(H_iH_i|\text{affected}) + 0.5\sum_{j\neq i} P_r(H_iH_j|\text{affected})$$

$$= \frac{f_2P_r(H_iH_i) + 0.5\sum_{j\neq i, j\leq n} f_2P_r(H_iH_j) + 0.5f_1P_r(H_iH_{n+1})}{P_r(\text{affected})}$$

$$= \frac{f_2P_r(H_iH_i) + 0.5\sum_{j\neq i, j\leq n} f_2)P_r(H_iH_j) + 0.5f_1P_r(H_iH_{n+1})}{\sum_{i=1}^{n}\sum_{j=1}^{n} f_2P_r(H_iH_i) + \sum_{i=1}^{n} f_1P_r(H_iH_{n+1}) + f_0P_r(H_{n+1}H_{n+1})}$$

Given the penetrances, we have

$$h_i = P_r(H_i|\text{affected}) = \frac{f_2p_i^0p_i^0 + f_2p_i^0(p - p_i^0) + f_1p_i^0(1-p)}{f_2p^2 + f_12p(1-p) + f_0(1-p)^2}$$

$$= \frac{f_2p_i^0p + f_1p_i^0(1-p)}{f_2p^2 + f_12p(1-p) + f_0(1-p)^2}.$$

Since we study rare risk haplotypes within a family. It is helpful to consider mode of inheritance. In the multiplicative mode, we assume that

$$f_2 = \eta f_0,\ f_1 = \sqrt{\eta} f_0,$$

which imply

$$P_r(H_i|affected) = \frac{\eta f_0p_i^0p + \sqrt{\eta}f_0p_i^0(1-p)}{\eta f_0p^2 + 2\sqrt{\eta}f_0p(1-p) + f_0(1-p)^2}$$

$$= \frac{\eta p_i^0p + \sqrt{\eta}p_i^0(1-p)}{\eta p^2 + 2\sqrt{\eta}p(1-p) + (1-p)^2}$$

$$= \frac{[\sqrt{\eta}p + (1-p)]\sqrt{\eta}p_i^0}{[\sqrt{\eta}p + (1-p)]^2} = \frac{\sqrt{\eta}p_i^0}{\sqrt{\eta}p + (1-p)}.$$

In the Dominant mode, we suppose that

$$f_2 = f_1 = \eta f_0,$$

which imply

$$P_r(H_i|affected) = \frac{\eta f_0 p_i^0 p + \eta f_0 p_i^0 (1-p)}{\eta f_0 p^2 + 2\eta f_0 p(1-p) + f_0(1-p)^2}$$

$$= \frac{\eta p_i^0}{\eta p(2-p) + (1-p)^2}.$$

In the recessive mode, we suppose that

$$f_2 = \eta f_0, f_1 = f_0,$$

which implies

$$P_r(H_i|affected) = \frac{\eta f_0 p_i^0 p + f_0 p_i^0 (1-p)}{\eta f_0 p^2 + 2 f_0 p(1-p) + f_0(1-p)^2} = \frac{[p(\eta-1)+1]p_i^0}{p^2(\eta-1)+1}.$$

At stage 1 of this method, the risk set S can be defined as

$$S = \{H_i | h_i - h_i^0 > \mu \sqrt{\frac{h_i^0(1-h_i^0)}{2N}}\},$$

where N is the number of cases used for co-classification, μ is a predefined constant and h_i is the frequency of rare risk haplotype H_i in unrelated cases. We can estimate h_i^0 from controls, if it is unknown practically. At stage 2, we use the remaining cases and controls to refine the haplotypes in S by using fisher's exact test.

2.7.2 Standard multiple logistic regression

Many studies in the literature have used the standard multiple logistic regression (SL) to analyse the genotype data. In this subsection, we review a standard way of fitting the logistic regression to the data in order to find the disease-associated genotypes (David et al., 2000). The multiple logistic regression model can be written as follows.

$$\log \frac{p(X_i)}{1-p(X_i)} = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}, \tag{2.9}$$

where $1 \leq i \leq n$, and $X_i = (x_{i1}, x_{i2}, ..., x_{iJ})^T$, namely the design variables, J is the total number of genotypes under study and n is the total number of the individuals. Here, $p(X_i)$ can be calculated by

$$p(X_i) = \frac{e^{\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}}}. \tag{2.10}$$

Generally, in the logistic regression model, if there are J design variables with J values, then $J-1$ design variables will be needed to allow an automatic adjustment for the intercept coefficient β_0. Therefore, we let $x_{ij}, 1 \leq j \leq J-1$ take 1 if the genotype G_j is present in the individual i and 0 if not. Let $\beta = (\beta_0, \beta_1, ..., \beta_J)^T$

The likelihood equations may be expressed as follows

$$L(\beta) = \prod_{i=1}^{n} p(X_i)^{y_i} (1 - p(X_i))^{(1-y_i)}$$

and the log-likelihood can be calculated by

$$l(\beta) = \sum_{i=1}^{n} \{y_i \log p(X_i) + (1 - y_i) \log(1 - p(X_i))\}.$$

On equating the first derivative of $l(\beta)$ to zero, we find

$$\sum_{i=1}^{n} (y_i - p(X_i)) = 0$$

and

$$\sum_{i=1}^{n} x_{ij}(y_i - p(X_i)) = 0$$

for $j = 0, 1, 2, ..., J$. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_J)^T$ denote the solution for these equations.

As each individual will have only one genotypes, the covariances of $\underline{\hat{\beta}}$ will be zeros, whereas the variances can be expressed as follows:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 p(X_i)(1 - p(X_i))$$

The observed information matrix $M(\beta)$ will only have the variances of $\hat{\beta}$, and the estimated standard error of the estimated coefficients will be expressed as follows:

$$\widehat{SE}(\hat{\beta}_j) = \left[\widehat{Var}(\hat{\beta}_j)\right]^{1/2},$$

for $j = 0, 1, ..., J$. To this end, finding the importance of the exposure variables in the model needs to perform a test statistics. The most used ones are likelihood ratio test, Wald statistics. The likelihood ratio test can be written as

$$\omega_j = -2\log(\frac{L_0}{L_j}),$$

where L_0 denote the likelihood of the model without j_{th} variable and L_j denote the likelihood of the model with j_{th}. ω_j will follow χ^2 distribution with two degree of freedom, under the hypothesis that the coefficient $\beta_j = 0$.

Wald's test statistics can be expressed as follows:

$$W_j = \hat{\beta}_j / \hat{SE}(\hat{\beta}_j).$$

Under the hypothesis that the coefficient $\beta_j = 0$, these statistics will be distributed as standard normal.

In our later applications of the standard logistic regression to the simulation, we used the generalized linear model (GLM) package in python. In declaring the risk genotypes, we find P-value that corresponds each coefficient. If any of them is less than a specific significant level, we would declare the corresponding genotype is potentially risk.

The tests that we discussed previously required large samples to insure asymptotic normality or χ^2 distributions under the null hypothesis that stated there is no risk haplotypes or genotypes in the samples. The small sample, on the other hand, can also examine by permutation test based on exact test or hypergeometric distribution. The latter test can be performed by representing two-way contingency table of haplotypes or genotypes counts, providing that the row and column margins are fixed at their observed values under the null hypothesis. The null hypothesis here assumes independence of the row and column variables. This procedure can also be applied to case-control samples by permuting the disease status within the individual and calculating P-value(Manly, 2007).

This test can be extended to more complicated cases. For example, it can be used to test the three disease modes: recessive, dominant and multiplicative. In addition, it can be used to calculate empirical P-value for other tests when it is hard to be calculated analytically. We will propose a permutation test for inferring disease-associated genotypes in Chapter 5 of this thesis.

2.8 Study design

Choosing a study design is determined by the nature of objectives of the study. For example, selecting a convenient design to undertake genetic data-based analysis is one of the difficulties that researchers need to overcome. Such difficulties can arise from the biological complexity underlying such a study, such as the rarity and high dimensionality of genotypes/haplotypes. One of frequently used designs, called cohort design is based on choosing a random sample that represents subgroup from the target population. The sample is then divided into sub-samples based on absence or presence of the genetic factors of interest. Subsequently, the absence and presence of a disease are measured. The other popular design, called case-control design is based on selection of two random samples from the target population, representing the diseased and non-diseased populations respectively. However, in practice the cost of the cohort design and the low proportion of a particular disease in the target population force researches to adopt the case-control design (Nicholas, 2003).

In epidemiological studies many issues need to be handled carefully to prevent the bias in measuring disease-exposure association. They are related to: 1) The estimated measure of association based on a randomly selected sample from a population of interest; 2) Assessing the uncertainty of estimation the model parameters in a random sample; 3) Determining whether an observed association in the sample is replicable in the population. To study these issues, we first choose the population to which we conduct our estimation and inference regarding disease-exposure association. Due to the cost of such analyses, it may be difficult to find a population that can avoid all these issues. Therefore, we might choose subgroup of the population that can represent as many as possible features of the population of interest. Let's call this subgroup as *representative population*, the population that we would like to sample from. The study sample than can be chosen from the representative population such that it comprises actual sampled individuals from the representative population. For the individuals of the latter sample, we collect data regarding disease (or any physical trait), exposures or any other factors of interest (e.g. genetic factors and environmental factors).

In this work, we are studying a binary trait in which an individual is carrying disease or not. There are several ways of sampling individuals from the representative population depending on how we scale the exposure variables or the disease prevalence. We consider only two designs in our work which are the most suitable designs for studying genetic data in terms of the high cost of providing data in the reality. To explain how to sample according to these designs, assume we would like

to study association of a disease (D) with several exposures, say $E_1, E_2, ..., E_k$. The two design for this purpose will be described below.

2.8.1 Prospective studies

The first design is called *exposure-based sampling or (cohort design).* In this design, the sampling is undertaking separately for each distinct cohort which may or may not vary in it's exposure level from the others exposures. In sampling individuals according to this design, we first identify k subgroups of the representative population based on the presence of each exposure E_i. We then take a random sample from each subgroup which represents individuals carrying E_i. Finally, we measure subsequently the absence and the presence of D for individuals in the k samples of k different exposures.

2.8.2 Retrospective studies

The second design is called *disease-based sampling (or case-control design).* In this design, two samples will be selected from the representative population; one represents the diseased individuals or cases (D) and the other represents the non-diseased individuals or controls ($\overline{D}$). We the subsequently measure for the presence of each exposure E_i in the cases and the controls.

The most of the genetic data providers are following the case-control design due to the high dimensionality and the rarity of the genetic factors (e.g. WTCCC data). In both designs, we represent the measurements in a 2×2 contingency table and carrying the analysis of association of disease-exposure.

2.9 Specificity and sensitivity

An important way to compare the performances of various test statistics is by using sensitivity and specificity. Consider the below table as the possible outcomes of examining m haplotypes. So, under the null hypothesis H_0 that a haplotype is not associated with a disease, we assume that V refers to the number of false positives or type I errors, S refers to the number of true positives, T refers to the number of false negatives or type II errors, U refers to the number of true negatives and R is the number of rejected null hypotheses.

Tab. 2.4: Outcomes when clustering m hypotheses

Hypothesis	*Accept* H_0	*Reject* H_0	*Total*
H_0 *True*	U	V	m_0
H_1 *True*	T	S	m_1
	W	R	m

.

Hence, we can measure the adequacy of a test statistics by calculating sensitivity and specificity, that is:

$$\text{sensitivity} = \frac{S}{S+T} \text{ and specificity} = \frac{U}{U+V}.$$

We are going to use sensitivity and specificity in comparing our methods to the multiple testing method and the standard logistic regression method.

2.10 Population substructure

Population substructure refers to the characteristics of a population that make expected allele frequencies vary across individuals in a population. Although population substructure is present, the equation 2.3 still leads to unbiased estimation of the allele frequency, but that would hold only if the sampling from the target population is equally likely. However, sample selection can result in unequal probability of representing all the subjects in the sample. As a consequence, the sample estimate may be biased if the genotype frequencies may differ significantly over subgroups. More than that, the distribution of test statistics, which is computed based on allele frequencies and the variance of the estimators, would be affected by population substructure even though there is no bias.

Population substructure can result from population stratification by which individuals in a population sub-classified into mutually exclusive strata. In this type of population substructure the allele frequency for all individual is same, whereas it differs between strata. Here the strata could be racial, ethnic or geographic subgroups. Another reason for population substructure could be admixture in a population that come from the mixing of two populations which have different genetic ancestries. Take migration as an example, it can lead to large difference in allele frequencies in the population (Hartl and Clark, 1997).

2.11 Handling population substructure

Several approaches, in the literature, have been proposed to deal with population substructure. In this thesis, we mention three types of them.

The first approach is genomic control. Many problematic issues have been addressed in this approach such as the effect of population admixture on the variance of the test statistics (Devlin and Roeder, 1999). They proposed the fact that the variance in the χ^2 statistics, which is computed at the null markers, can be empirically estimated rather than focusing on calculating the theoretical variance which is in many occasions not necessarily to be correct.

The second approach is based on using a model and data on the additional markers to control for population substructure to infer the latent population structure and to involve it into the analysis. More to the point, this approach will be more efficient if the population substructure affects the variance of the test statistics. To ensure that the approach will work correctly, the model should be based on strong population admixture in which date is selected from several distinct ethnicities or a large number of ancestry informative markers (Pritchard et al., 2000).

The third approach is by fitting linear or logistic regression model to data based on covariates that represent null markers or linear combinations of null markers. This can be good control of population substructure if the allele frequencies and disease probabilities differ across subgroups (Chen et al., 2003).

3. HAPLOTYPE MIXTURE MODEL-BASED APPROACH (HM)

3.1 Introduction

The advanced genotyping technology and the availability of a large number of dense single nucleotide polymorphisms (SNPs) across human genome have enabled the design of genome-wide association studies (GWAS) for complex diseases.

These studies have progressed from genotyping the SNPs over thousands of case and control subjects (Hindorff et al., 2009), producing large, high-dimensional genotype datasets. The rapid increase in the number of GWAS provides an unprecedented opportunity to examine the effects of rare SNPs on disease susceptibility by the integrative analysis of these data under the assumption that both common and rare SNPs contribute to the underlying genetic mechanisms of complex diseases(Li et al. 2010; Zhu et al. 2010).

It is generally believed that jointly analyzing rare SNPs within a region of strong linkage disequilibrium can be more informative and effective than individual SNP analysis, as multiple SNPs influence the risk of complex diseases in aggregate (Tzeng et al., 2005; Morris et al., 2006; Li et al., 2011; Stranger et al., 2011). The multilocus haplotype, the ordered allele sequences on a chromosome, provides a nature unit of analysis for capturing linear and non-linear correlations in SNPs (Zhang et al., 2003).

Unfortunately, the multi-locus analysis discussed above can suffer from high-dimensional problems that are associated with many predictors, some of which are highly correlated. A popular strategy, suggested by the block-like structure of the human genome, is to divide each chromosome into a list of genetically meaningful regions to reduce the dimensions of these genotype data. Direct, laboratory-based haplotyping to infer the unknown phase are expensive ways to obtain haplotypes. So, in a typical haplotype-based association analysis, people infer haplotypes together with their population frequencies in cases and controls from observed genotypes by using the software such as fastPHASE (Stephens et al., 2001; Scheet and Stephens, 2006). The empirical evidence suggests that the majority of the polymorphism is

concentrated on a relatively small number of haplotypes while the rest is sparsely spread over a number of categories. These non-common haplotypes can be rare and thus hard to assess their disease-susceptibility.

Haplotype clustering offers a promising avenue for addressing the above issue. Over the past decade, enormous progress has been made in this direction and various methods of clustering have been developed on the basis of haplotype similarity and evolution characteristics (Molitor et al., 2003; Tzeng et al., 2006; Browning, 2007; and references therein). However, none of them except Zhu, (2010) has explored advantage of the haplotype similarity in terms of their contributions to disease risks. Zhu, (2010) implemented a method for co-classifying rare risk haplotypes by performing multiple Z-tests for the significant differences between retrospective haplotype frequencies in cases and controls, on the basis that rare risk haplotypes can be enriched in cases. There is a thorny issue of false positive when performing a large number of multiple testing. Moreover, these existing methods are heuristically motivated and not model-based. As model-based algorithms can often provide a principled alternative to heuristic-based algorithms in gene microarray studies (Yeung et al., 2001), it is desirable to develop model-based counterparts of the existing methods mentioned above. Here, to deal with these issues we propose a haplotype mixture model based on the prospective frequencies, as the exact distribution of the differences between retrospective haplotype frequencies are hardly derived. The rationale behind the proposal is as follows. We arrange the haplotype frequencies derived from a case-control study by a contingency table, where rows stand for the disease status (case or control) and columns for haplotypes. Then, we can directly assess whether two haplotypes belong to the same group by their column similarity in the table. Formally, we fit each column by a binomial distribution with the disease-penetrance as the success probability, inferring the grouping of these columns through use of binomial mixtures.

The main advantage of the proposed model over the other existing methods is that it allows the clustering to be directly linked to the disease-penetrances of haplotypes. Moreover, using the estimated prospective haplotype frequencies derived from a retrospective study to estimate disease odds ratio is known to be asymptotically consistent even though the disease-penetrance estimators may not be (Prentice and Pyke, 1979).

We employ the expectation-maximization (EM) algorithm to calculate the maximum likelihood estimator for the proposed mixture model. The EM algorithm can guarantee monotone convergence to a local maximum. On the other hand, it needs to choose initial values in order to reach a local maximum which is close to

the global maximum. The existing ways to choose initial values include: multiple random initializations, initially grouping the data and estimating the corresponding parameters for each mixture component, and among others. See Karlis and Xekalaki (2003) for a review. In this Chapter, we propose a new procedure to regularize the EM algorithm by posterior sampling and compare it to the methods of multiple random initializations and data partition.

We conduct simulation studies on the proposed clustering method in both prospective and retrospective design settings, showing that the proposed method can outperform Zhu et al.'s approach in most cases. We apply both the proposed method and Zhu et al's method to the Coronary Artery Disease (CAD) and Hypertension (HT) data in the Wellcome Trust Case Control Consortium (WTCCC), identifying potential risk haplotypes for each pre-specified chromosomal region.

The rest of the chapter is organized as follows. The proposed methodology and some theory are introduced in Section 3.2. The simulation studies and real data applications are presented in Sections 3.3 and 3.4. Discussions and conclusion are made in Section 3.5.

3.2 Methods

Consider a case-control sample with N_0 controls and N_1 cases, typed at m SNP markers in a candidate region, yielding unphased genotypes $\mathbf{G}$. The disease status y_i of individual i is set to 1 if affected and 0 if unaffected. Let $H_j, 1 \leq j \leq J$ denote the distinct haplotypes inferred from $\mathbf{G}$. We introduce the following approaches for identifying risk haplotypes. The former is a variation of Zhu et al. (2010) and the latter is new.

3.2.1 Multiple testing method (MT)

Following Zhu et al. (2010), a subsample A containing $N_0^{(a)}$ and $N_1^{(a)}$ individuals are randomly chosen from the controls and cases respectively. These individuals are used in the screening stage and the remaining forms a validation subsample B to be used in the validation stage.

Let $(r_{0j}^{(a)}, r_{1j}^{(a)})$, $1 \leq j \leq J^{(a)}$ be the respective frequencies of $J^{(a)}$ haplotypes in controls and cases derived inferred from A. A respective frequencies-based screening

is then performed, giving an estimated risk haplotype set defined by

$$S^{(a)} = \{H_j : z_j^{(a)} > c_0, 1 \leq j \leq J^a\},$$

where c_0 is a pre-specified constant ($c_0 = 1$ in our later simulations) and

$$z_j^{(a)} = \frac{r_{1j}^{(a)} - r_{0j}^{(a)}}{\sqrt{r_{0j}^{(a)}(1 - r_{0j}^{(a)})/(2N_1^{(a)})}}.$$

In the validation stage, the $S^{(a)}$ is refined by using the subsample B, where Fisher's exact testing is performed for each haplotype in $S^{(a)}$, obtaining a final risk haplotype set denoted by $S^{(b)}$.

3.2.2 Mixture model-based method

We now describe the mixture model-based approach, which includes two stages. In Stage 1, we conduct a haplotype grouping by use of the binomial mixtures, while in Stage 2, we refine the selected risk haplotype set by multiple odds ratio (OR) thresholding.

Haplotype grouping can effectively reduce the size of candidate risk-haplotypes, improving the accuracy of multiple thresholding in Stage 2. Note that genetic association studies often assess the overall association between a disease and multilocus-haplotypes in a population by an association test. However, the clinical relevance of an association depends on the magnitude of risk conferred to the carriers of haplotypes. Therefore, an advantage of using OR thresholding over Zhu et al.'s multiple Z-test screening is that we directly select risk haplotypes based on their disease-susceptibility. *Stage 1 (Grouping).* Let $H_j, 1 \leq j \leq J$ denote the distinct haplotypes inferred from $\mathbf{G}$ with haplotype counts $n_{0j}, 1 \leq j \leq J$ and the total $2N_0$ in controls, and $n_{1j}, 1 \leq j \leq J$ and the total $2N_1$ in cases respectively. Then, on the basis of the whole sample, the respective frequencies of the jth haplotype in controls and cases can be estimated by $r_{0j} = n_{0j}/(2N_0)$ and $r_{1j} = n_{1j}/(2N_1)$ respectively. Similarly, the prospective frequencies of the jth haplotype in controls and cases can also be estimated by $p_{0j} = n_{0j}/(n_{0j} + n_{1j})$ and $p_{1j} = n_{1j}/(n_{0j} + n_{1j})$ respectively.

We hypothesize that haplotypes are either risk or non-risk. Under this assumption, the counts $\mathbf{n} = \{(n_{0j}, n_{1j})^T : 1 \leq j \leq J\}$ follow the two-component binomial

mixture,

$$f((n_{0j}, n_{1j})^T|p_r, p_{\bar{r}}, \pi) = \pi f((n_{0j}, n_{1j})^T|p_r) + (1-\pi) f((n_{0j}, n_{1j})^T|p_{\bar{r}}),$$

where $p_r = P(\text{affected}|H_r)$ and $p_{\bar{r}} = P(\text{affected}|H_{\bar{r}})$ are the disease-penetrances of risk haplotype H_r and non-risk haplotype $H_{\bar{r}}$ respectively, and

$$\begin{aligned} f((n_{0j}, n_{1j})^T|p_r) &= \binom{n_{0j}+n_{1j}}{n_{1j}} p_r^{n_{1j}} (1-p_r)^{n_{0j}}, \\ f((n_{0j}, n_{1j})^T|p_{\bar{r}}) &= \binom{n_{0j}+n_{1j}}{n_{1j}} p_{\bar{r}}^{n_{1j}} (1-p_{\bar{r}})^{n_{0j}}. \end{aligned}$$

The unknown parameter $\theta = (p_r, p_{\bar{r}}, \pi)^T$ can be estimated by maximizing the log-likelihood

$$l(\theta|\mathbf{n}) = \sum_{j=1}^{J} \log\left(\pi f((n_{0j}, n_{1j})^T|p_r) + (1-\pi) f((n_{0j}, n_{1j})^T|p_{\bar{r}})\right).$$

Note that the direct calculation of the above maximum likelihood estimator (MLE) is not possible. Instead, we calculate it indirectly by the EM algorithm (McLachlan and Basford, 1988). For this purpose, we introduce the following group membership indicators I_{jr} and $I_{j\bar{r}}$,

$$I_{jr} = \begin{cases} 1, & H_j \text{ in the risk group} \\ 0, & \text{otherwise} \end{cases}, \quad I_{j\bar{r}} = \begin{cases} 1, & H_j \text{ in the non-risk group} \\ 0, & \text{otherwise} \end{cases}$$

for $1 \le j \le J$. Set $\mathbf{I} = \{(I_{jr}, I_{j\bar{r}})^T : 1 \le j \le J\}$. Then, the so-called complete-data log-likelihood can be written as

$$l(\theta|\mathbf{n}, \mathbf{I}) = \sum_{j=1}^{J} \left\{ I_{jr} \log(\pi f((n_{0j}, n_{1j})^T|p_r)) + I_{j\bar{r}} \log((1-\pi) f((n_{0j}, n_{1j})^T|p_{\bar{r}})) \right\}.$$

Given the current value $\theta^{(t)} = (p_r^{(t)}, p_{\bar{r}}^{(t)}, \pi^{(t)})^T$ and the data $\mathbf{n}$, we first calculate the current log-likelihood $l(\theta^{(t)}|\mathbf{n})$. Then, in the E-step, we calculate the expectation of the complete-data log-likelihood with respect to $\mathbf{I}$,

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E[l(\theta|\mathbf{n}, \mathbf{I})|\mathbf{n}, \theta^{(t)}] \\ &= \sum_{j=1}^{J} (\tau_{jr}^{(t)} \log(\pi) + \tau_{j\bar{r}}^{(t)} \log(1-\pi)) \\ &\quad + \sum_{j=1}^{J} (\tau_{jr}^{(t)} \log(f((n_{0j}, n_{1j})^T|p_r)) + \tau_{j\bar{r}}^{(t)} \log(f((n_{0j}, n_{1j})^T|p_{\bar{r}}))), \end{aligned}$$

where

$$\tau_{jr}^{(t)} = P(I_{jr}=1|(n_{0j},n_{1j})^T,\theta^{(t)}) = \frac{\pi^{(t)} f((n_{0j},n_{1j})^T|p_r^{(t)})}{\pi^{(t)} f((n_{0j},n_{1j})^T|p_r^{(t)}) + (1-\pi^{(t)}) f((n_{0j},n_{1j})^T|p_{\bar{r}}^{(t)})},$$

$$\tau_{j\bar{r}}^{(t)} = P(I_{j\bar{r}}=1|(n_{0j},n_{1j})^T,\theta^{(t)}) = \frac{\pi^{(t)} f((n_{0j},n_{1j})^T|p_{\bar{r}}^{(t)})}{\pi^{(t)} f((n_{0j},n_{1j})^T|p_r^{(t)}) + (1-\pi^{(t)}) f((n_{0j},n_{1j})^T|p_{\bar{r}}^{(t)})}.$$

In the M-step, we update $\theta^{(t)}$ by solving the partial derivatives equations

$$\frac{\partial Q}{\partial \pi}=0, \quad \frac{\partial Q}{\partial p_r}=0, \quad \frac{\partial Q}{\partial p_{\bar{r}}}=0.$$

We obtain

$$\pi^{(t+1)} = \frac{\sum_{j=1}^J \tau_{jr}^{(t)}}{J}, \quad p_r^{(t+1)} = \frac{\sum_{j=1}^J \tau_{jr}^{(t)} n_{1j}}{\sum_{j=1}^J \tau_{jr}^{(t)}(n_{1j}+n_{0j})}, \quad p_{\bar{r}}^{(t+1)} = \frac{\sum_{j=1}^J \tau_{j\bar{r}}^{(t)} n_{1j}}{\sum_{j=1}^J \tau_{j\bar{r}}^{(t)}(n_{1j}+n_{0j})}.$$

Let $\text{err}^{(t+1)}$ denote the absolute distance of $l(\theta^{(t+1)}|\mathbf{n})$ to the previous $l(\theta^{(t)}|\mathbf{n})$. We calculate the updated log-likelihood $l(\theta^{(t+1)}|\mathbf{n})$ and $\text{err}^{(t+1)}$.

Start with the initial value $\theta^{(0)}$, we alternatively run the E-step and the M-step for $t=0,1,...$, till $\text{err}^{(t+1)}$ is less than a pre-specified value d_0 (we set $d_0=0.0001$ in our codes). Suppose that the algorithm stops at $(t+1)$th iteration. Based on $\tau_{jr}^{(t+1)}$ and $\tau_{j\bar{r}}^{(t+1)}$, the estimated risk and non-risk haplotype clusters can be defined by

$$S_r^{(t+1)} = \{H_j : \tau_{jr}^{(t+1)} > \tau_{j\bar{r}}^{(t+1)}\}, \quad S_{\bar{r}}^{(t+1)} = \{H_j : \tau_{jr}^{(t+1)} \le \tau_{j\bar{r}}^{(t+1)}\}.$$

Stage 2 (OR thresholding): We are going to refine the above selected risk haplotype set on the basis of their odds ratios. Let n_{0H} and n_{1H} be control- and case-counts of the haplotype H. Let $n_{0\bar{r}} = \sum_{H_* \in S_{\bar{r}}^{(t+1)}} n_{0H_*}$ and $n_{1\bar{r}} = \sum_{H_* \in S_{\bar{r}}^{(t+1)}} n_{1H_*}$. The corrected OR statistic is defined by

$$\text{OR}_H = \frac{(n_{1H}+0.5)(n_{0\bar{r}}+0.5)}{(n_{0H}+0.5)(n_{1\bar{r}}+0.5)}.$$

Then, the risk haplotype set $S_r^{(t+1)}$ is updated by

$$\hat{S}_r = \left\{ H \in S_r^{(t+1)} : \text{OR}_H \ge \exp(c_1 \phi(n_{0H}, n_{1H}, n_{0\bar{r}}, n_{1\bar{r}})) \right\},$$

where

$$\phi(n_{0H}, n_{1H}, n_{0\bar{r}}, n_{1\bar{r}}) = \sqrt{1/(n_{0H}+0.5) + 1/(n_{1H}+0.5) + 1/(n_{0\bar{r}}+0.5) + 1/(n_{1\bar{r}}+0.5)}$$

and c_1 is a pre-specified constant (in our later simulations, we set $c_1 = 2.6$, while in the real data analysis, invoking the Bonferroni adjustment, we set $c_1 = 5.3$).

The non-risk haplotype set is updated by

$$\hat{S}_{\bar{r}} = S_{\bar{r}} \cup (S_r - \hat{S}_r).$$

Given the clusters $\hat{S}_r$ and $\hat{S}_{\bar{r}}$, the estimators of π, p_r, and $p_{\bar{r}}$ are updated by

$$\hat{\pi} = \frac{|\hat{S}_r|}{|\hat{S}_r| + |\hat{S}_{\bar{r}}|}, \quad \hat{p}_r = \frac{\sum_{H \in \hat{S}_r} n_{1H}}{\sum_{H \in \hat{S}_r}(n_{1H} + n_{0H})}, \quad \hat{p}_{\bar{r}} = \frac{\sum_{H \in \hat{S}_{\bar{r}}} n_{1H}}{\sum_{H \in \hat{S}_{\bar{r}}}(n_{1H} + n_{0H})}.$$

The population frequencies of $\hat{S}_r$ and $\hat{S}_{\bar{r}}$ (i.e., $P(H \in \hat{S}_r)$ and $P(H \in \hat{S}_{\bar{r}})$ are estimated by their retrospective frequencies in controls,

$$\hat{P}(\hat{S}_r) = \frac{\sum_{H \in \hat{S}_r} n_{0H}}{\sum_{H \in \hat{S}_r \cup \hat{S}_{\bar{r}}} n_{0H}}, \quad \hat{P}(\hat{S}_{\bar{r}}) = 1 - \hat{P}(\hat{S}_r).$$

3.2.3 Example

To apply our method, we choose one example from our simulation whose disease status is simulated according to multiplicative model. The number of the risk haplotypes is 10. The actual probabilities for

$$\theta = (\pi, p_{\bar{r}}, p_r) = (0.8846, 0.1154, 0.28218),$$

and the incomplete log likelihood value is -249.028, given the above values for the parameters of the model. We applied the mixture model to classify these haplotypes into risk and non-risk group according to their association with the disease status. In this method, we use the EM algorithm to maximize the likelihood. The results of the final iterations corresponding to different sets of the initial values obtained by the EM algorithm are shown in Table 3.1. In the same table, we also showed the result of the specificity and the sensitivity that correspond each set of the initial values.

As it can be seen from Table 3.1, the likelihood value is vary depending on the initial values that the algorithm starts with. However, the one that is close to the actual likelihood value can be seen in the row 7 of this table. More than that, the corresponding parameters estimators in the final iteration are also close to the real one plus good result for the specificity and sensitivity. In this table, we can clearly see the problematic issue of the high dimensionality.

Tab. 3.1: The table shows the first and the last iteration of the EM on Example 3.2.3 starting from different random initial values.

	Initial values				Final iteration						
Rep.	π	$(1-\pi)$	$p_{\bar{r}}$	p_r	π	$(1-\pi)$	$p_{\bar{r}}$	p_r	Inc $l(\theta)$	Spec.	Sens.
1	0.24902	0.75098	0.34972	0.94064	0.91168	0.08832	0.15292	0.85198	-426.7096	0.0	0.92754
2	0.59258	0.40742	0.18077	0.71109	0.89168	0.10832	0.14929	0.50297	-380.56764	0.11111	0.92754
3	0.02434	0.97566	0.0168	0.65403	0.57707	0.42293	0.10392	0.28041	-244.03129	0.88889	0.63768
4	0.56222	0.43778	0.07759	0.74431	0.8751	0.1249	0.14888	0.49672	-376.40283	0.22222	0.91304
5	0.06436	0.93564	0.36483	0.98282	0.90368	0.09632	0.1529	0.83649	-427.18425	0.0	0.91304
6	0.11409	0.88591	0.39284	0.83895	0.83598	0.16402	0.15292	0.57163	-406.62222	0.0	0.7971
7	0.75457	0.24543	0.11636	0.31322	0.7454	0.2546	0.10523	0.28112	-242.08635	0.77778	0.85507
8	0.71065	0.28935	0.43501	0.11713	0.3119	0.6881	0.30853	0.12025	-256.04382	0.66667	0.66667
9	0.15036	0.84964	0.37186	0.65663	0.78362	0.21638	0.15278	0.41214	-344.93251	0.11111	0.72464
10	0.50194	0.49806	0.53793	0.20123	0.1557	0.8443	0.47222	0.14898	-363.21942	0.22222	0.91304
11	0.74117	0.25883	0.71336	0.38348	0.15892	0.84108	0.52891	0.15292	-403.88364	0.0	0.7971
12	0.83943	0.16057	0.47591	0.01123	0.51767	0.48233	0.24734	0.09528	-253.70363	0.88889	0.47826
13	0.15849	0.84151	0.95281	0.26195	0.04946	0.95054	0.9876	0.15307	-424.34528	0.0	0.94203
14	0.6285	0.3715	0.14724	0.17534	0.64626	0.35374	0.10612	0.25421	-246.51166	0.77778	0.91304
15	0.92719	0.07281	0.60895	0.39665	0.29494	0.70506	0.33568	0.152	-321.05469	0.22222	0.69565
16	0.85725	0.14275	0.93536	0.43433	0.1121	0.8879	0.75424	0.15292	-422.11437	0.0	0.91304
17	0.45401	0.54599	0.39472	0.48626	0.79425	0.20575	0.15031	0.36531	-325.23985	0.22222	0.7971
18	0.52902	0.47098	0.05072	0.58287	0.76284	0.23716	0.12055	0.31215	-257.3879	0.55556	0.84058
19	0.04303	0.95697	0.30131	0.63928	0.71148	0.28852	0.14936	0.41862	-345.23502	0.33333	0.71014
20	0.25866	0.74134	0.17324	0.20927	0.41699	0.58301	0.10381	0.2443	-256.31434	1.0	0.31884

3.2.4 Improving the mixture approach

There are various ways to improve the proposed mixture model such as choosing its initial values for the EM algorithm by multiple random starts, a grid search, and the data partition.See Karlis and Xekalaki (2003) for a review. Here, we consider the three methods to improve Stage 1 above. The first two aim to improve the EM algorithm by the choice of initial values, while the last one uses the Bayesian posteriors to relax the model assumption. Note that in the proposed mixture model, the haplotypes in each group are assumed to have the same disease-penetrance, which may not be true in practice. So, to allow for the disease-penetrance variations within each group, the disease-penetrances of haplotypes within each group are assumed to follow a prior distribution. *Method 1 (random initialization)*: We randomly choose i_0 initial values (say $i_0 = 100$) of θ and run the EM algorithm in Stage 1 with each chosen initial value. We take the best one among these runs in terms of maximizing the log-likelihood. *Method 2 (data initial partition)*: In order to initialize the parameters p_r, $p_{\bar{r}}$ and π, we partition the prospective haplotype frequencies $p_{1j}, 1 \leq j \leq J$. Without loss of generality, we assume that $0 < p_{1j} < 1$ for $1 \leq j \leq J$. Otherwise, we only select $0 < p_{1j} < 1$ in our calculation.

We partition these frequencies into two sets with low and high values respectively,

by defining

$$T_r = \{k : p_{1k} > (\max_j p_{1j} - \min_j p_{1j})/2\}, \quad T_{\bar{r}} = \{k : p_{1k} \leq (\max_j p_{1j} - \min_j p_{1j})/2\}$$

Set the initial values,

$$p_r^{(0)} = \frac{\sum_{k \in T_r} p_{1k}}{|T_r|}, \quad p_{\bar{r}}^{(0)} = \frac{\sum_{k \in T_{\bar{r}}} p_{1k}}{|T_{\bar{r}}|}, \quad \pi_r^{(0)} = \frac{|T_r|}{|T_r| + T_{\bar{r}}}$$

where $|T_r|$ and $|T_{\bar{r}}|$ stand for the cardinalities of T_r and $T_{\bar{r}}$ respectively. We then run the EM algorithm in Stage 1. *Method 3 (Bayesian regularization)*: We first randomly generate i_0 (say $i_0 = 100$) initial values at which we calculate the log-likelihoods, and take the one which attains the maximum as the initial value $\theta^{(0)} = (p_r^{(0)}, p_{\bar{r}}^{(0)}, \pi^{(0)})^T$ for the posterior sampling. Motivated by the Gibbs sampling, we employ the posterior of θ to improve each iteration of the EM. Here, we draw $q_r^{(t)}$ and $q_{\bar{r}}^{(t)}$ from the posteriors of p_r and $p_{\bar{r}}$ at the iteration t. Start with the initial $\theta^{(0)}$ and set $q_r^{(0)} = p_r^{(0)}$ and $q_{\bar{r}}^{(0)} = p_{\bar{r}}^{(0)}$. At the iteration $t+1$, given $\theta^{(t)} = (p_r^{(t)}, p_{\bar{r}}^{(t)}, \pi^{(t)})^T$, we have the expected values of I_{jr} and $I_{j\bar{r}}$, say $\tau_{jr}^{(t)}$ and $\tau_{j\bar{r}}^{(t)}$. Haplotype grouping can be defined by

$$S_r^{(t)} = \{H_j : \tau_{jr}^{(t)} > \tau_{j\bar{r}}^{(t)}\}, \quad S_{\bar{r}}^{(t)} = \{H_j : \tau_{jr}^{(t)} \leq \tau_{j\bar{r}}^{(t)}\}.$$

Collapse haplotypes in S_r and calculate the counts of the collapsed S_r in controls and cases, s_{0r} and s_{1r}. Similarly, collapse $S_{\bar{r}}$ and calculate the counts of the collapsed $S_{\bar{r}}$ in controls and cases, $s_{0\bar{r}}$ and $s_{1\bar{r}}$. Based on these counts, the likelihood functions of p_r and $p_{\bar{r}}$ can be written as

$$l(p_r|(s_{0r}, s_{1r})^T) \propto p_r^{s_{1r}}(1-p_r)^{s_{0r}}, \quad l(p_{\bar{r}}|(s_{0\bar{r}}, s_{1\bar{r}})^T) \propto p_{\bar{r}}^{s_{1\bar{r}}}(1-p_{\bar{r}})^{s_{0\bar{r}}}.$$

Let $p_r^{\delta_1}(1-p_r)^{\delta_0}$ and $p_{\bar{r}}^{\delta_0}(1-p_{\bar{r}})^{\delta_1}$ denote the conjugate priors for p_r and $p_{\bar{r}}$ respectively, with the pre-specified pseudo-counts δ_0 and δ_1. We expect that a risk haplotype appears more frequently in cases than does any non risk haplotype. So, the pseudo-counts should satisfy the constrain $\delta_1 > \delta_0$. They should also be small compared to the number of cases. In this work, we set $\delta_1 = 8$ and $\delta_0 = 2$. In our simulations, we found the results are not very sensitive to the choice of these constants. After setting the above priors, we then derive the posteriors,

$$p(p_r|(s_{0r}, s_{1r})^T) \propto \text{Beta}(\delta_1 + s_{1r}, \delta_0 + s_{0r}), \quad p(p_{\bar{r}}|(s_{0\bar{r}}, s_{1\bar{r}})^T) \propto \text{Beta}(\delta_0 + s_{0\bar{r}}, \delta_1 + s_{1\bar{r}})$$

We draw $q_r^{(t+1)}$ from $p(p_r|(s_{0r}, s_{1r})^T)$ and $q_{\bar{r}}^{(t+1)}$ from $p(p_{\bar{r}}|(s_{0\bar{r}}, s_{1\bar{r}})^T)$. We update

the estimates of p_r, $p_{\bar{r}}$ and π by posterior averaging,

$$p_r^{(t+1)} = \frac{1}{t+2}\sum_{k=0}^{t+1} q_r^{(k)}, \quad p_{\bar{r}}^{(t+1)} = \frac{1}{t+2}\sum_{k=0}^{t+1} q_{\bar{r}}^{(k)}, \quad \pi^{(t+1)} = \frac{|S_r^{(t)}|}{|S_r^{(t)}| + |S_{\bar{r}}^{(t)}|}.$$

Finally, we repeat the above procedure until the absolute difference between the estimates of θ in two consecutive iterations is less than a pre-specified value, say 0.0001. In the late section, we show the superiority of Method 3 over the other two methods by simulations. In light of this, we replace the M-step in the EM by Method 3 to form a hybrid EM algorithm. In summary, we opt for the following *two-stage hybrid mixture approach* for association analysis in the remaining framework: *Stage 1 (Grouping):* Use the hybrid EM algorithm to estimate the two-component binomial mixture model. *Stage 2 (OR thresholding):* Use the OR statistic to screen haplotypes further as in the previous section.

3.2.5 Model justification

To make the proposed model identifiable, we need to assume that the disease-penetrance ratio $p_r/p_{\bar{r}} > 1$, that is, risk haplotypes are more enriched in cases than non-risk haplotypes. In this section, under the commonly used inheritance models, we prove the above hypothesis holds when the so-called relative risk measure is larger than one. For this purpose, let S_r and $S_{\bar{r}}$ denote the risk and non-risk haplotype sets in the population. Suppose that the disease-penetrance of a genotype depends only on the number of risk haplotypes contained in that genotype. Then, we have three types of penetrance:

$$f_0 = P(\text{affected}|H_{\bar{r}}H_{\bar{r}}), \quad f_1 = P(\text{affected}|H_rH_{\bar{r}}), \quad f_2 = P(\text{affected}|H_rH_r),$$

where $H_r \in S_r$ and $H_{\bar{r}} \in S_{\bar{r}}$. Denote the relative risk measures $\lambda_1 = f_1/f_0$ and $\lambda = f_2/f_0$. In the following, we show that the haplotype disease-penetrances, $P(\text{affected}|H_r)$ and $P(\text{affected}|H_{\bar{r}})$ are linear functions of the relative risk measures of genotypes and the population haplotype frequencies. Note that if an individual has the haplotype H, his/her genotype can be: HH (homozygous); HH_r, $H_r \in S_r$ and $H \neq H_r$ (heterozygous); $HH_{\bar{r}}$, $H_{\bar{r}} \in S_{\bar{r}}$ and $H \neq H_{\bar{r}}$ (heterozygous). Therefore,

under the Hardy-Weinberg equilibrium, we have

$$\begin{aligned}
P(\text{affected}, H_r) &= P(\text{affected}, H_rH_r) + 0.5 \sum_{H \in S_r, H \neq H_r} P(\text{affected}, H_rH) \\
&\quad +0.5 \sum_{H \in S_{\bar{r}}, H \neq H_r} P(\text{affected}, H_rH) \\
&= f_2 P(H_r)P(H_r) + f_2 P(H_r) \sum_{H \in S_r, H \neq H_r} P(H) + f_1 P(H_r) \sum_{H \in S_{\bar{r}}, H \neq H_r} P(H) \\
&= P(H_r)(f_2 P(H_r) + f_2 \sum_{H \in S_r, H \neq H_r} P(H) + f_1 \sum_{H \in S_{\bar{r}}, H \neq H_r} P(H)) \\
&= P(H_r)(f_2 P(H \in S_r) + f_1 P(H \in S_{\bar{r}})),
\end{aligned}$$

where $P(H_r)$, $P(H \in S_r)$ and $P(H \in S_{\bar{r}})$ are the population frequencies of H_r, S_r and $S_{\bar{r}}$. Consequently,

$$\begin{aligned}
P(\text{affected}|H_r) &= \frac{P(\text{affected}, H_r)}{P(H_r)} \\
&= f_2 P(H \in S_r) + f_1 P(H \in S_{\bar{r}}) \\
&= f_0 \left\{ \lambda P(H \in S_r) + \lambda_1 P(H \in S_{\bar{r}}) \right\}.
\end{aligned}$$

In much the same spirit, we can show that

$$P(\text{affected}|H_{\bar{r}}) = f_0 \left\{ \lambda_1 P(H \in S_r) + P(H \in S_{\bar{r}}) \right\}.$$

The disease-penetrance ratio between risk and non-risk haplotypes,

$$\frac{P(\text{affected}|H_r)}{P(\text{affected}|H_{\bar{r}})} = \frac{\lambda_1 \{ \lambda P(H \in S_r)/\lambda_1 + P(H \in S_{\bar{r}}) \}}{\lambda_1 P(H \in S_r) + P(H \in S_{\bar{r}})}.$$

We can further show that under the commonly used models of inheritance (multiplicative, dominant, and recessive), the haplotype relatively risk (i.e., the disease-penetrance ratio between the risk and non-risk haplotypes) is larger than one if and only if the corresponding genotype relative risk is larger than one. The details are as follows. In a multiplicative model, where $\lambda = \lambda_1^2$, we have

$$\frac{P(\text{affected}|H_r)}{P(\text{affected}|H_{\bar{r}})} = \sqrt{\lambda},$$

which is larger than 1 if and only if $\lambda > 1$. In a dominant model, where $\lambda = \lambda_1$, we have

$$\frac{P(\text{affected}|H_r)}{P(\text{affected}|H_{\bar{r}})} = \frac{\lambda}{\lambda P(H \in S_r) + P(H \in S_{\bar{r}})},$$

which is larger than 1 if and only if $\lambda > 1$. In a recessive model, where $\lambda_1 = 1$, we have

$$\frac{P(\text{affected}|H_r)}{P(\text{affected}|H_{\bar{r}})} = \lambda P(H \in S_r) + P(H \in S_{\bar{r}}),$$

which is larger than 1 if and only if $\lambda > 1$. The above results imply that when the genotype relative risk $\lambda > 1$, the individuals carrying the risk haplotype H_r will have more chance of getting the disease than do non-risk haplotype carriers; when $\lambda < 1$, the individuals carrying H_r have the less chance of getting the disease than do non-risk haplotype carriers and thus H_r plays a disease-protective role.

3.2.6 Testing for haplotype inheritance modes

In the previous subsection, we develop a theory on the identification of the proposed model under certain inheritance assumption on hyplotypes. However, the biological justification for the choice of an inheritance model is seldom available and lack of a statistical justification for the specific genetic model is customary practice. To address the issue, we introduce a statistical test as follows. We begin with deriving non-parametric estimators of the genotype disease-penetrances. Suppose $\hat{S}_r$ and $\hat{S}_{\bar{r}}$ are the estimated risk and non-risk haplotype sets obtained from our hybrid mixture approach.

Let $\mathbf{G}_0$ be the set containing the observed genotypes which consist of two haplotypes in $\hat{S}_{\bar{r}}$, $\mathbf{G}_1$ the set containing the observed genotypes which consist of one haplotype in $\hat{S}_r$ and one in $\hat{S}_{\bar{r}}$, and $\mathbf{G}_2$ containing the observed genotypes which consist of two haplotypes in $\hat{S}_r$. For $k = 0, 1, 2$, we then calculate the total haplotype frequencies of $\mathbf{G}_k$ in controls and cases, denoted by $(n_{02}, n_{12}), (n_{01}, n_{11}), (n_{00}, n_{10})$ respectively. Then the disease-penetrances of genotypes can be estimated non-parametrically by

$$\hat{f}_0 = \frac{n_{10}}{n_{10} + n_{00}}, \quad \hat{f}_1 = \frac{n_{11}}{n_{01} + n_{11}}, \quad \hat{f}_2 = \frac{n_{12}}{n_{02} + n_{12}}.$$

Let A denote the set of the above three inheritance modes: the multiplicative, the dominant, and the recessive. We assume that genotypes are linked their underlying haplotype pairs via the Hardy-Weinberg equilibrium. To test for an inheritance mode, for $a \in A$ and $k = 0, 1, 2$, we first derive a parametric estimator of f_k, say $\hat{f}_k^{(a)}$ by using the estimators $\hat{p}_r$, $\hat{p}_{\bar{r}}$, $\hat{P}(\hat{S}_r)$ obtained in the previous subsection. We then calculate the statistic

$$D_a = |\hat{f}_0 - \hat{f}_0^{(a)}| + |\hat{f}_1 - \hat{f}_1^{(a)}| + |\hat{f}_2 - \hat{f}_2^{(a)}|.$$

We calculate the minimum $D_A = \min_{a \in A} D_a$ and record $\hat{a}$ at which D_a attains the minimum. We expect that D_A takes small values when one of modes in A is true. We can quantitatively justify the significance by use of the following parametric bootstrap test: We re-sampling genotypes M times on the basis of the estimated mode $\hat{a}$ with the estimated penetrances $\hat{f}_k^{(\hat{a})}$, $k = 0, 1, 2$. We set $M = 100$ in our simulation. Each bootstrap dataset contains the original genotypes (and their haplotype pairs) but with new sets of case and control counts. We apply the two-stage hybrid mixture approach to these datasets respectively, obtaining M bootstrap values $D_{Am}, m = 1, ..., M$. The empirical p-value $\sum_{m=1}^{M} I(D_A > D_{Am})/M$ can be used to judge the significance of the test. To conclude this section, we now state the formulas for estimating the disease-penetrances under the three inheritance models. The proofs are straightforward and thus omitted. We use the notations $\lambda = f_2/f_0$ and $\lambda_1 = f_1/f_0$ introduced before.

- *Multiplicative model, where* $\lambda = \lambda_1^2$. We have

$$\hat{\lambda} = \left(\frac{\hat{p}_r}{\hat{p}_{\bar{r}}}\right)^2, \quad \hat{f}_0 = \frac{\hat{p}_{\bar{r}}}{(\sqrt{\hat{\lambda}} - 1)\hat{P}(\hat{S}_r) + 1}, \quad \hat{f}_2 = \hat{\lambda}\hat{f}_0, \quad \hat{f}_1 = \sqrt{\hat{\lambda}}\hat{f}_0.$$

- *Dominant model, where* $\lambda = \lambda_1$. We have

$$\hat{\lambda} = \frac{\hat{P}(\hat{S}_{\bar{r}})}{\hat{p}_{\bar{r}}/\hat{p}_r - \hat{P}(\hat{S}_r)}, \quad \hat{f}_0 = \frac{\hat{p}_{\bar{r}}}{(\hat{\lambda} - 1)\hat{P}(\hat{S}_r) + 1}, \quad \hat{f}_1 = \hat{f}_2 = \hat{\lambda}\hat{f}_0.$$

- *Recessive model, where* $\lambda_1 = 1$. We have

$$\hat{\lambda} = (\hat{p}_r/\hat{p}_{\bar{r}} - \hat{P}(\hat{S}_{\bar{r}}))/P(\hat{S}_r), \quad \hat{f}_1 = \hat{f}_0 = \hat{p}_{\bar{r}}, \quad \hat{f}_2 = \hat{\lambda}\hat{f}_0.$$

3.3 Simulation studies

In this section, via simulations we will examine the performance of the proposed methods in terms of the estimated L_1 bias and the average of sensitivity and specificity under various scenarios. Let $\hat{\theta}$ be the estimator of θ, and $\hat{S}_r$ and $\hat{S}_{\bar{r}}$ the estimators of the true risk and non-risk haplotype sets S_r and $S_{\bar{r}}$ respectively. Then, by the L_1 bias, we mean the L_1 distance between $\hat{\theta}$ and θ. The sensitivity and specificity of $\hat{S}_r$ and $\hat{S}_{\bar{r}}$ are defined as sen $= \frac{|\hat{S}_r \cap S_r|}{|S_r|}$ and spe $= \frac{|\hat{S}_{\bar{r}} \cap S_{\bar{r}}|}{|S_{\bar{r}}|}$. We take the average AVSS $=$ (sen + spe)/2 to assess the performance of the haplotype classification above.

3.3.1 Performance of the proposed Bayesian regularization

To compare Method 3 (the Bayesian regularization) to Methods 1 and 2 (the initialization), we simulated 30 genotype datasets on 10 SNPs, each dataset containing N_0 controls and N_1 cases was obtained by the following two steps:

In the step 1, we used the software MS (Hudson, 2002) to simulate $2(N_0 + N_1)$ haplotypes with a mutation rate of 2. We randomly chose m_r of these haplotypes and labeled them as risk haplotypes. To save the space, we considered only $N_0 + N_1 = 5000$ and $m_r = 10$. The results for other values of $N_0 + N_1$ and m_r are similar.

In the step 3, the disease states of the above genotypes were simulated from the multiplicative inheritance model with $f_0 = 0.1$ and $\lambda = 3$. Note that the number of genotypes depends on the mutation rate and was varying across 30 datasets.

We applied the three methods mentioned in Subsection 2.3 to each of these datasets and recorded their corresponding L_1 distances between the estimated and the true values of θ, the log-likelihoods, the AVSS values, and the CPU-time costs in seconds. Finally, multiple box-whisker plots of these quantities across the three methods are presented in Figure 3.1. The results demonstrate that although Method 3 did not give a global maximum of the log-likelihood, it outperformed the other two methods in terms of estimated biases.

This is not surprising because the log-likelihood in Method 3 have been regularized by the prior. Methods 1 and 3 performed similar and were better than Method 2 in terms of the AVSS. Methods 2 and 3 was much less costly in terms of CPU-time compared to Methods 1.

Overall, Method 3 was ranked to the first. So, we decided to replace the M-step in Stage 1 by Method 3 to form a hybrid mixture approach, which was used in the simulations and real data analysis below.

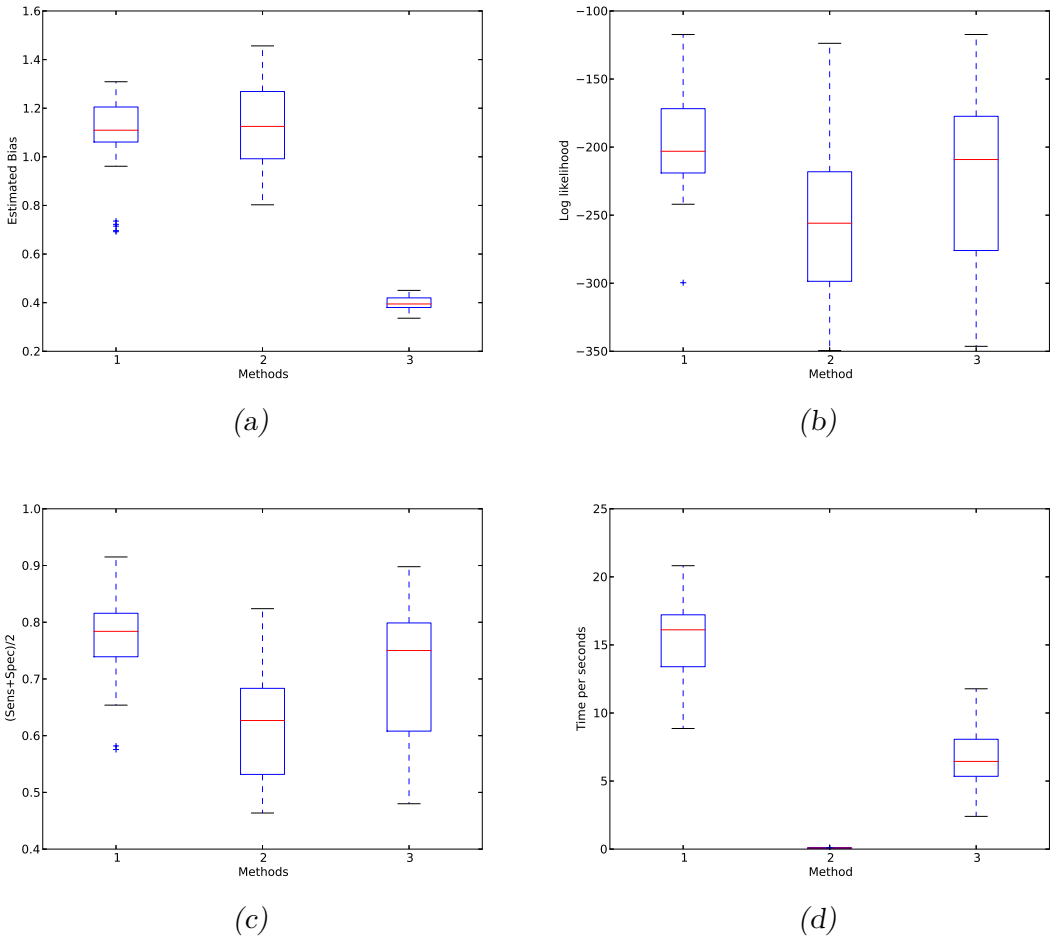


Fig. 3.1: Performance of the three modification methods for Stage 1. The figures show the box-whisker plots of the estimated biases of the parameter θ, the averages of specificity and sensitivity, the attained log-likelihoods, and time-costs for the three modifications.

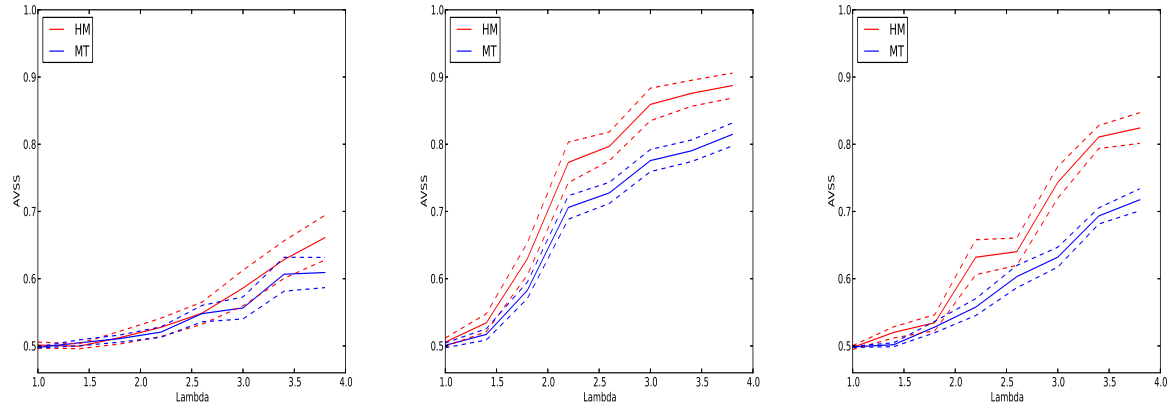
3.3.2 Performance of the proposed hybrid mixture approach

Note that the proposed hybrid mixture method is based on the prospective likelihood model although real data can be from retrospective studies. By the simulations below, we addressed whether the proposed hybrid mixture approach could outperform Zhu et al' multiple-testing procedure in both prospective (i.e., cohort) and retrospective (i.e., case-control) studies.

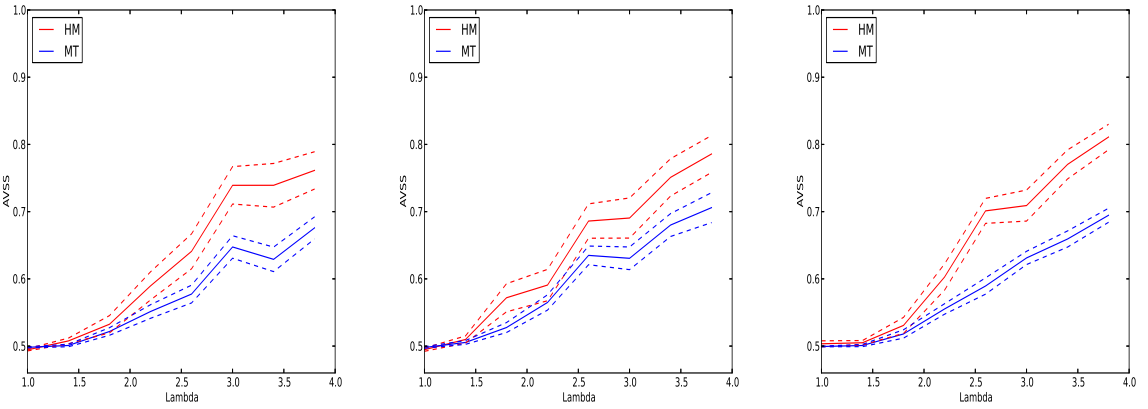
Setting 1 (cohort design): We generated 30 datasets, each with N_1 case-genotypes and N_0 control-genotypes. They were obtained by the following steps. In the first two steps, we adopted the same approach for generating $N_0 + N_1$ genotypes which contained m_r risk haplotypes as we did before. In the third step, we simulated

the disease status of each genotype by sampling from a Bernoulli distribution. The Bernoulli distribution took f_0, or $\lambda_1 f_0$, or λf_0 as a success probability according to whether the genotype contained zero, one or two risk haplotypes. We considered the three inheritance models coded by IM: the multiplicative (IM = 1), the dominant (IM = 2) and the recessive (IM = 3). Note that the values of (N_0, N_1) may vary across different datasets. We considered the scenarios with various combinations of $(N_0 + N_1, m_r, \text{IM}, f_0, \lambda)$, where $N_0 + N_1 = 3000, 5000$, $m_r = 5, 10, 20$, IM $= 1, 2, 3$, $f_0 = 0.1$, $\lambda = 1, 1.4, 1.8, 2.2, 2.6,\ 3, 3.4$, and 3.8 respectively.

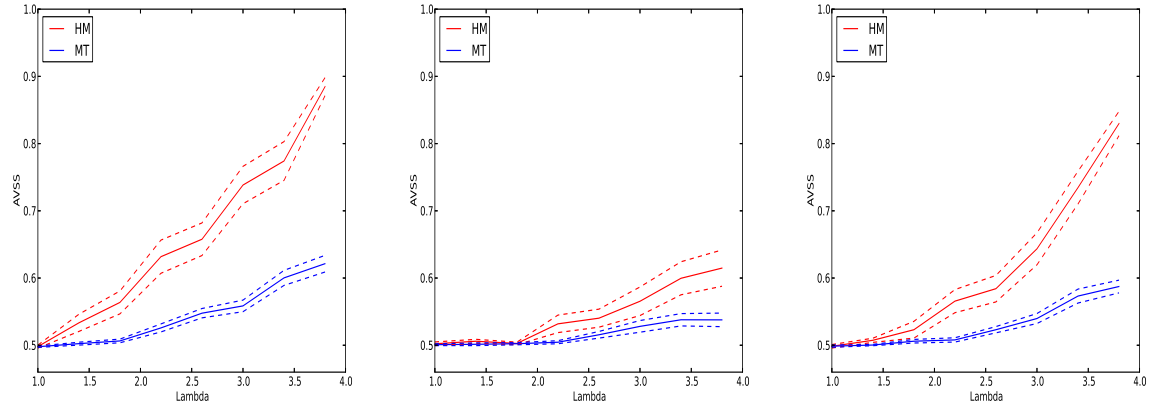
For each scenario, we applied both the hybrid mixture method and the multiple testing method to 30 datasets and calculated their AVSS values respectively. For each of the three inheritance models, we plotted the means of these AVSS values over 30 datasets against λ. In the plots of the figure 3.2, the red and the blue solid curves, showing means of the AVSS values (i.e., the values of (specificity and sensitivity)/2)) over 30 datasets, were plotted against the values of λ for the hybrid mixture method and the multiple testing method respectively. The two red dash curves are one standard error up and down from the red mean curves. Similarly, the two blue dash curves are one standard error up and down for blue mean curves. The plots in the columns from the top to the bottom are for the cases where there were 5, 10, and 20 risk haplotypes in the underlying haplotypes. While from the left to the right, the plots represent recessive, dominant and multiplicative mode of inheritance. The sample size of the cases and the controls altogether for each plot is 5000. The plots are based on different modes of inheritance. Similarly for the figure 3.3 except the considered sample size is 3000. On the cohort data, the hybrid mixture method performed substantially better than the multiple testing method in all the scenarios defined above.



(a) N=5000, m_r=5,rece (b) N=5000,m_r=5,dom (c) N=5000,m_r=5,mult

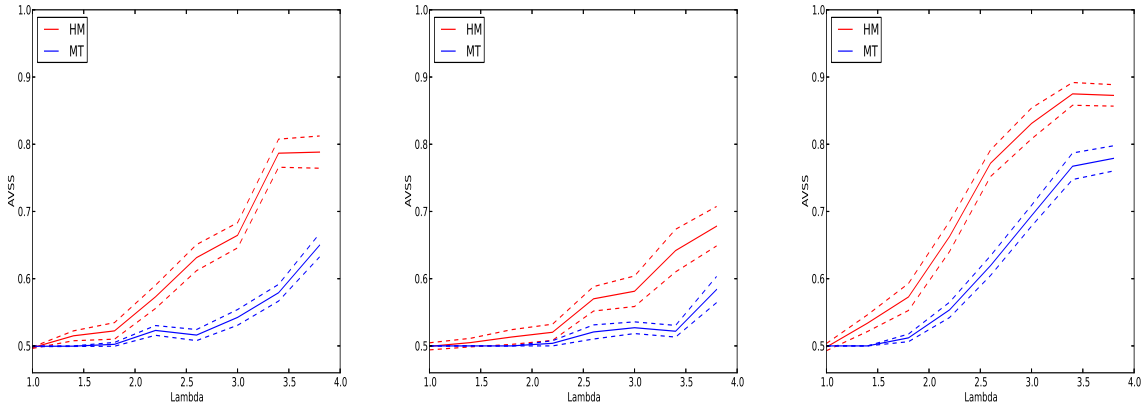


(d) N=5000,m_r=10,rece (e) N=5000,m_r=10,dom (f) N=5000,m_r=10,mult

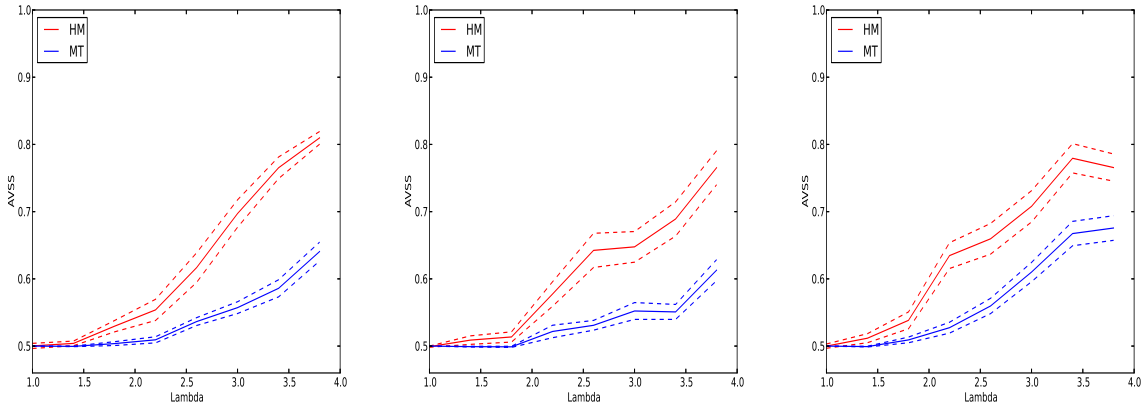


(g) N=5000,m_r=20,rece (h) N=5000,m_r=20,dom (i) N=5000,m_r=20,mult

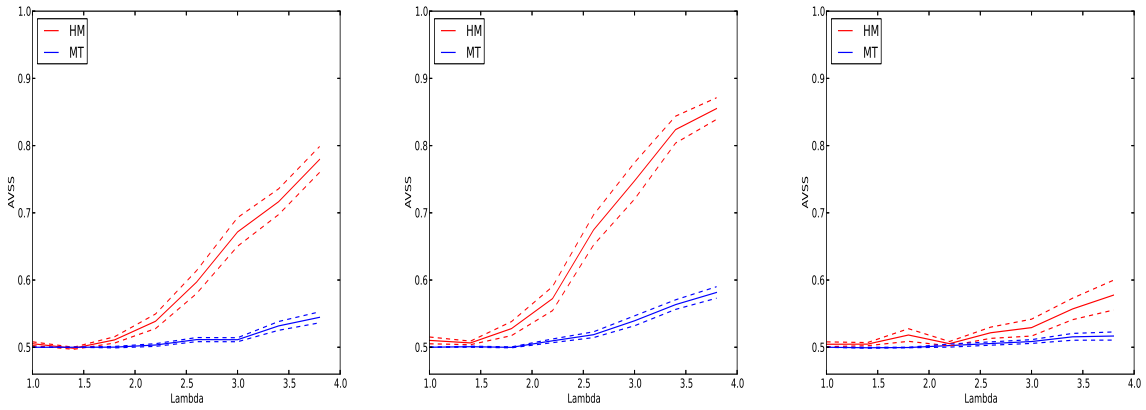
Fig. 3.2: Performances of the proposed hybrid mixture method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models with sample size $N = 5000$..



(a) N=3000,m_r=5,mult (b) N=3000,m_r=5,rece (c) N=3000,m_r=5,dom



(d) N=3000,m_r=10,mult (e) N=3000,m_r=10,rece (f) N=3000,m_r=10,dom



(g) N=3000,m_r=20,mult (h) N=3000,m_r=20,rece (i) N=3000,m_r=20,dom

Fig. 3.3: Performances of the proposed hybrid mixture method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models with sample size $N = 3000$..

Setting 2 (case-control design): We generated 30 datasets, each of which were simulated by the following two steps. Step 1, to generate N_1 case-genotypes, we first drew $2N_1$ haplotypes by the software MS with mutation rate of 2, of which m_r haplotypes were labeled as risk haplotypes. We then randomly paired these haplotypes to form N_1 case-genotypes. Let G_j, $1 \leq j \leq J$ be all the different genotypes contained in the N_1 cases and $r_{1j}, 1 \leq j \leq J$ be the retrospective frequencies. These case-genotypes formed three groups according to the number of risk haplotypes which each genotype contained: Each genotype in Groups 0, 1 and 2 contained two non-risk haplotypes, only one risk-haplotype, and two risk haplotypes respectively. Step 2, we generated N_0 control-genotypes, which also had genotypes G_j, $1 \leq j \leq J$ but with population retrospective frequencies $q_{0j}, 1 \leq j \leq J$. We first let $q_{0j}, 1 \leq j \leq J$ depend on the pre-specified constant d by

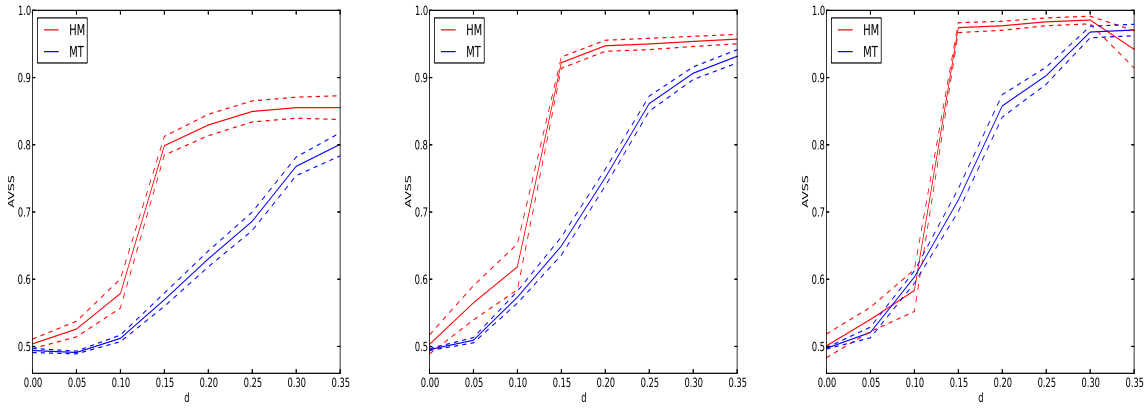
$$q_{0j} = \begin{cases} r_{1j}(1 - d/r_{1g_2}), & G_j \text{ belongs to Group 2} \\ r_{1j}(1 - 0.5d/r_{1g_1}), & G_j \text{ belongs to Group 1} \\ r_{1j}(1 + 1.5d/r_{1g_0}), & G_j \text{ belongs to Group 0} \end{cases}$$

where $r_{1g_k} = \sum_{G_j \in \text{ Group}_k} r_{1j}$ for $k = 0, 1, 2,$, and $d \geq 0$ is a parameter to reflect the effects of risk haplotypes on genotype frequencies. We simulated N_0 control-genotype counts from the multinomial model $\text{MN}(N_0, (q_{01}, ..., q_{0J})^T)$ and calculated the corresponding retrospective frequencies $r_{0j}, 1 \leq j \leq J$. We considered the cases where $m_r = 5, 10, 20,$ and $d = 0, 0.05, 0.1, 0.1,\ 0.15, 0.2, 0.25, 0.3,$ and 0.35 respectively.

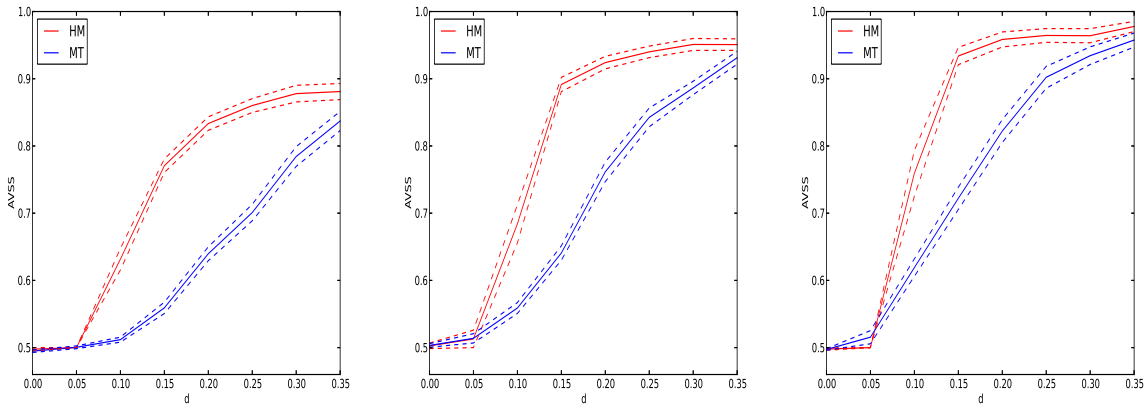
For each dataset, the cumulative genotype frequencies of Groups 0, 1, and 2 in controls are $r_{1g_0} + 1.5d$, $r_{1g_1} - 0.5d$, and $r_{1g_2} - d$ respectively, whereas the corresponding frequencies in cases are r_{g_0}, r_{g_1} and r_{g_2} respectively. This implies that due to the impacts of risk haplotypes, the cumulative frequencies of Groups 2 and 1 in cases have been increased compared to those in controls. The odds ratios between Groups 2 and 0 and between Group 1 and Group 0, $(1 + 1.5d/r_{g_0})/(1 - d/r_{g_2})$ and $(1 + 1.5d/r_{1g_0})/(1 - 0.5d/r_{1g_1})$, are larger than one. Similarly, the odds ratio between the risk haplotype group and the non-risk haplotype group can be expressed as $(1+1.25d/(r_{1g_0}+0.5r_{1g_1}))/(1-1.25d/(r_{1g_2}+0.5r_{1g_1}))$. All these ratios are increasing in d.

We applied the hybrid mixture method and the multiple testing method to these case-control data. In the figure 3.4, The plots in the columns from the left to the right are for the scenarios, where the underlying number of risk haplotypes $m_r = 20, 10,$ and 5. The top row stands for the cases, where $(N_0, N_1) = (3000, 2000)$, while the bottom row stands for the cases, where $(N_0, N_1) = (2000, 1000)$. In these

plots, the red and the blue solid curves show mean curves of the AVSS values over 30 datasets as functions of $d = 0, 0.05, 0.1, 0.1,\ 0.15, 0.2, 0.25, 0.3$, and 0.35 for the hybrid mixture method and the multiple testing method respectively. The dash curves are one standard error up or down from the mean curves. The results again demonstrate that the hybrid mixture method can be more powerful than the multiple testing method in detecting risk haplotypes.



(a) N1=2000, N0=3000,m_r=20 (b) N1=2000, N0=3000,m_r=10 (c) N1=2000, N0=3000,m_r=5



(d) N1=1000, N0=2000,m_r=20 (e) N1=1000, N0=2000,m_r=10 (f) N1=1000, N0=2000,m_r=5

Fig. 3.4: Performances of the proposed hybrid mixture and the multiple testing method on the case-control data.

3.3.3 *Performance of the proposed inheritance mode test*

For each of the three inheritance models, we generated 30 datasets. Each dataset was simulated as follows. Following the cohort design, we first simulated $N_0 + N_1$ genotypes, where the underlying haplotypes contained $m_r = 10$ risk-haplotypes and

followed the Hardy-Weinberg equilibrium. We then simulated their disease status by use of the inheritance models with $f_0 = 0.1$ and $\lambda = 1, 1.4, 1.8, 2.2, 2.6, 3, 3.4,$ and 3.8 respectively as we did in the previous subsection.

For each dataset, we calculated D_A and the optimal mode $\hat{a}$. We generated 100 parametric bootstrap samples of the genotype frequencies based on the mode $\hat{a}$ and calculated the corresponding values of the inheritance testing statistic, $D_A^{(k)}$, $k = 1, ..., 100$. Based on these values, we obtained the empirical p-value.

We calculated the success rates by counting how many times that $\hat{a}$ is the true mode over the 30 datasets for each λ. These success rates and the empirical p-values are displayed in Figure 3.5. In this figure, the plots in the columns from the left to the right are for the dominant, the multiplicative and the recessive models respectively. The top row shows the box-whisker plots of the empirical p-values (based on 100 bootstrap samples) against for the inheritance test statistic Dmin over 30 datasets, while the bottom row shows the success rate of identifying the true inheritance mode against over 30 datasets. The results indicate that the success rates are increasing as λ is increasing. The box-whisker plots in Figure 3.5 show that almost all the empirical p-values are above 0.20, suggesting that almost all the tests are not significant. Therefore, the bootstrap test has a very high power in finding the true inheritance modes in the data.

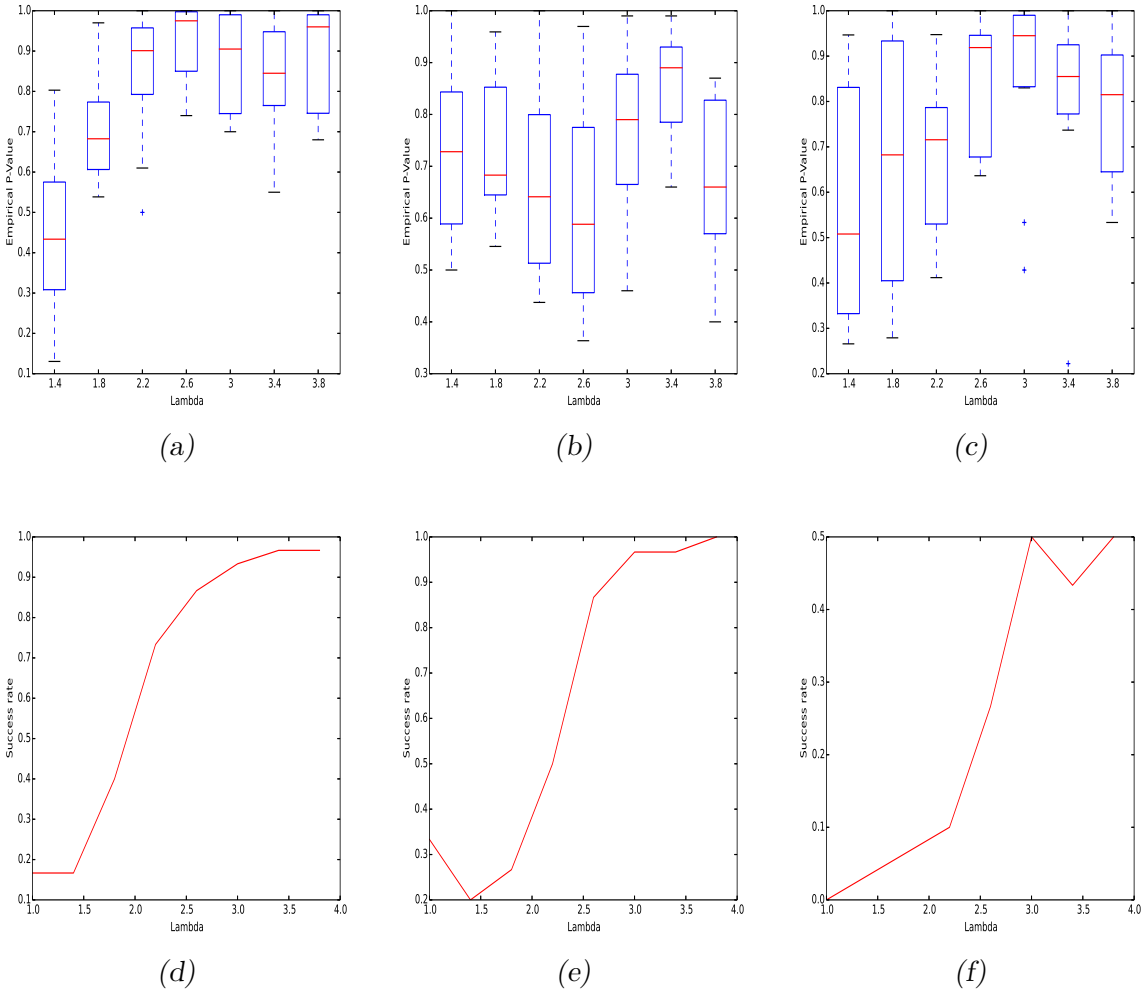


Fig. 3.5: Performances of the proposed test for inheritance patterns.

3.4 Quality control of haplotypes

The aim of this section is to choose a threshold for the p-values derived from the chi-square tests for the association of the population sub-structures with risk haplotypes.

We used Setting 2 in the simulation to generate 30 datasets for each of the following scenarios. For each d in $\{0, 0.05, 0.1\}$, we considered 6 scenarios.

Scenario 1: $(N_0, N_1) = (2000, 3000), m_r = 5$.

Scenario 2: $(N_0, N_1) = (2000, 3000), m_r = 10$.

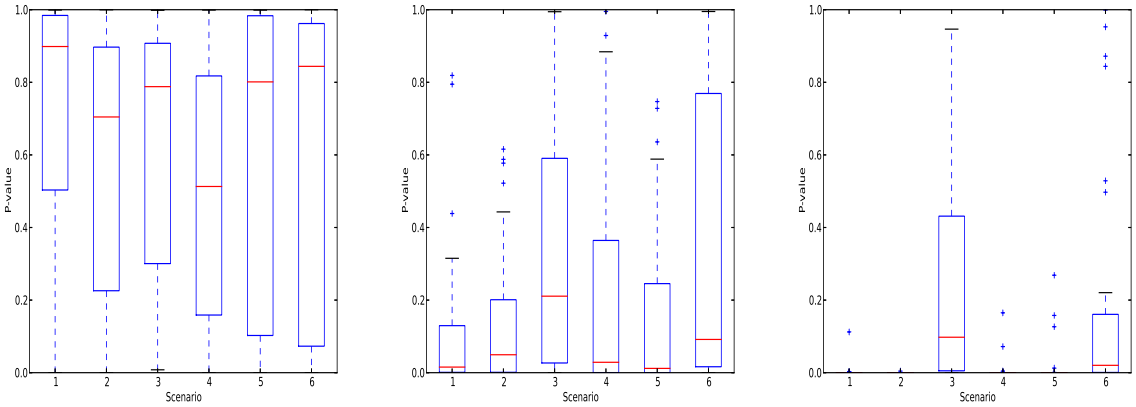
Scenario 3: $(N_0, N_1) = (2000, 3000), m_r = 20$.

Scenario 4: $(N_0, N_1) = (1000, 2000), m_r = 5$.

Scenario 5: $(N_0, N_1) = (1000, 2000), m_r = 10$.

Scenario 6: $(N_0, N_1) = (1000, 2000), m_r = 20$.

For each dataset, we first obtained a haplotype contingency table by using the software PHASE. We then calculated the p-value of chi-squared test of the association. We ended up with 30 p-values for each scenario. The box-whisker plots are displayed in Figure 3.6. In this figure, the plots from the left to right are for $d = 0, 0.05, 0.1$ respectively. The x-axis was labeled by 6 scenarios and y-axis was labeled by p-values.



(a) (b) (c)

Fig. 3.6: The box-whisker plots of p-values of the chi-squared tests on 30 datasets which represent the above six scenarios.

3.5 Real data analysis

We applied the proposed hybrid mixture approach to the GWAS genotype datasets on coronary artery disease (CAD) and hypertension (HT) obtained by Affymetrix 500K SNP chips in the WTCCC study (WTCCC, 2007). Each dataset contained 2000 unrelated cases as well as 3000 unrelated controls. The table (3.2) shows the original format of such data.

Tab. 3.2: The table shows the format of the genotype data format of WTCCC. The first column represents the SNP id, the second column represents individual id, the third column represents is the genotype of the corresponding at the corresponding individual and the score column shows the quality of SNPs calling.

SNP	SAMPLE	GENOTYPE	SCORE
rs1234567	WTCCC12345	CC	0.9262
rs1234568	WTCCC12345	TC	0.8650
rs1234569	WTCCC12345	AA	0.9117

The controls came from two sources: 1500 from the 1958 British Birth Cohort (58C) and 1500 from the three National UK Blood Services (NBS). There were about 500600 SNPs across the human genome. These data were downloaded from the WTCCC website. We first pre-processed the data by excluding the SNPs which meet one of the following criteria: (1) the HWE Fisher test p-value is less than 10^{-8} in controls; (2) the chi-square test p-value between 58C and NBS is less than 10^{-8}; (3) the minor allele frequency is less than 1%; (4) the calling score is less than 95%. After the exclusion, around 4897746 SNPs remained for the analysis. We divided the genome into regions (or blocks) of 8 SNPs according to their positions on the chromosomes, obtaining 61218 regions.

Note that the long block will dilute the effects of risk SNPs whereas the short block will miss interactions between SNPs. The block length of 8 was chosen to achieve a compromise between the above aspects. Also note that as we excluded the SNPs with bad callings, the numbers of cases and controls are varying across the different regions.

For all regions, we first reconstructed the haplotype pairs of genotypes by use of the software PHASE, to which we applied Stage 1 of the hybrid procedure. It led to 902909 haplotypes and 961942 haplotypes to be declared as risk haplotypes at Stage 1 for the CAD and the HT respectively. We then calculated the OR tests on these haplotypes at Stage 2. At Stage 2, According to the Bonferroni adjustment, the individual significance level was set at the levels of $0.05/902909 = 5.5 \times 10^{-8}$ and $0.05/961942 = 5.2 \times 10^{-8}$ for the CAD and the HT respectively.

These individual significance levels were then used to determine the thresholding level c_1 in the multiple OR thresholding, which is $c_1 = 5.3$.

After performing the proposed hybrid mixture procedure on the datasets, we obtained the estimated risk and non-risk haplotype sets, ($\hat{S}_r$ and $\hat{S}_{\bar{r}}$), for the CAD and the HT respectively.

Note that there were two sub-populations in controls. Any estimated risk haplotype which is significant in differing two control sub-populations should be viewed as an artifact. By using this, we made further quality control on the selected haplotypes by running the chi-square tests on the association of two control sub-populations with each selected risk haplotype. We eliminated these risk haplotypes whose p-values for the above chi-square tests were $< 30\%$. Here, 30% was chosen by the simulations as shown in the figure 3.6, aiming to filter out these artificial risk haplotypes with parameter $d \geq 0.05$. From the simulations, we can see that when $d = 0.05$, these p-values would be less than or equal to 0.30 most times.

Finally, we calculated the ORs for all the estimated haplotypes and thresholded them by using the bound

$$\exp(c_1\sqrt{1/(n_{0H}+0.5)+1/(n_{1H}+0.5)+1/(n_{0\bar{r}}+0.5)+1/(n_{1\bar{r}}+0.5)})$$

with $c_1 = 5.3$. This gave the final risk-haplotype set as displayed in Tables 3.3, 3.4, 3.5 and 3.6 below. In the tables, each haplotype has been assigned to a physically closest gene on the basis of the information provided the GWAS catalog and the genetic information from the British 1958 Birth cohort. See Welter et al. (2014) and the web page at http://www2.le.ac.uk/projects /birthcohort/1958bc. In the CAD case, we did rediscover the CAD risk gene CDKN2B and the risk haplotype $GGTGCCAG$ found by the previous study (WTCCC, 2007; Zhu et al., 2010). We also tested the inheritance modes for these risk haplotypes. Taking the gene CDKN2B as an example, we obtained $D_A = 0.4087$ with $\hat{a}$ = "dominant mode" and the empirical p-value of 0.97, suggesting that the hyplotype $GGTGCCAG$ in the gene followed the dominant inheritance mode.

Tab. 3.3: The predicted risk haplotypes for CAD by use of the WTCCC data. In the table, the P-values were derived from the chi-square test of the frequencies of H_i against the collapsed frequencies of the estimated non-risk haplotypes.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
1	202166400 − 202187685	$rs6692041 - rs1041311$	$AAATGGGA$	0.07815	0.05083	1.95856	2.8×10^{-13}	LOC284577
1	237650028 − 237672617	$rs6683639 - rs10802930$	$TCAAATGC$	0.05256	0.02763	2.57538	6.1×10^{-13}	RGS7
3	102073696 − 102093722	$rs973309 - rs4928094$	$TAACCTTT$	0.07591	0.06898	7.73184	5.6×10^{-15}	ABI3BP
3	142488272 − 142537277	$rs7643346 - rs2871887$	$CGCCCATC$	0.05008	0.03809	11.90617	2.0×10^{-15}	ACPL2
3	147806667 − 147828893	$rs17433833 - rs17434589$	$CCGGGGGC$	0.03363	0.01291	3.14753	3.2×10^{-15}	PLSCR5
4	132550 − 344051	$rs11735742 - rs17719492$	$TGGCACTC$	0.05993	0.04793	1.9902	7.6×10^{-11}	LOC654254
4	4464610 − 4499426	$rs16835627 - rs4234727$	$TCGAGCAT$	0.04072	0.0251	3.92994	1.1×10^{-14}	ZNF509
			$CTAAGCAT$	0.09413	0.07343	3.10696	1.2×10^{-13}	
4	180659963 − 180699763	$rs6811556 - rs17090633$	$CCCCCACT$	0.01782	0.00755	7.33583	5.5×10^{-15}	LOC391719
5	157267571 − 157303032	$rs10071157 - rs17055168$	$GTGAGCAA$	0.02135	0.00701	3.93074	4.0×10^{-13}	CLINT1
7	77725471 − 77739291	$rs10485891 - rs7803705$	$AACATGCG$	0.03652	0.04027	3.67364	6.8×10^{-13}	MAGI2
			$AACATGTA$	0.01312	0.01117	4.76163	2.1×10^{-11}	
			$AGTGCACA$	0.01312	0.00846	6.27027	1.2×10^{-14}	
7	130749877 − 130784667	$rs4728224 - rs4728225$	$AGAACCGG$	0.14061	0.13197	4.05796	1.0×10^{-12}	LOC647030
8	104190450 − 104202402	$rs2515173 - rs3019159$	$GGCCATCT$	0.14195	0.08768	2.20746	2.5×10^{-27}	BAALC
9	22088619 − 22120515	$rs2891168 - rs10965245$	$GGTGCCAG$	0.34939	0.29298	1.90115	2.7×10^{-11}	CDKN2B
9	77341767 − 77366988	$rs2889774 - rs3780296$	$ATGAGAGT$	0.01936	0.01072	5.31687	5.0×10^{-18}	GNA14
			$ATGAAGAC$	0.03898	0.03923	2.93116	4.5×10^{-13}	
			$ATGGAAAT$	0.06672	0.042	4.68028	4.9×10^{-30}	
			$GCGAAGAT$	0.14207	0.14656	2.85712	4.9×10^{-19}	
9	131714465 − 131751663	$rs3012758 - rs11243551$	$CGAATTGC$	0.06641	0.04652	2.41478	6.2×10^{-13}	RAPGEF1
			$CGAACTGC$	0.02448	0.01227	3.36929	4.4×10^{-12}	
10	64409674 − 64442476	$rs1509952 - rs2842286$	$TTTCTTAC$	0.02299	0.0073	9.37291	1.6×10^{-16}	NRBF2
10	112527724 − 112597595	$rs17763100 - rs1341055$	$GCCTCCCG$	0.07752	0.07383	1.85031	6.2×10^{-11}	RBM20
			$ACCTCCCG$	0.24688	0.21703	2.00368	6.7×10^{-23}	
10	129835144 − 129894934	$rs11016102 - rs1335014$	$AAGAACTT$	0.02987	0.01529	4.40461	6.2×10^{-14}	MKI67
11	36361306 − 36410807	$rs330255 - rs331485$	$GCGATTAA$	0.0309	0.00779	4.87953	1.5×10^{-21}	FLJ14213
11	133079508 − 133113640	$rs4937817 - rs4937826$	$GTAGTGCC$	0.04216	0.02425	2.69929	5.9×10^{-17}	LOC646522
			$CCGGCCCG$	0.05747	0.04018	2.22186	1.4×10^{-15}	
			$GTAGCCCG$	0.04001	0.02779	2.23683	8.3×10^{-12}	

Tab. 3.4: The continuation of Table 3.3.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
11	133914862 – 133953680	$rs12417998 - rs10894845$	$GTTAGCCC$	0.12907	0.13389	3.70503	1.4×10^{-12}	IQSEC3
			$GTTAATCC$	0.09778	0.09576	3.92451	3.3×10^{-13}	
			$GTCAGCTC$	0.06932	0.07079	3.76457	7.6×10^{-12}	
12	24250132 – 24288211	$rs3922562 - rs17412555$	$CTGTGCCT$	0.07253	0.06027	5.51363	6.0×10^{-15}	SOX5
			$TCGCGCCC$	0.05454	0.03857	6.47638	9.9×10^{-17}	
			$TCGCGTCC$	0.02399	0.01788	6.14651	1.0×10^{-12}	
12	51469295 – 51501190	$rs17738862 - rs876407$	$CACCCTCG$	0.14455	0.13704	2.25981	2.6×10^{-13}	KRT3
12	127083338 – 127105747	$rs7960047 - rs9668398$	$GTGCGTCT$	0.06573	0.06076	3.67491	2.7×10^{-15}	TMEM132C
15	37962389 – 38014169	$rs11633436 - rs534757$	$TTACAACC$	0.07798	0.03763	2.66998	3.9×10^{-26}	EIF2AK4
16	79852394 – 79892297	$rs6564863 - rs11639552$	$TTCGTTAT$	0.02663	0.01053	5.1576	7.7×10^{-16}	BCMO1
17	29052246 – 29089136	$rs2046899 - rs17783280$	$AGTCAATC$	0.11305	0.0966	2.10899	5.7×10^{-14}	LOC646202
17	52973696 – 53057256	$rs17834557 - rs3744089$	$TGGTTAAC$	0.05825	0.03915	2.15515	8.7×10^{-14}	MSI2
18	9649377 – 9700554	$rs1965881 - rs1455587$	$TCACATGT$	0.06243	0.04149	2.15776	6.3×10^{-13}	RAB31
18	60647495 – 60688045	$rs1595904 - rs17678507$	$CAGTATAT$	0.09403	0.0848	2.55691	1.2×10^{-11}	C18orf20
18	72313651 – 72356779	$rs17059443 - rs8084536$	$GCGAGACC$	0.08958	0.08373	2.43635	1.0×10^{-11}	FLJ44313
19	4625799 – 4746342	$rs11670570 - rs1044409$	$AGCAACCG$	0.05419	0.02332	3.3426	6.7×10^{-25}	DPP9
19	56075162 – 56127664	$rs187930 - rs1654545$	$ACATGTGA$	0.03532	0.02898	7.24575	3.3×10^{-13}	KLK2
19	58460745 – 58519652	$rs1978611 - rs7408137$	$AGGTAGTG$	0.05628	0.042	1.99812	4.0×10^{-12}	VN1R4
22	35324014 – 35335429	$rs7410412 - rs12160203$	$TCCTAGGG$	0.44488	0.50199	3.09116	1.6×10^{-21}	CACNG2
			$GCCTAGAG$	0.03358	0.02891	4.05372	6.2×10^{-17}	

Tab. 3.5: The predicted risk haplotypes of hypertension by use of WTCCC data. In the table, the P-values were derived from the chi-square test of the frequencies of H_i against the collapsed frequencies of the estimated non-risk haplotypes.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
1	1586208 – 1753641	$rs6603791 - rs2272908$	$AACCCATC$	0.03406	0.01973	2.45812	2.7×10^{-12}	SSU72
1	227569611 – 227620956	$rs7514972 - rs9431663$	$CGTATAGG$	0.03377	0.00926	7.08695	2.8×10^{-32}	TRIM67
1	227914995 – 228040530	$rs16854388 - rs1655296$	$CAAGGTAG$	0.04372	0.04622	2.90643	1.9×10^{-13}	TSNAX
1	236986859 – 237020204	$rs12137158 - rs16840310$	$GCTGTGGG$	0.02424	0.01534	2.95857	1.7×10^{-11}	GREM2
			$ATTTAGGG$	0.08733	0.05437	3.00646	3.0×10^{-26}	
			$GCTTTGAG$	0.0756	0.06745	2.09936	1.1×10^{-12}	
3	101569551 – 101696774	$rs277640 - rs4928098$	$CCCAGGCG$	0.02137	0.00908	6.27332	1.9×10^{-13}	TOMM70A
3	142488272 – 142537277	$rs7643346 - rs2871887$	$AGCTCATC$	0.17323	0.17868	2.2344	4.4×10^{-11}	ACPL2
3	142878508 – 142912781	$rs12485838 - rs16851691$	$GCATAGAG$	0.02089	0.00902	5.09818	1.3×10^{-13}	LOC646730
4	21080985 – 21131665	$rs1495517 - rs358574$	$GTCGCACG$	0.05716	0.04649	7.36766	4.2×10^{-13}	KCNIP4
			$GTTGCACG$	0.06033	0.04669	7.74264	7.1×10^{-14}	
4	23359572 – 23389742	$rs10008808 - rs1976201$	$AGTTCTTA$	0.03874	0.01347	3.68417	1.5×10^{-20}	PPARGC1A
5	10695437 – 10746687	$rs2062200 - rs6891527$	$GTCACACG$	0.16002	0.14866	6.18799	2.8×10^{-23}	LOC651746
5	32084851 – 32103155	$rs438834 - rs10065850$	$TGCTCCCA$	0.02254	0.01065	14.40157	1.6×10^{-24}	PDZD2
6	139560239 – 139612833	$rs7765885 - rs9495394$	$GCGCAACG$	0.0487	0.01774	4.54602	2.2×10^{-35}	HECA
			$ACGAAATG$	0.01641	0.00709	3.82046	6.4×10^{-12}	
			$GTACAATA$	0.14141	0.13391	1.75292	4.9×10^{-16}	
6	139693238 – 139758634	$rs11155050 - rs9373237$	$TTGCGGCT$	0.01924	0.00686	5.16669	1.1×10^{-14}	TXLNB
			$CTAAGATT$	0.25795	0.24508	1.9524	6.6×10^{-11}	
7	48232027 – 48237897	$rs17729647 - rs2362301$	$AGACTGGT$	0.07901	0.07156	3.41729	4.7×10^{-15}	ABCA13
			$AGATTGAC$	0.03345	0.02897	3.57621	3.7×10^{-12}	
			$AGATTGGC$	0.35755	0.38319	2.88725	3.0×10^{-14}	
7	77695246 – 77717237	$rs2215379 - rs4515471$	$CTTAAAAA$	0.03102	0.01998	4.32524	2.1×10^{-21}	MAGI2
			$TCTAAAAA$	0.02943	0.01786	4.58962	5.0×10^{-22}	
			$CTTGGAAA$	0.02094	0.01061	5.49009	8.4×10^{-21}	
			$CCTAGAAA$	0.05541	0.05534	2.79199	2.4×10^{-16}	
			$CCGAAAAA$	0.13203	0.13667	2.6926	4.0×10^{-21}	
9	77269212 – 77301387	$rs17063627 - rs7032444$	$GCGGACAG$	0.03393	0.01858	3.58867	1.6×10^{-12}	GNA14
10	119535731 – 119568729	$rs4752106 - rs10787797$	$TATTCACA$	0.09968	0.06304	2.91842	4.8×10^{-19}	RAB11FIP2

Tab. 3.6: The continuation of Table 3.5.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
11	125683058 – 125763272	$rs2096915 - rs7118117$	$CACACGAG$	0.07736	0.04727	2.42988	4.9×10^{-12}	ST3GAL4
12	27155055 – 27179334	$rs841636 - rs841613$	$TAAAGGGT$	0.05414	0.04075	2.81343	7.5×10^{-15}	LOC729222
12	112703139 – 112738033	$rs11066758 - rs7137339$	$GGGGTCCC$	0.06128	0.04048	2.52574	2.3×10^{-18}	RBM19
12	114038450 – 114074493	$rs1828384 - rs35346$	$TGTACCTG$	0.09952	0.10526	3.07564	1.5×10^{-11}	TBX3
			$TCCAATTG$	0.04718	0.03821	4.01761	2.2×10^{-13}	
13	23708179 – 23726596	$rs881428 - rs2760374$	$AGAAGTTT$	0.12142	0.07922	1.89748	5.7×10^{-19}	SPATA13
			$GAAAGCTT$	0.2454	0.19993	1.51979	6.8×10^{-15}	
13	70170848 – 70209722	$rs17087430 - rs12876111$	$CGGGTTAT$	0.13996	0.13226	3.39099	3.0×10^{-14}	ATXN8OS
			$CGGGTCCT$	0.02217	0.01356	5.23473	1.9×10^{-14}	
			$CGGGTCAT$	0.13141	0.13526	3.11367	2.3×10^{-12}	
			$CGGACTCT$	0.04728	0.0398	3.8075	1.9×10^{-13}	
14	21674996 – 21704333	$rs12050442 - rs1894369$	$GGGGTTAC$	0.03075	0.00968	6.13598	8.7×10^{-19}	TRA@
14	36411583 – 36421982	$rs10872897 - rs2564848$	$ATCCACTT$	0.02299	0.00637	4.45891	8.9×10^{-16}	SLC25A21
			$TACCTCCC$	0.02712	0.01101	3.05584	1.8×10^{-12}	
16	4881048 – 4960784	$rs760117 - rs9937749$	$CTTCCCCA$	0.0847	0.08126	4.18237	1.8×10^{-12}	SEC14L5
16	17231173 – 17272606	$rs754067 - rs17277691$	$CGGACCCT$	0.02658	0.02179	3.37015	1.1×10^{-11}	XYLT1
17	69565860 – 69595387	$rs7406930 - rs8080915$	$CTGTACGC$	0.0413	0.02484	2.54279	8.3×10^{-14}	RPL38
19	3315188 – 3432578	$rs758257 - rs1860192$	$GTTTGATT$	0.27769	0.23516	1.99935	3.4×10^{-28}	NFIC
19	8475735 – 8540766	$rs2967603 - rs11259990$	$CCGCTCTT$	0.06824	0.04351	3.23984	2.1×10^{-17}	ZNF414
19	17595848 – 17649789	$rs10419511 - rs7252308$	$TTGGTGTG$	0.07791	0.05267	2.40001	1.9×10^{-23}	UNC13A
			$TTGGTATG$	0.04536	0.01971	3.72872	2.3×10^{-28}	
19	38822176 – 38857206	$rs2059876 - rs16968366$	$CAAATGCG$	0.06455	0.05252	2.83486	6.2×10^{-20}	CHST8
20	10019135 – 10038764	$rs552048 - rs670562$	$TATGAGGG$	0.04043	0.02307	7.32891	4.5×10^{-21}	ANKRD5
			$TATAAGAA$	0.03726	0.03549	4.39728	2.7×10^{-12}	
			$TGTGAGGG$	0.27299	0.294	3.88507	1.0×10^{-13}	
			$TGTATGGG$	0.19239	0.1808	4.4523	3.1×10^{-16}	

3.6 Discussion and conclusion

The GWAS and sequencing studies have produced a huge amount of high-dimensional data. Analyzing these data offers many challenges to statistical inference. Several empirical studies have demonstrated the superiority of SNP region-based association analysis over single-SNP strategy (see Zakharov et al., 2013 and reference therein). However, even restricted to a region, we may still obtain many sparsely distributed haplotypes derived from phasing the genotypes. In the presence of sparsely distributed haplotypes, haplotype clustering is very useful for performing statistical analysis on the data. Most of the existing methods of haplotype clustering are heuristic and not disease-penetrance based. To overcome this drawback, we have proposed a hybrid mixture model-based approach for grouping and identifying risk haplotypes. The key ingredient of the approach is a prospective mixture model with priors. The proposal includes two stages: in the stage 1, one groups haplotypes and therefore reduce the haplotype sparsity, while in the second stage, one conducts a two-sample Z-test based screening on the haplotypes derived from the previous stage. We have also provided a test for genetic inheritance modes.

We have examined the performance of the proposed procedure by a theoreti-

cal analysis, simulations and a real data analysis. We have showed that under the Hardy-Weinberg equilibrium, the risk haplotype group is identifiable if genotype relative risk is not equal to one. Compared to the standard multiple Z-testing method, the proposed procedure is more efficient in terms of sensitivity and specificity. We applied our procedures to the WTCCC CAD and hypertension data, rediscovering some existing risk gene and haplotypes and identifying many more risk haplotypes than did the multiple Z-test based approach. This is not surprising as the simulations have already demonstrated that the model-based clustering often performs better than does the multiple Z-test approach.

We note that the proposed method is applicable to address the problem of column clustering (i.e., collapsing) in analyzing contingency table data.

4. GENOTYPE MIXTURE MODEL-BASED APPROACH (GM)

4.1 Introduction

The advanced genotyping technology has made it possible to conduct genome-wide association studies (GWAS) on complex diseases in recent years (Hindorff et al., 2009; Stranger et al., 2011). Genome-wide association studies systematically analyze genetic variation across the genome by its effects on phenotypic traits. The early landmark study using these technologies was the Wellcome Trust Case Control Consortium (WTCCC), which reported genetic association results for over 500,000 single nucleotide polymorphisms (SNPs) in seven disease sample sets of 2000 individuals each and 3000 control individuals (WTCCC, 2007). Most of these studies were based on the so-called common-disease-common-variant hypothesis that the variants being sought are common to many individuals with the disease. To date, these studies have identified hundreds of signposts associated with disease. But the search for causative variants derived from them has been remarkably less successful, with only a handful of causative variants discovered in follow-up sequencing studies. Many of the variants found have had only a weak effect on the risk of disease and therefore explained only a small proportion of the risk. Moreover, the signals in these studies might not always be pointing to a few common genetic variants but instead to many rare variants, each of which causes relatively few cases (Robinson, 2010; Li et al., 2010). The rapid increase in the number and the volume of GWAS provides an unprecedented opportunity to examine effects of rare variants on disease susceptibility. This also gives rise to a challenging problem of search for multiple variant sets in a high-dimensional genotype space. A popular strategy, suggested by the block-like structure of the human genome, is to segment each chromosome into a list of genetically meaningful regions. The multilocus haplotype, the ordered allele sequences on a chromosome, provides a unit of analysis for capturing linear and non-linear correlations among variants (Schaid et al., 2002; Zhang et al., 2003; Greevenbroek et al., 2008; Li et al., 2011). Unfortunately, haplotypes are often unknown and sparsely distributed. Many existing procedures suffers from the problem caused by sparsely distributed genotypes. Direct, laboratory-based haplotyping

to infer the unknown phase are expensive ways to obtain haplotypes. So, people prefer to infer haplotypes from observed genotypes by using the computational software such as PHASE (Stephens et al., 2001; Scheet et al., 2006). To deal with the haplotype distribution sparsity, a number of haplotype clustering methods have been developed in literature (Molitor et al., 2003; Tzeng et al., 2006; Browning and Browning, 2007; Zhu et al., 2010, and references therein). However, computational inferred haplotypes may contain both true and false haplotypes, resulting in a high false discovery rate of risk haplotypes.

Here, to deal with the above issue we propose a finite mixture model for directly clustering genotypes on the basis of their prospective frequencies. The rationale behind the proposal is as follows. We arrange the genotype frequencies derived from a case-control study by a contingency table, where rows stand for the disease status (case or control) and columns for genotypes. Then, we can directly assess whether two genotypes belong to the same group by their column similarity in the table. Formally, we fit each column by a binomial distribution with the disease-penetrance as the success probability, inferring the grouping of these columns through use of three-component binomial mixtures. The main advantage of the proposed model over the other existing methods is that it can avoid haplotyping-error effects on grouping rare haplotypes. Moreover, using the estimated prospective frequencies derived from a retrospective study to estimate genotype (and haplotype) disease odds ratio is known to be asymptotically consistent even though the prospective frequency estimators may not be (Prentice and Pyke, 1979).

We employ the expectation-maximization (EM) algorithm to calculate the maximum likelihood estimator for the proposed mixture model. The EM algorithm can guarantee monotone convergence to a local maximum. On the other hand, it needs to choose initial values in order to reach a local maximum which is close to the global maximum. The existing methods for initialization include: multiple random initializations, initially grouping the data and among others (Karlis and Xekalaki,2003). In this Chapter, we propose a new initialization procedure by grouping the estimated genotype frequencies. We conduct simulation studies on the proposed method in both prospective and retrospective design settings, showing that our method can outperform Zhu et al.'s approach in most cases. We also apply both the proposed method and Zhu et al's method to the Coronary Artery Disease (CAD) and Hypertension (HT) data in the Wellcome Trust Case Control Consortium (WTCCC), identifying potential risk haplotypes for each pre-specified chromosomal region.

The rest of the chapter is organized as follows. The proposed methodology is introduced in Section 4.2. The simulation studies and real data applications are

presented in Sections 4.3 and 4.4. Discussions and conclusion are made in Section 4.5.

4.2 Methodology

Consider a case-control sample with N_0 controls and N_1 cases, typed at m SNP markers in a candidate region, yielding unphased genotypes $\mathbf{G}$. Suppose that $\mathbf{G}$ contains distinct genotypes $G_j, 1 \leq j \leq J^*$ with counts N_{0j}, N_{1j} in controls and cases respectively. To tackle the issue of extremely rare genotypes, we first collapsed these genotypes by defining the set

$$G_c = \left\{ G_j | N_{0j} = 0 \text{ or } N_{1j} = 0 \text{ or } \frac{N_{0j} + N_{1j}}{N_0 + N_1} \leq 0.005, j = 1, ..., J^* \right\},$$

where we say that G_j is extremely rare if its prospective frequency is less than 0.05%. With a slight abuse of notation, we still denote these non-extreme genotypes as $G_1, ..., G_{J-1}$ with accounts $N_{0j}, N_{1j}, 1 \leq j \leq J-1$, and the set G_c by G_J with the collapsed account N_{0J} and N_{1J} in controls and cases respectively. We write $\mathbf{N} = \{(N_{0j}, N_{1j}) : 1 \leq j \leq J\}$ and rewrite $\mathbf{G} = \{G_1, ..., G_J\}$. Then, the prospective frequencies of G_j in the controls and cases can be estimated by

$$\hat{p}_{0j} = \frac{N_{0j}}{N_{0j} + N_{1j}}, \quad \hat{p}_{1j} = \frac{N_{1j}}{N_{0j} + N_{1j}}$$

respectively. Let $\mathbf{H}^2$ denote all haplotype pairs reconstructed from $\mathbf{G}$ by using the software PHASE (Stephens and Donnelly, 2003).

4.2.1 Two-stage procedure

We introduce the following two-stage approach for finding risk haplotypes. In Stage 1, genotypes are clustered and risk genotypes are derived, whereas in Stage 2 the odds ratio thresholding is employed to infer risk-haplotypes. As the reconstructed haplotypes may contain errors, to avoid the effect of hapolotying errors on clustering, we co-classify genotypes instead of the inferred haplotypes in Stage 1. The details are given below.

Stage 1 (Genotype clustering): We assume that haplotypes can be annotated by two categories: risk and non-risk, where non-risk category include both neutral and protective risk haplotypes. As each genotype consists of a haplotype pair, the

observed genotypes can be clustered into three categories according to the numbers of risk haplotypes which they have. In light of the above fact, we fit the following three-component binomial mixture model to the genotype account data:

$$f((N_{0j}, N_{1j})^T|\theta) = \pi_0 f((N_{0j}, N_{1j})^T|q_0) + \pi_1 f((N_{0j}, N_{1j})^T|q_1) + \pi_2 f((N_{0j}, N_{1j})^T|q_2),$$

where $\theta = (q_0, q_1, q_2, \pi_0, \pi_1)^T$ with $0 \leq q_\nu \leq 1, 0 \leq \pi_\nu \leq 1, \nu = 0, 1, 2, \pi_0+\pi_1+\pi_2 = 1$, and

$$f((N_{0j}, N_{1j})^T|q_\nu) = \binom{N_j}{N_{1j}} q_\nu^{N_{1j}}(1-q_\nu)^{N_{0j}}, \nu = 0, 1, 2; N_j = N_{0j} + N_{1j}.$$

The (incomplete) likelihood of θ given data $\mathbf{N}$ can be calculated by

$$L(\theta|\mathbf{N}) = \prod_{j=1}^{J} f((N_{0j}, N_{1j})^T|\theta).$$

We use the so-called expectation-maximization algorithm (McLachlan and Basford, 1988) to calculate the maximum likelihood estimator (MLE) $\hat{\theta}$. To this end, we introduce the following complete log-likelihood

$$l(\theta|\mathbf{N}, \mathbf{I}) = \sum_{j=1}^{J}\sum_{\nu=0}^{2} I_{\nu j} \log \left[\pi_\nu \ f((N_{0j}, N_{1j})^T|q_\nu)\right],$$

where $\mathbf{I} = \{(I_{0j}, I_{1j}, I_{2j})^T : 1 \leq j \leq J\}$ and $(I_{0j}, I_{1j}, I_{2j})^T$ are unknown group membership indicators defined by

$$I_{\nu j} = \begin{cases} 1, & \text{if } G_j \text{ in the group } \nu \\ 0, & \text{otherwise} \end{cases} \qquad \nu = 0, 1, 2.$$

The EM algorithm consists of two steps.

E-Step: Given the current estimator $\theta^{(t)}$ and the data, the conditional expectation of the complete log-likelihood can be calculated by

$$Q(\theta, \theta^{(t)}) = E\left[l(\theta|\mathbf{N}, \mathbf{I})|\mathbf{N}, \theta^{(t)}\}\right] = \sum_{j=1}^{J}\sum_{\nu=0}^{2} \tau_{\nu j}^{(t)} \log\left[\pi_\nu^{(t)} f((N_{0j}, N_{1j})^T|q_\nu^{(t)})\right],$$

where the expectation is taken with respect to the distribution of $\mathbf{I}$ and the estimated posterior probability of the j-th genotype being in the group ν, $\tau_{\nu j}^{(t)}$ admits

$$\tau_{\nu j}^{(t)} = P(I_{\nu j} = 1|(N_{0j}, N_{1j})^T, \theta^{(t)}) = \frac{\pi_\nu^{(t)} \ f((N_{0j}, N_{1j})^T|q_\nu^{(t)})}{\sum_{\nu=0}^{2} \pi_\nu^{(t)} f((N_{0j}, N_{1j})^T|q_\nu^{(t)})}.$$

M-Step: We update the current estimate $\theta^{(t)}$ by maximizing Q with respect to θ. This is equivalent to solving the following equations

$$\frac{\partial Q}{\partial \pi_\nu} = 0, \quad \frac{\partial Q}{\partial q_\nu} = 0, \ \nu = 0, 1, 2,$$

subject to $\pi_0 + \pi_1 + \pi_2 = 1$. For $\nu = 0, 1, 2$, we obtain the updated estimate $\theta^{(t+1)}$ via

$$\pi_\nu^{(t+1)} = \sum_{j=1}^{J} \tau_{\nu j}^{(t)} / J, \quad q_\nu^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{\nu j}^{(t)} N_{1j}}{\sum_{i=1}^{J} \tau_{\nu j}^{(t)} N_j}.$$

The existing EM theory suggests that the value of the log-likelihood function at the updated estimate is not decreasing in the sense that $l(\theta^{(t+1)}|\mathbf{N}) \geq l(\theta^{(t)}|\mathbf{N})$. We alternatively repeat the E- and M-steps until $l(\theta^{(t+1)}|\mathbf{N}) - l(\theta^{(t)}|\mathbf{N})$ is less than a pre-specified number η, say $\eta = 0.0001$.

Note that if the j-th genotype is not risk to the disease, then $X = (N_{0j} + N_{1j})\hat{p}_{1j}$ approximately follows a binomial distribution $B \sim f((N_{0j}, N_{1j})^T|\hat{q}_0)$. In light of this fact, the risk-genotype group (which consists of genotypes with at least one risk haplotype) can be estimated by

$$\mathbf{G}_r = \{G_j : \hat{p}_{1j} > w_j, j = 1, ..., J\},$$

where

$$w_j = \hat{q}_0 + \mu_j \sqrt{\hat{q}_0(1 - \hat{q}_0)/(N_{0j} + N_{1j})}$$

satisfying

$$P\left(X \geq (N_{0j} + N_{1j})w_j\right) < \varepsilon, \tag{4.1}$$

where ε is a pre-specified constant. In the simulation studies later, around 100 different genotypes will be involved in each dataset. Using the Bonferroni correction, we set $\varepsilon = 0.05/J$ so that the total probability of type I errors involved in the thresholding is less than 0.0005. Similarly, in the real data analysis section below, we will use the Bonferroni correction to set a different value of ε.

Stage 2 (haplotype thresholding): We introduce the following approach for identifying risk haplotypes. Let $\mathbf{H}_a^2$ be all haplotype pairs corresponding to $\mathbf{G}_r$, which are derived from $\mathbf{H}^2$ directly by taking advantage that $\mathbf{G}_r$ is a subset of $\mathbf{G}$. Let $\mathbf{H}_a = (h_1, ..., h_K)^T$ be all the distinct haplotypes in $\mathbf{H}_a^2$ with counts n_{0k} and n_{1k}, $k = 1, ..., K$ in controls and cases respectively. The subset $\mathbf{H}_a - \{h_k\}$ represents

non-risk background with controls and cases counts

$$n_{0\bar{r}} = \sum_{h_t \notin \mathbf{H}_a} n_{0t}, \quad n_{1\bar{r}} = \sum_{h_t \notin \mathbf{H}_a} n_{1t},$$

respectively. Note that $\mathbf{H}_a$ may contain non-risk haplotypes when $\mathbf{G}_r$ carries genotypes of a risk haplotype paired with a non-risk haplotype. For example, in the so-called dominant inheritance mode, risk haplotypes are often paired with non-risk haplotypes in producing genotypes. Therefore, to find risk haplotypes, we need to further threshold $\mathbf{H}_a$. It is well-known that the prospective frequency-based penetrance estimators with case-control data can be biased. However, the odds ratio estimator based on the prospective frequencies is asymptotically unbiased (Prentice and Pyke, 1979). So, we use the odds ratio to judge whether a haplotype is risk or not. Here, non-risk haplotypes are defined as haplotypes which are neutral or protective to the disease. The technical details are described as follows.

We calculate the odds ratio between h_k and $\mathbf{H}_a - \{h_k\}$ by

$$\mathrm{OR}_k = \frac{(n_{1k} + 0.5)(n_{0\bar{r}} + 0.5)}{(n_{0k} + 0.5)(n_{1\bar{r}} + 0.5)},$$

where adding 0.5 to the OR for the continuity correction. Then, the risk haplotype set $\mathbf{H}_r$ is calculated by

$$\mathbf{H}_r = \{h_k \in \mathbf{H}_a : \mathrm{OR}_k \geq \exp(c_1 \phi(n_{0k}, n_{1k}, n_{0\bar{r}}, n_{1\bar{r}}))\}, \tag{4.2}$$

where c_1 is a pre-specified constant and

$$\phi(n_{0k}, n_{1k}, n_{0\bar{r}}, n_{1\bar{r}}) = \sqrt{1/(n_{0k} + 0.5) + 1/(n_{1k} + 0.5) + 1/(n_{0\bar{r}} + 0.5) + 1/(n_{1\bar{r}} + 0.5)}.$$

4.2.2 Example

In applying the three components binomial mixture model to the previous example that we discussed briefly in Example 3.2.3, we found that the incomplete log likelihood taken a value of -392.224 at the true parameter values $\theta = (\pi_0, \pi_1, \pi_2, q_0, q_1, q_2) = (0.6980, 0.2370, 0.0650, 0.096, 0.23, 0.319)$ which we set in the simulation. However, when we applied the EM algorithm to the above example starting from different sets of initial values, we found out that the EM algorithm converged to different values. Table 4.1 shows the initial iteration and the final one resulted from applying the EM algorithm.

Tab. 4.1: The table shows the random initial values and the estimated ones of the final iteration when applying the EM algorithm to genotype data.

	Initial values						Final iteration						
Iter.	π_0	π_1	π_2	q_0	q_1	q_2	π_0	π_1	π_2	q_0	q_1	q_2	Inc $l(\theta)$
1	0.10105	0.69669	0.20226	0.43202	0.0484	0.70173	0.11658	0.76562	0.11779	0.27226	0.09459	0.89773	-358.76387
2	0.42563	0.50748	0.06689	0.42673	0.23196	0.10098	0.24941	0.54845	0.20214	0.49366	0.19009	0.07287	-437.84477
3	0.12049	0.3183	0.56121	0.89382	0.02339	0.11772	0.09249	0.2709	0.63661	0.9868	0.01428	0.15767	-361.48989
4	0.30887	0.05712	0.63401	0.91955	0.38633	0.15956	0.10226	0.04521	0.85252	0.97709	0.30794	0.11812	-360.85976
5	0.95638	0.02541	0.01821	0.59947	0.92822	0.51645	0.74782	0.02295	0.22923	0.15748	0.99659	0.13786	-416.1461
6	0.00043	0.78261	0.21695	0.3443	0.56379	0.12384	0.01466	0.36751	0.61784	0.29292	0.49687	0.11017	-446.93735
7	0.05152	0.13456	0.81392	0.75908	0.06098	0.73824	0.01825	0.68415	0.2976	0.7367	0.13292	0.59515	-477.64734
8	0.95643	0.0365	0.00707	0.49998	0.13264	0.10433	0.61282	0.30248	0.0847	0.32042	0.13339	0.06955	-474.33923
9	0.52399	0.30245	0.17355	0.3381	0.62554	0.50207	0.77443	0.13569	0.08987	0.13567	0.77644	0.2548	-397.65649
10	0.69936	0.03294	0.2677	0.12468	0.05317	0.84755	0.8265	0.06215	0.11135	0.14809	0.01116	0.95643	-369.26013
11	0.09258	0.16839	0.73903	0.56329	0.10294	0.78082	0.07161	0.69889	0.2295	0.30735	0.11685	0.74198	-398.54006
12	0.30666	0.34839	0.34494	0.30907	0.53513	0.85531	0.70726	0.18957	0.10317	0.13458	0.20015	0.93332	-387.73741
13	0.5981	0.03963	0.36227	0.0876	0.58177	0.79732	0.83928	0.02774	0.13298	0.11774	0.31716	0.90498	-372.95683
14	0.33234	0.4426	0.22506	0.50312	0.97698	0.54127	0.67259	0.08882	0.2386	0.13597	0.98438	0.13037	-412.80278
15	0.8683	0.06662	0.06508	0.3814	0.09615	0.84624	0.62512	0.32542	0.04946	0.23372	0.09276	0.98886	-366.4288
16	0.4205	0.23499	0.34452	0.11028	0.09444	0.12736	0.3988	0.22865	0.37254	0.1158	0.07779	0.22624	-468.20586
17	0.97859	0.0168	0.00461	0.5732	0.52413	0.93077	0.82498	0.16072	0.0143	0.14704	0.14153	0.99938	-422.67232
18	0.77082	0.12206	0.10711	0.59657	0.89334	0.26974	0.39596	0.0493	0.55474	0.23223	0.9788	0.13442	-384.29741
19	0.0154	0.01736	0.96724	0.32966	0.94419	0.70633	0.45034	0.01507	0.53459	0.13532	0.99714	0.26353	-402.36242
20	0.52574	0.32263	0.15163	0.80129	0.01057	0.66451	0.16505	0.70215	0.1328	0.80318	0.09466	0.27857	-374.0017

As it can be seen from Table 4.1 that the convergence is not really accurate in terms of the distance of the estimated parameter and the true ones underlying the model. The incomplete likelihood value is not a global maxima because it is less than the likelihood value at the true parameters. In fact there are two main reasons for that is we start from random initial values and there are many rare genotypes. We cope with these issues by proposing two ways of setting the initial values for EM algorithm.

4.2.3 EM algorithm initialization

Choosing initial values for the EM algorithm is an important step in finding a maximum of the likelihood. There are various ways to do that such as random initialization and data partition. See Karlis and Xekalaki (2003) for a review. Here, we consider the following two methods to initialize the EM algorithm.

Method 1 (random initialization): We randomly choose i_0 initial values (say $i_0 = 100$) of θ and run the EM algorithm with each chosen initial value. We take the best one among these runs in terms of maximizing the log-likelihood.

Method 2 (data partition): We first exclude the outlying frequencies in $\{\hat{p}_{1j}\}$, which have values of 0 or 1, to obtain robust means of a partition. Then, letting $c = (\max\{\hat{p}_{1j}\} - \min\{\hat{p}_{1j}\})/3$, we partition the frequencies into three sets as follows:

$$S_0 = \{\hat{p}_{1j};\ p_{1j} \leq \min\{\hat{p}_{1j}\} + c\},$$

$$S_1 = \{\hat{p}_{1j};\ \min\{\hat{p}_{1j}\} + c \leq \hat{p}_{1j} <\ \min\{\hat{p}_{1j}\} + 2c\},$$

and

$$S_2 = \{\hat{p}_{1j};\ \hat{p}_{1j} > \min\{\hat{p}_{1j}\} + 2c\}.$$

Note that the prospective frequency is increasing in the number of risk haplotypes which it carries. So, we expect that S_2, S_1 and S_0 mainly contain the frequencies corresponding the sets of genotypes with two risk haplotypes, with one risk haplotype, and with no risk haplotypes respectively. We choose the following initial values for estimating q_ν and π_ν, $\nu = 0, 1, 2$:

$$q_\nu^0 = \frac{\sum_{p_{1j} \in S_\nu} p_{1j}}{|S_\nu|},\ \text{and}\ \pi_\nu^0 = \frac{|S_\nu|}{m},$$

where $|S_\nu|$ denotes the cardinality of S_ν, $1 \leq j \leq m$ and m is equal to J minus the

excluded outliers in $\{\hat{p}_{1j}\}$.

4.2.4 Multiple testing method

To compare the proposed method to the multiple testing procedure of Zhu et al. (2010), we briefly describe their procedure as follows. In their procedure, a subsample A containing $N_0^{(a)}$ and $N_1^{(a)}$ individuals are randomly chosen from the controls and cases respectively. These individuals are used in the screening stage and the remaining forms a validation subsample B to be used in the validation stage. Suppose that there are K different haplotypes inferred from A by using the PHASE. Let $(r_{0k}^{(a)}, r_{1k}^{(a)})$, $1 \leq k \leq K$ be their respective frequencies in controls and cases respectively.

Screening stage: We perform a respective frequencies-based screening by calculating an estimated risk haplotype set as follows:

$$S^{(a)} = \{h_k : z_k^{(a)} > c_0, 1 \leq k \leq K\},$$

where c_0 is a pre-specified constant ($c_0 = 1$ in our later simulations) and

$$z_k^{(a)} = \frac{r_{1k}^{(a)} - r_{0k}^{(a)}}{\sqrt{r_{0k}^{(a)}(1 - r_{0k}^{(a)})/(2N_1^{(a)})}}.$$

Validation stage: The $S^{(a)}$ is refined by performing Fisher's exact test based on subsample B for each haplotype in $S^{(a)}$. This gives a final risk haplotype set denoted by $S^{(b)}$.

4.3 Simulation studies

In this section, via simulations we will examine the performance of the proposed methods in terms of the estimated L_1 bias and the average of sensitivity and specificity under various scenarios. Let $\hat{\theta}$ be the estimator of θ, and $\mathbf{H}_r$ and $\mathbf{H}_{\bar{r}}$ the estimated true risk and non-risk haplotype sets respectively. Let $\mathbf{T}_r$ and $\mathbf{T}_{\bar{r}}$ be the true risk and non-risk haplotype sets. Then, by the L_1 bias we mean the L_1 distance between $\hat{\theta}$ and θ. By the sensitivity and specificity of $\mathbf{H}_r$ and $\mathbf{H}_{\bar{r}}$, we mean

the positive discovery rate and the negative discovery rate:

$$\text{sen} = \frac{|\mathbf{H}_r \cap \mathbf{T}_r|}{|\mathbf{T}_r|} \text{ and spe} = \frac{|\mathbf{H}_{\bar{r}} \cap \mathbf{T}_{\bar{r}}|}{|\mathbf{T}_{\bar{r}}|}.$$

We take the average AVSS = (sen + spe)/2 to assess the performance of a haplotype classification procedure.

4.3.1 Performance of the proposed data partition-based initialization

To compare the proposed data partition-based initialization (Method 2) to the random initialization (Method 1), we generated 30 genotype datasets on 10 single nucleotide polymorphisms (SNPs), each dataset, containing N_0 controls and N_1 cases, was obtained by the following two steps: In the step 1, we used the software MS (Hudson, 2002) to simulate $2(N_0 + N_1)$ haplotypes with a mutation rate of 2. We randomly chose m_r of these haplotypes and labeled them as risk haplotypes. To save the space, we considered only $N_0 + N_1 = 5000$ and $m_r = 10$. The results for other values of $N_0 + N_1$ and m_r were similar. In the step 2, the disease states of the above genotypes were simulated from the multiplicative inheritance model with $q_0 = 0.1$ and $\lambda = 3$. Note that the number of genotypes depends on the mutation rate and was varying across 30 datasets.

The comparison was based on the log-likelihood, the run time, estimated bias and classification error rate (CER). The estimated bias can be calculated by sum all the absolute values of the differences between $\hat{\theta}$ and the true θ. Note that genotypes in each dataset could be divided into three (true) groups, say $\mathbf{G}_\nu$, $\nu = 0, 1, 2$ as we knew the number of risk haplotypes which each genotype contained in the simulation. On the other hand, if we pretended that we did not know which haploypes were risk (therefore, we did not know the group memberships of these genotypes). To infer their memberships, we fitted a three-component binomial mixture model to each of 30 datasets. By using the estimated posterior probabilities, $\tau_{\nu j}, \nu = 0, 1, 2$, of group memberships derived from the EM algorithm, we assigned the j-th genotype to the group $\hat{\mathbf{G}}_\nu$, $\nu = 0, 1, 2$ if $\tau_{\nu j} = \max_t \tau_{tj}$. Here, we labeled three estimated groups according to the ordered penetrances $\hat{q}_0 \leq \hat{q}_1 \leq \hat{q}_2$. The accuracy of three estimated groups was evaluated by the CER defined as

$$\text{CER} = \sum_\nu \left(1 - \frac{|\mathbf{G}_\nu \cap \hat{\mathbf{G}}_\nu|}{|\mathbf{G}_\nu|}\right),$$

where we counted the total number of misclassified genotypes divided by the total

number of the genotypes. The results were summarized in Figure 4.1 in terms of the box-whisker plots of the estimated biases, the CERs, likelihood values, and time-costs over 30 datasets for Methods 1 and 2 respectively. Methods 1 and 2 denote the random initialization and the data partition-based methods. In this figure, the panels show the box-whisker plots of the estimated biases in estimating θ, the CERs, the attained log-likelihoods, and the time-costs for Methods 1 and 2 respectively. The result shows that Method 2 substantially outperformed Method 1. Therefore, we decided to initialize the EM algorithm by use of Method 2 in the remaining simulations as well as the real data analysis below.

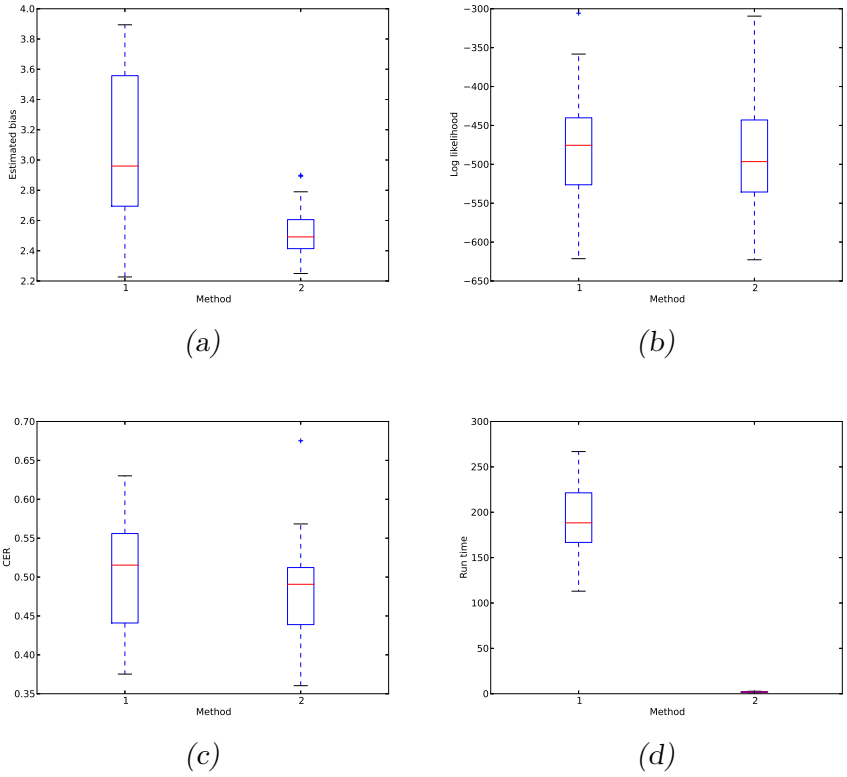


Fig. 4.1: Performance of two initialization methods.

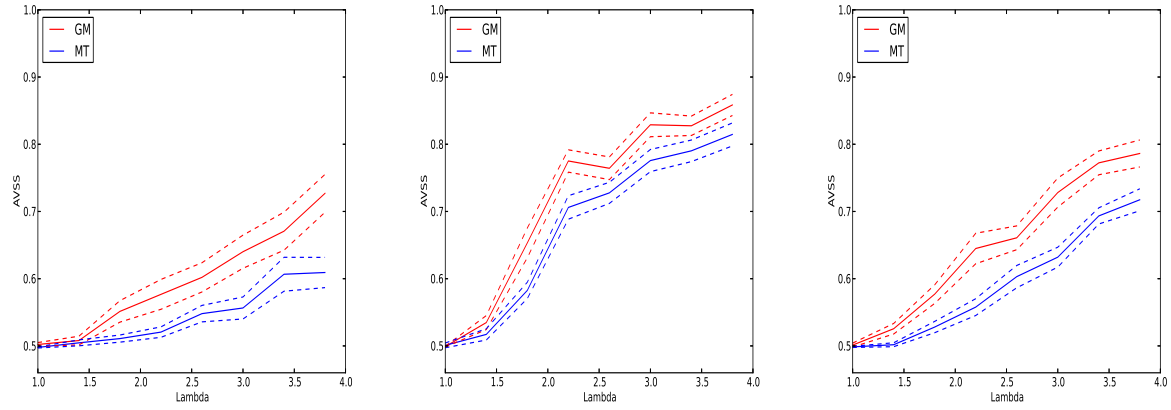
4.3.2 Performance of the proposed two-stage method

Note that the proposed two-stage method is based on the prospective likelihood model although real data were obtained from retrospective studies. By the simulations below, we addressed whether the proposed method could outperform Zhu

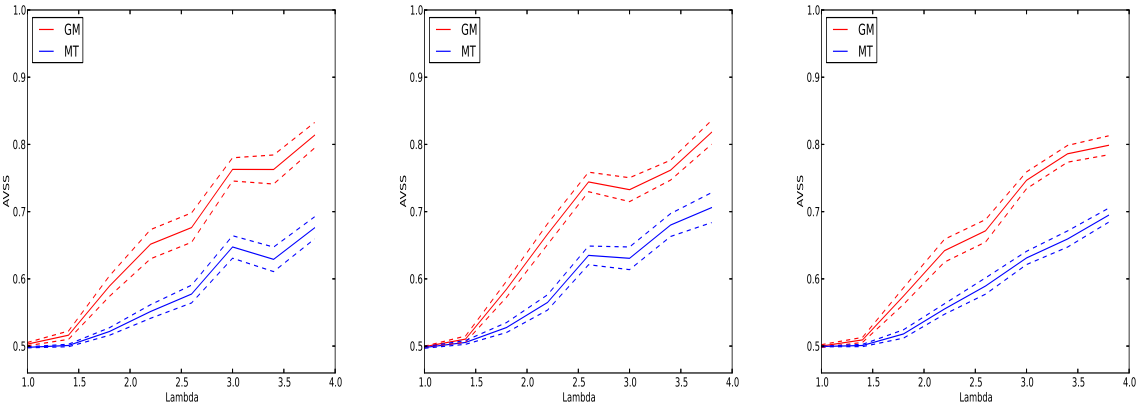
et al's multiple-testing procedure in both prospective (i.e., cohort) and retrospective (i.e., case-control) studies. We used the same ways that we described in the section 3.3.2 to generate data according to the cohort design and case-control design.

Setting 1 (cohort design): In this design, we considered various combinations of $(N_0 + N_1, m_r, \text{IM}, f_0, \lambda)$, where $N_0 + N_1 = 3000, 5000$, $m_r = 5, 10, 20$, IM $= 1, 2, 3$, $f_0 = 0.1$, $\lambda = 1, 1.4, 1.8, 2.2, 2.6,\ 3, 3.4,$ and 3.8 respectively.

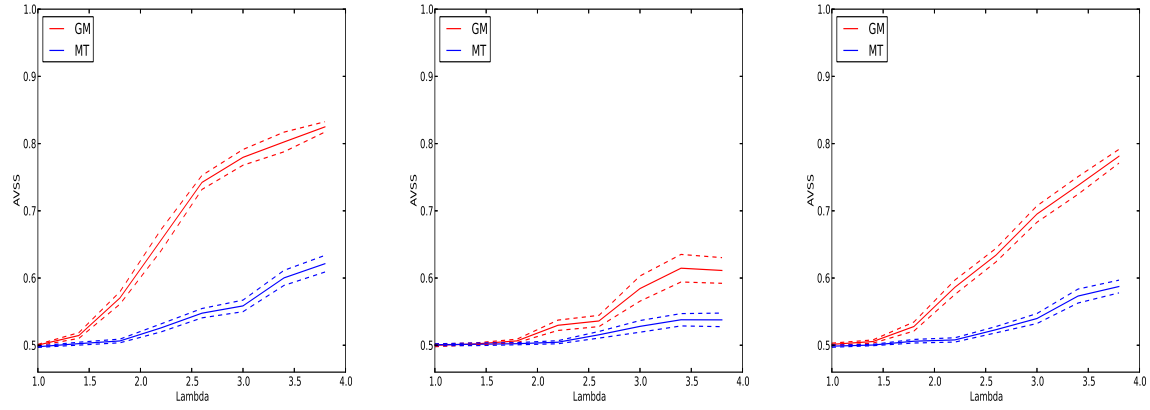
For each scenario, we applied both the proposed method and the multiple testing method to 30 datasets and calculated their AVSS values respectively. For each of the three inheritance modes, we plotted the means of these AVSS values over 30 datasets against λ. In the plots displayed in Figures 4.2, the red and the blue solid curves show means of the AVSS values (i.e., the values of (specificity and sensitivity)/2)) over 30 datasets are plotted against the values of λ for the proposed method and the multiple testing method respectively. The two red dash curves are one standard error up and down from the red mean curves. Similarly, the two blue dash curves are one standard error up and down for blue mean curves. The plots in the rows from the top to the bottom are for the cases where there were 5, 10, and 20 risk haplotypes in the underlying haplotypes and based on different modes of inheritance. The result represents the sample size of 5000 for cases and controls altogether. Similarly, in Figure 4.4, the plots showing the results of the same above scenarios but for sample size of 3000 for cases and controls altogether. The figure shows that on the cohort data, the proposed two stage method performed substantially better than the multiple testing method in all the scenarios defined above. The improvement was achieved by using model-based genotype clustering. This is not surprising, because Yeung et al. (2001) has already showed that the model-based clustering is often superior over non-model based clustering.



(a) N=5000, m_r=5, rece (b) N=5000,m_r=5,dom (c) N=5000,m_r=5,mult

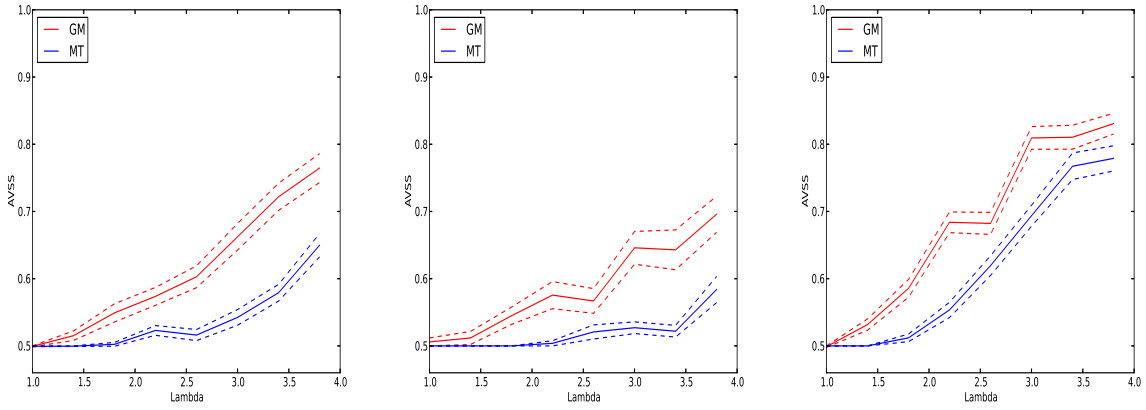


(d) N=5000,m_r=10,rece (e) N=5000,m_r=10,dom (f) N=5000,m_r=10,mult

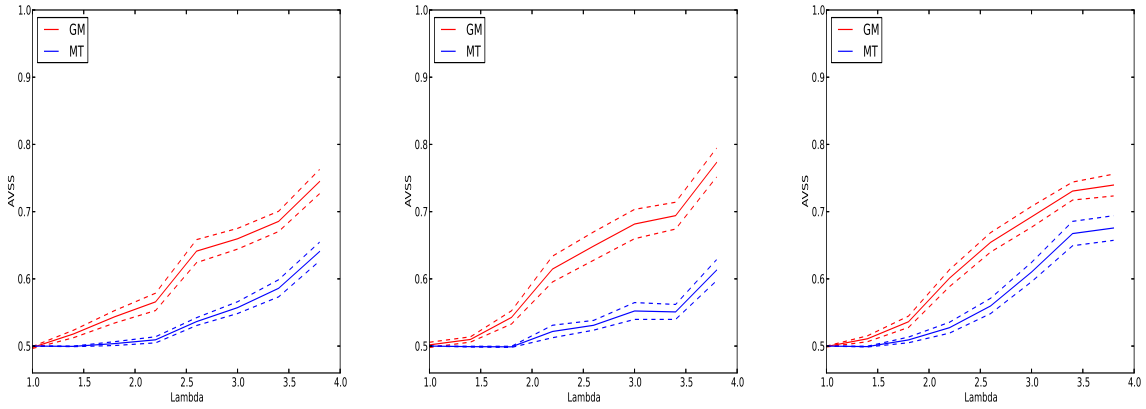


(g) N=5000,m_r=20,rece (h) N=5000,m_r=20,dom (i) N=5000,m_r=20,mult

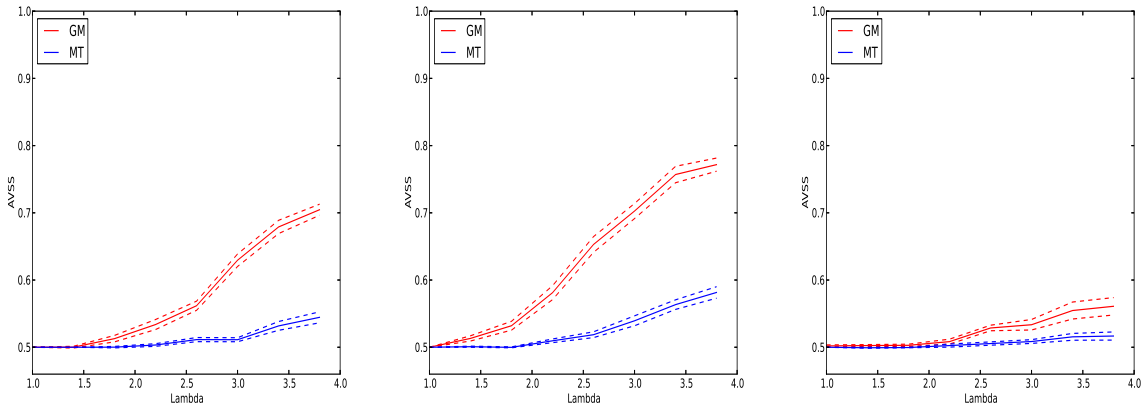
Fig. 4.2: Performances of the proposed two-stage method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance modes with sample size $N = 5000$.



(a) N=3000,m_r=5,mult (b) N=3000,m_r=5,rece (c) N=3000,m_r=5,dom



(d) N=3000,m_r=10,mult (e) N=3000,m_r=10,rece (f) N=3000,m_r=10,dom

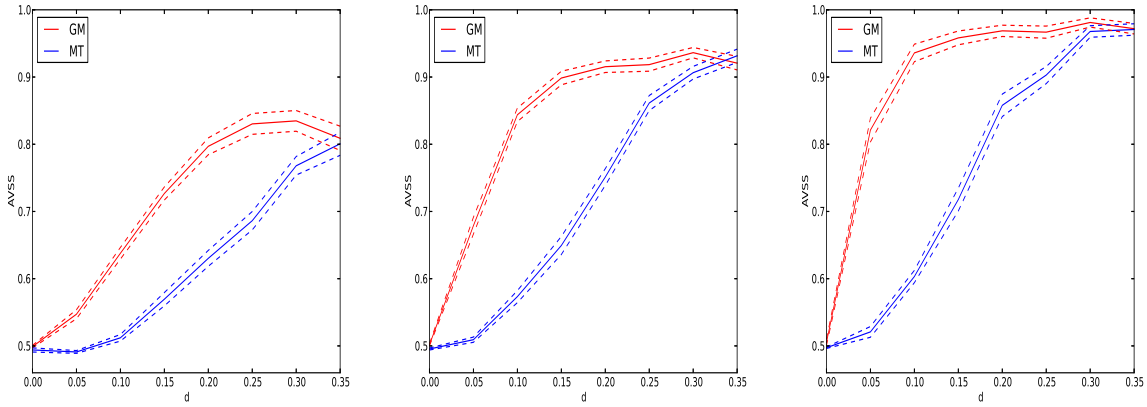


(g) N=3000,m_r=20,mult (h) N=3000,m_r=20,rece (i) N=3000,m_r=20,dom

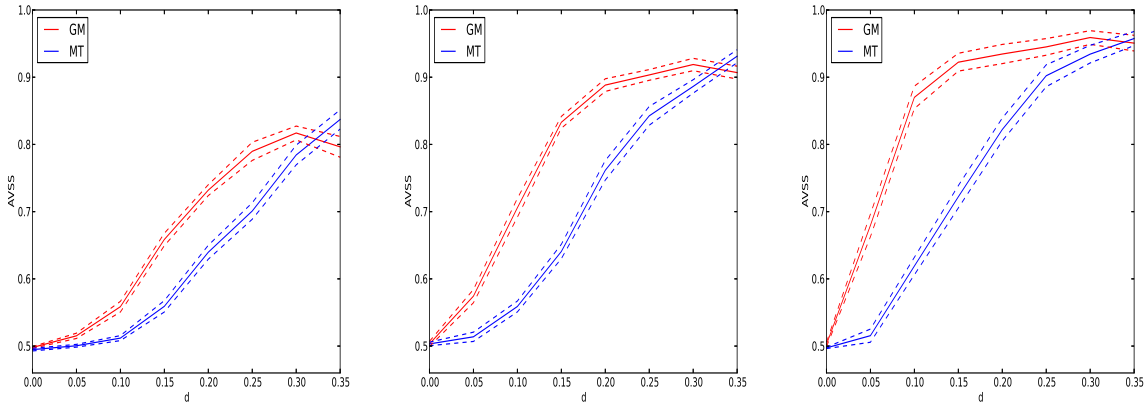
Fig. 4.3: Performances of the proposed two-stage method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models with sample size $N = 3000$..

Setting 2 (case-control design):

We applied the proposed two-stage method and the multiple testing method to these case-control data. The plots in Figure 4.4, the mean curves of the AVSS values with one standard error up and down were plotted against the d values $\{0, 0.05, 0.1, 0.1,\ 0.15, 0.2, 0.25, 0.3, 0.35\}$. The columns from the left to the right are for the scenarios, where the underlying number of risk haplotypes $m_r = 5, 10,$ and 20. The top row stands for the cases, where $(N_0, N_1) = (2000, 3000)$, while the bottom row stands for the cases, where $(N_0, N_1) = (1000, 2000)$. The results again demonstrate that the proposed two-stage method can be more powerful than the multiple testing method in detecting risk haplotypes. However, the AVSS gain was decreasing in the number of risk haplotypes m_r as well as the underlying odds ratios of Groups 1 and 2. In particular, the AVSS gain can be negative when there were many risk-haplotypes presented in the data. This is due to the effect of unbalanced case and control sample sizes in the finite sample size setting, because our model in Stage 1 is a prospective model.



(a) N1=2000, N0=3000,m_r=20 (b) N1=2000, N0=3000,m_r=10 (c) N1=2000, N0=3000,m_r=5



(d) N1=1000, N0=2000,m_r=20 (e) N1=1000, N0=2000,m_r=10 (f) N1=1000, N0=2000,m_r=5

Fig. 4.4: Performances of the proposed two-stage method and the multiple testing method on the case-control data.

4.4 Real data analysis

We applied the proposed permutation method to the GWAS genotype datasets on coronary artery disease (CAD) and hypertension (HT) obtained by Affymetrix 500K SNP chips in the WTCCC study (WTCCC, 2007). The datasets are prepared in the same way we described in 3.5. To reduce the dimension of the genotypes, we segmented the genome into regions of 8 SNPs according to their positions on the chromosomes, obtaining 61218 regions and the corresponding genotype datasets $\mathbf{G}_k, k = 1, 2, ..., 61218$. Note that the long region will dilute the effects of risk SNPs and can result in many rare genotypes, whereas the short region will miss interactions between SNPs. The region length of 8 was chosen to achieve a compromise between

the above aspects by using a pilot study. Also note that as we excluded the SNPs with bad callings, the numbers of cases and controls are varying across the different regions.

Note that $\{\mathbf{G}_k : k = 1, ..., 61218\}$ contained 1983537 genotypes in total for the CAD data and 2097111 genotypes in total for the HT data respectively. The proposed procedure includes two stages. In Stage 1, we obtained the estimated risk genotypes, while in Stage 2, we further inferred haplotype pairs from the estimated risk genotypes. In Stage 1, we first fitted a three-component binomial mixture model to each $\mathbf{G}_k$ and then thresholded the genotypes based on the smallest penetrance in the three components. The thresholding would involve 1983537 tests for the CAD data and 2097111 tests for the HT data. So in equation (4.1), we set $\varepsilon = 0.05/1983537 = 2.52 \times 10^{-8}$ for the CAD data and $\varepsilon = 0.05/2097111 = 2.38 \times 10^{-8}$ for the HT data. In Stage 2, we employed the PHASE to infer the haplotypes from the risk genotypes derived from the previous stage. This gave rise to 201528 potential risk haplotypes out of 1448586 in CAD data and 213578 potential risk haplotypes out of 1463838 in HT data. We further conducted the OR thresholding for these haplotypes. There would involve 201528 tests in the CAD case and 213578 tests in the HT case. By using the Bonferroni adjustment, we set the corresponding individual test level at $0.05/201528 = 2.48 \times 10^{-7}$ and $0.05/213578 = 2.34 \times 10^{-7}$ for the CAD and the HT respectively. These individual test levels were then used to determine the tuning constant c_1 in equation (4.2). This yielded $c_1 \approx 5$. After performing the proposed two-stage method on the datasets, we obtained the estimated risk and non-risk haplotype sets, $\hat{\mathbf{H}}_r$ and $\hat{\mathbf{H}}_{\bar{r}}$, for the CAD and the HT respectively.

Finally, we carried out a genome control on the above results by taking advantage of the fact that there were two sub-populations in controls. The genome control can eliminate these false haplotypes generated by the PHASE and population substructures from the selected list of risk haplotypes. In the genome control, we run the chi-square tests on the association of two control sub-populations with each estimated risk haplotype. We eliminated these estimated risk haplotypes with p-values for the above chi-square tests less than $< 30\%$. Here, 30% was chosen by the simulations which are shown in Figure 3.6, aiming to filter out false risk haplotypes.

The genome control gave the final risk-haplotype set as displayed in Tables 4.2, 4.3, 4.4, and 4.5 below. In the tables, each haplotype has been assigned to a physically closest gene on the basis of the information provided the GWAS catalog and the genetic information from the British 1958 Birth cohort. See Welter et al. (2014) and the web page at

http://www2.le.ac.uk/projects/birthcohort/1958bc. In the CAD case, we did rediscover the CAD risk genes TNIK in chromosome 3, CDKN2B in chromosome 9, BTG1 in chromosome 12, and A2BP1 in chromosome 16, which were found by the previous study (WTCCC, 2007; Zhu et al., 2010; Welter et al., 2014). In the HT case, we also identified a number of variants which were potentially associated with hypertension. However, we were not able to confirm any existing discoveries in the literature (Welter et al., 2014). A possible reason is that we set a very stringent level for the odds ratio thresholding based on the Bonferroni adjustment for multiple testing. It is well-known that the Bonferroni adjustment is very conservative.

Tab. 4.2: The predicted risk haplotypes for CAD by use of the WTCCC data. In the table, the P-values were derived from the chi-squared test of the frequencies of H_i against the collapsed frequencies of the estimated non-risk haplotypes.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
1	17921479 – 17955334	$rs11203219 - rs638425$	$AATGCCGC$	0.04602	0.01388	3.05038	4.1×10^{-12}	ACTL8
1	75974016 – 76018681	$rs3806162 - rs5745391$	$TCTATCAA$	0.05105	0.01954	3.18049	1.2×10^{-12}	MSH4
2	49934439 – 50000082	$rs6736617 - rs17039375$	$CCAAAGGT$	0.02347	0.00757	3.08898	6.6×10^{-10}	NRXN1
2	81387425 – 81525659	$rs4401229 - rs2862499$	$TTGCTCCA$	0.0451	0.02468	2.54951	1.8×10^{-12}	LOC442021
2	222486954 – 222527591	$rs16863087 - rs2392937$	$CCAAACGG$	0.04059	0.02497	2.09348	4.3×10^{-08}	LOC402120
2	230201571 – 230228527	$rs6755403 - rs13391903$	$AGTTTGCC$	0.1132	0.04164	2.78377	2.3×10^{-08}	DNER
2	239420300 – 239491966	$rs4545955 - rs13008279$	$TTCCAGGA$	0.05558	0.02584	2.17494	1.3×10^{-12}	FLJ43879
2	241821720 – 241873661	$rs4675991 - rs935262$	$CGGGGTTT$	0.03735	0.01659	2.32538	1.4×10^{-10}	PPP1R7
3	4927181 – 5001898	$rs17041733 - rs11925620$	$CCTCCTCC$	0.04287	0.01795	2.16999	1.2×10^{-07}	BHLHB2
3	14422977 – 14471151	$rs4684216 - rs9834629$	$GATGATGC$	0.01815	0.00509	3.63785	1.7×10^{-09}	SLC6A6
3	60586653 – 60641652	$rs7432576 - rs1716739$	$CTATAAGC$	0.15989	0.11374	1.55681	9.4×10^{-12}	FHIT
3	63365648 – 63390235	$rs17068494 - rs1403700$	$TCCTTCGG$	0.08979	0.04741	2.04072	7.1×10^{-09}	SYNPR
3	67509601 – 67525645	$rs9867659 - rs17046411$	$ACGATGTT$	0.05192	0.03019	1.95683	5.1×10^{-09}	SUCLG2
3	103285842 – 103325614	$rs7623627 - rs9844712$	$GTCCCTAT$	0.02744	0.00999	3.15138	1.6×10^{-09}	NFKBIZ
3	106353367 – 106411138	$rs16850901 - rs9846852$	$TATCGAGA$	0.02931	0.0065	4.87306	7.5×10^{-18}	ALCAM
3	144925558 – 144993828	$rs4330252 - rs12233446$	$TGGGATAC$	0.02976	0.00733	5.71824	1.8×10^{-16}	SLC9A9
3	145364476 – 145471873	$rs9854202 - rs3925560$	$AACGGACT$	0.37409	0.29638	2.25725	5.5×10^{-34}	C3orf58
3	172422863 – 172457251	$rs954749 - rs16856054$	$TTCTTACT$	0.12948	0.08707	1.50219	2.2×10^{-08}	TNIK
3	192463499 – 192526004	$rs7644510 - rs293871$	$GACGCGTA$	0.04375	0.01075	3.69505	1.3×10^{-18}	UTS2D
3	197256495 – 197339533	$rs6583286 - rs9834962$	$TAGACTTA$	0.0498	0.02364	2.27577	2.7×10^{-10}	TFRC
4	3636361 – 3700212	$rs10025237 - rs16844722$	$GGGGAGGG$	0.22491	0.15492	1.65607	1.9×10^{-07}	FLJ35424
5	120487082 – 120547238	$rs11956204 - rs17514347$	$ATTGGGAG$	0.02739	0.00735	3.8359	1.5×10^{-13}	LOC728682
5	166764561 – 166801933	$rs6863935 - rs7724862$	$CTATGTGT$	0.09145	0.05448	1.69398	8.7×10^{-09}	ODZ2
7	4779368 – 4930112	$rs2942566 - rs4320451$	$CGGGTCAT$	0.10433	0.06243	1.66428	5.5×10^{-10}	RBAK
7	10052046 – 10079446	$rs10225194 - rs11768931$	$GGTTCGCT$	0.04951	0.0245	2.64149	9.4×10^{-15}	LOC340268
7	34178282 – 34260002	$rs17169771 - rs16878925$	$AGGTTGCG$	0.05229	0.02631	2.71386	3.3×10^{-13}	AAA1
7	42931717 – 42940671	$rs2024125 - rs2330742$	$AGTGTAGA$	0.09745	0.0513	1.90132	2.0×10^{-10}	HECW1
7	153564509 – 153621369	$rs869490 - rs6953905$	$TCGTATCG$	0.0667	0.03524	1.93779	6.6×10^{-11}	LOC653748
8	5482876 – 5498858	$rs2189889 - rs4875607$	$CGGACCGA$	0.07873	0.0533	1.64615	2.4×10^{-08}	LOC648237
8	17486464 – 17509327	$rs2705093 - rs2588121$	$CCTGCGAG$	0.05925	0.02338	2.67404	1.6×10^{-15}	PDGFRL
8	38345434 – 38449100	$rs16887343 - rs12677355$	$ACGTACCT$	0.09472	0.05661	1.82381	7.0×10^{-13}	WHSC1L1
8	104190450 – 104202402	$rs2515173 - rs3019159$	$GGCCATCT$	0.14195	0.08768	1.62006	1.5×10^{-08}	BAALC

Tab. 4.3: The continuation of Table 4.2.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
9	22088619 – 22120515	$rs2891168 - rs10965245$	$GGTGCCAG$	0.34939	0.29298	1.52609	1.0×10^{-07}	CDKN2B
9	74180343 – 74241329	$rs10114124 - rs17081046$	$GTATTTAT$	0.21608	0.13046	1.61055	1.2×10^{-07}	RORB
9	114777214 – 114805868	$rs1322060 - rs10121268$	$GAGCCTAA$	0.09498	0.06007	1.56664	2.3×10^{-08}	TNFSF8
9	119506057 – 119537035	$rs2191675 - rs10984648$	$GTTGGCTA$	0.08762	0.03361	2.41642	3.0×10^{-16}	CDK5RAP2
10	11879196 – 11924252	$rs6602535 - rs11257355$	$TCTGCCGG$	0.1694	0.12811	1.41273	6.4×10^{-08}	C10orf47
10	11879196 – 11924252	$rs6602535 - rs11257355$	$TCTGCCGG$	0.1694	0.12811	1.41273	6.4×10^{-08}	C10orf47
10	64409674 – 64442476	$rs1509952 - rs2842286$	$TTTCTTAC$	0.02299	0.0073	4.03039	1.6×10^{-09}	NRBF2
11	8165969 – 8200374	$rs4758310 - rs11041816$	$ATAATGGG$	0.36298	0.3164	1.3306	1.1×10^{-08}	LOC644497
11	21323965 – 21363331	$rs17233214 - rs1945444$	$GGTAACAT$	0.08147	0.04232	1.98043	8.6×10^{-12}	NELL1
11	69213458 – 69295251	$rs1192923 - rs3168175$	$TCGTGGCA$ $TTGTGGCA$	0.10225 0.05213	0.05587 0.02803	1.98038 2.01202	8.9×10^{-14} 5.6×10^{-09}	FGF4
11	83230307 – 83256927	$rs1878266 - rs1878264$	$TATATTCA$	0.03571	0.01807	2.11905	2.5×10^{-07}	CCDC90B
12	90721177 – 90758721	$rs10745571 - rs17193868$	$GGGCTATA$	0.0351	0.00949	3.88035	1.7×10^{-16}	BTG1
12	114038450 – 114074493	$rs1828384 - rs35346$	$TGTACCCT$	0.03245	0.01341	2.52817	2.3×10^{-07}	TBX3
12	127146384 – 127182360	$rs10847535 - rs10773498$	$TTGTCGCG$	0.10562	0.07049	1.50842	1.3×10^{-07}	TMEM132C
12	129086441 – 129129809	$rs713149 - rs1027557$	$AAAGCGGT$	0.18839	0.11206	1.74867	4.4×10^{-14}	FLJ31485
13	26845975 – 26875430	$rs11616513 - rs17085553$	$TACGCACA$	0.04431	0.02025	2.30656	7.1×10^{-10}	MTIF3
13	31414174 – 31438047	$rs17076954 - rs169410$	$CCTCCCGT$	0.30306	0.29469	2.6188	6.9×10^{-08}	LOC196549
13	48154476 – 48209065	$rs7330127 - rs9562843$	$ACGATAGA$	0.02762	0.0048	5.63922	2.7×10^{-10}	RCBTB2
14	25140850 – 25159405	$rs8020556 - rs1951062$	$AGTACATA$ $AGTAAACT$ $GCTACATA$	0.24934 0.09084 0.04608	0.2259 0.02999 0.01682	1.41488 3.87615 3.50368	3.5×10^{-08} 1.0×10^{-41} 3.4×10^{-22}	LOC401767
14	32591680 – 32606647	$rs12883961 - rs10140504$	$CATGGGAG$	0.03736	0.01879	2.21665	1.1×10^{-08}	NPAS3
14	65343491 – 65401760	$rs3924222 - rs12896836$	$TATAACTC$	0.0462	0.01904	2.55404	5.2×10^{-14}	FUT8
15	20592297 – 20610835	$rs4778334 - rs1991922$	$TAGCCCAT$	0.04494	0.01488	2.75061	1.1×10^{-12}	NIPA1
15	20624103 – 21246055	$rs7166056 - rs8024346$	$GTGACGTG$	0.08093	0.04109	2.10848	2.4×10^{-13}	NIPA1
15	21610088 – 21670901	$rs824163 - rs7181211$	$TTTTCAAC$	0.22034	0.15435	1.43864	4.9×10^{-09}	MAGEL2
15	37962389 – 38014169	$rs11633436 - rs534757$	$TTACAACC$	0.07798	0.03763	1.99235	2.7×10^{-11}	GPR176
15	64637416 – 64669062	$rs1030986 - rs4776800$	$CACGTCGT$	0.04575	0.01594	2.65924	2.2×10^{-09}	LCTL
15	79193543 – 79223619	$rs1317059 - rs6495541$	$CTCGGACC$	0.02813	0.00459	6.34974	2.2×10^{-15}	C15orf26
15	90365510 – 90400043	$rs12906289 - rs992838$	$ACGTAAGG$	0.07777	0.02342	3.50153	1.1×10^{-26}	SLCO3A1
15	91435452 – 91473401	$rs4778099 - rs17526830$	$GATCCCTA$	0.07536	0.04084	1.94917	1.7×10^{-09}	RGMA

Tab. 4.4: The continuation of Table 4.3.

16	6155489 – 6181184	$rs11642397 - rs1946127$	$TTGGGTTG$	0.02433	0.00883	2.92587	1.7×10^{-07}	A2BP1
16	46937666 – 47050362	$rs11076564 - rs8054696$	$AACGGGCC$ $TGAAGGCT$	0.18717 0.04224	0.15302 0.02781	1.62027 2.01195	1.1×10^{-07} 2.3×10^{-07}	LONP2
16	51239337 – 51264345	$rs3112587 - rs4386133$	$CCTATGAG$	0.07702	0.0442	1.68656	7.3×10^{-08}	LOC643714
16	55207138 – 55253047	$rs8055724 - rs12447986$	$TTCTCCTC$	0.03044	0.01113	2.65805	9.0×10^{-09}	MT1L
17	73602775 – 73670122	$rs16970811 - rs9909570$	$CCCACTAG$	0.02022	0.00446	4.82821	3.1×10^{-13}	TNRC6C
17	74629176 – 74682195	$rs2612793 - rs8072667$	$CGAGGTTG$	0.06276	0.03471	1.95026	6.7×10^{-09}	FLJ21865
18	8212591 – 8279839	$rs10468776 - rs11876033$	$GGGACAAG$	0.02689	0.00982	2.86846	1.7×10^{-10}	PTPRM
18	8772147 – 8782163	$rs12606001 - rs8084401$	$TCAGTGAC$	0.09539	0.03649	2.66938	1.3×10^{-17}	KIAA0802
18	60647495 – 60688045	$rs1595904 - rs17678507$	$CAGCGTGC$	0.08119	0.04205	2.1482	6.5×10^{-16}	C18orf20
19	50064169 – 50153836	$rs17561351 - rs204907$	$AGGCAGAA$	0.05937	0.02583	2.35486	5.1×10^{-14}	PVRL2
19	52946204 – 53026777	$rs10402957 - rs4427918$	$CATTCAGC$	0.0741	0.04321	1.87681	1.7×10^{-11}	GLTSCR2
19	59113663 – 59296006	$rs7257613 - rs3760698$	$CCGGCCGC$ $CCGGCCAC$	0.06977 0.12473	0.0159 0.08441	5.01246 1.69429	2.7×10^{-43} 6.7×10^{-13}	CACNG7
20	5265473 – 5327486	$rs6085111 - rs6085143$	$ACCAATCC$	0.04815	0.02744	1.83971	1.3×10^{-07}	FLJ33544
20	42465269 – 42498442	$rs3181206 - rs6017342$	$GGCTTCCA$	0.12685	0.06245	2.08814	3.0×10^{-14}	HNF4A
20	44639977 – 44681497	$rs376438 - rs847096$	$AAGTCTGC$	0.09805	0.04784	1.90457	8.8×10^{-12}	SLC13A3
20	49937544 – 50006641	$rs6067996 - rs6021570$	$ATTGGACA$	0.03133	0.01165	2.82133	2.6×10^{-11}	SALL4
20	51762764 – 51798874	$rs4811452 - rs4811457$	$GATGTTCA$	0.05611	0.03099	1.87441	1.7×10^{-08}	ZNF217
20	57707915 – 57741702	$rs12481511 - rs16984986$	$TGTACCAG$	0.0773	0.0427	1.95199	1.2×10^{-07}	PHACTR3
21	2015127 – 13517135	$rs2847443 - SNP_A$	$TACAAGAT$	0.10999	0.09446	1.65501	2.4×10^{-08}	TPTE
22	16871076 – 16895136	$rs8142200 - rs975826$	$TCGGGAGG$	0.03219	0.00253	10.88401	1.8×10^{-19}	LOC729269
22	31354524 – 31372260	$rs8139704 - rs5749480$	$CGCTAGGG$	0.02584	0.00524	5.07641	3.4×10^{-16}	SYN3
22	35324014 – 35335429	$rs7410412 - rs12160203$	$TTTCAAGG$	0.17403	0.10746	1.67423	1.3×10^{-10}	CACNG2

Tab. 4.5: The predicted risk haplotypes of hypertension by use of WTCCC data. In the table, the P-values were derived from the chi-squared test of the frequencies of H_i against the collapsed frequencies of the estimated non-risk haplotypes.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
1	236986859 – 237020204	$rs12137158 - rs16840310$	$ATTTAGGG$	0.08733	0.05437	1.69625	3.4×10^{-10}	GREM2
4	3700382 – 3734797	$rs177772 - rs12641338$	$TACCGATT$	0.12978	0.08988	1.59997	7.7×10^{-12}	FLJ35424
4	170032303 – 170061525	$rs6822949 - rs17614553$	$GAACGGAA$	0.0425	0.01579	2.86663	4.8×10^{-10}	PALLD
6	152700181 – 152736079	$rs7747166 - rs7776399$	$CGGCTCCC$	0.52639	0.49931	3.36065	2.7×10^{-23}	SYNE1
			$CGGGTCCT$	0.04238	0.03768	3.58962	5.7×10^{-14}	
11	69213458 – 69295251	$rs1192923 - rs3168175$	$TTGTGGCA$	0.05532	0.02803	2.12665	3.4×10^{-10}	FGF4
12	116500495 – 116514298	$rs10850852 - rs1400593$	$CTCTCTTC$	0.28748	0.26232	2.46528	5.2×10^{-17}	NOS1
14	21674996 – 21704333	$rs12050442 - rs1894369$	$GGGGTTAC$	0.03075	0.00968	3.28277	1.8×10^{-11}	TRA@
14	25140850 – 25159405	$rs8020556 - rs1951062$	$AGTAAACT$	0.08475	0.02999	2.94949	6.6×10^{-27}	LOC401767
14	36411583 – 36421982	$rs10872897 - rs2564848$	$TACCTCCC$	0.02712	0.01101	2.63669	1.4×10^{-08}	SLC25A21
			$ATCCACTT$	0.02299	0.00637	3.84732	1.3×10^{-11}	
14	36969639 – 37032855	$rs10132119 - rs17106785$	$CTATGACA$	0.01914	0.00402	5.57575	6.1×10^{-10}	MIPOL1
19	17595848 – 17649789	$rs10419511 - rs7252308$	$TTGGTATG$	0.04536	0.01971	2.16516	1.7×10^{-10}	UNC13A

4.5 Discussion and conclusion

We are currently at an era of extraordinary growth in the data describing human genetic variation and its correlation with complex traits. The recent development of bio-technologies allows an international consortium of geneticists to revolutionize genetic research through large scale genome wide association studies (GWAS). Although these studies have identified hundreds of loci at very stringent levels of statistical significance across many different human traits, these loci are only able to explain a small fraction of the population risk. To address the issue, new models and new hypotheses have been proposed, which pose challenges to conventional statistics underlying much of our genetic analysis. For example, GWAS analyses are most commonly performed by testing the association of individual variants with the disease, ignoring the potential interactions between the variants. It is believed that the region or gene-based analysis is more powerful in capturing the collective activity of sets of variants by testing the association of the group instead of each component individually with the disease.

In this approach, we have adopted the region-based strategy that segments the genome into 61218 regions with around 8 SNPs each. For each region, a list of distinct genotypes with their frequencies in cases and controls have been worked out. The problem facing us is of the sparse distribution of these genotypes. To circumvent it, people often first infer haplotypes from the genotypes and then cluster the haplotypes into a number of groups. The association analysis is conducted on the basis of the inferred groups, for example, by using multiple Z-tests (Zhu et al., 2010). There is a drawback of the above approach: The in-silico reconstruction of

haplotypes can generate a proportion of false haplotypes which may hamper the finding of rare but true haplotypes. We have proposed an alternative two-stage approach to the association analysis with GWAS data. Our major contribution is to develop a method for co-classifying genotypes in terms of their penetrances to the disease. In Stage 1, we cluster the genotypes through a finite mixture model, followed by estimating the risk genotypes. In Stage 2, we infer the risk haplotypes from the estimated risk genotypes by using the software PHASE and the odds ratio thresholding. We have proposed a novel data-partition-based initialization for the associated EM algorithm.

We have examined the performance of the proposed procedure by simulations and applications to the CAD and HT data generated from the WTCCC. Compared to the standard multiple Z-testing method, the proposed procedure has been shown to be more powerful in terms of sensitivity and specificity for detecting the true risk haplotypes. In the real data analysis, we have rediscovered some existing risk gene and haplotypes and identifying many more risk haplotypes than did the multiple Z-test based approach. This is not surprising as the simulations have already demonstrated that the model-based clustering can perform better than the multiple Z-test. The Bonferroni adjustment for multiple testing has been applied when multiple tests or thresholding are involved. We note that the results may be further improved if we use advanced multiple testing adjustment methods.

5. PERMUTATION APPROACH

5.1 Introduction

Genome-wide association studies (GWAS) have played an important role in detection genetic risk variants since microarray gene chips became more efficient in screening DNA sequences. The early studies were dealing with SNP data to identify the risk ones. However, some of these SNP do not provide enough information about disease prevalence if they are conducted individually for the study. For that reason, considering multi-SNP studies provides more information about diseases. As a result, the combination of their alleles will form segments of genotypes/haplotypes that might cause disease.

As the haplotype structures are unknown. We need to infer their structures from the observed genotypes. Several softwares have been developed for this purpose such as PHASE (Stephens et al., 2001). Phase employed EM algorithm and Gibbs sampling to infer the most likelihood genotype for each individual in the sample, given all the possible haplotypes' pairs that are consistent with the observed genotypes. There is two way of reconstructing the haplotypes. The first one is by reconstructing the haplotypes of the cases and the controls separately. The second one is by reconstructing the haplotypes of the cases and the controls as one sample without loosing the risk identity of the individuals.

One of the statistical methods to be employed is permutation test which benefits from the rapid developments in the modern computers to minimise the necessary run time. Some of these frameworks were proposed to examine the independence in the distributions for categorical variables by using χ^2 test for the contingency table of observed frequencies (Finos and Salmaso, 2004). More to the point, permutation tests can also be used to investigate marker-disease association for case-control data (Zhao et al., 2000).

Permutation test is also commonly used in haplotypes-based studies to detect the risk haplotypes underlying some diseases by using case-control data. Like many

statistical techniques, permutation tests may face difficulty in case-control studies regarding adjustment for confounding variables such as rare variant in the sample (Epstein et al., 2012).

To cope with these difficulties, we propose a permutation method to detect disease-risk haplotypes by considering two ways of phasing the haplotypes. The first one by reconstruct the haplotypes of cases and controls separately, whereas in the second one, we reconstruct all the haplotypes of both cases and controls samples altogether.

5.2 Method

The proposed permutation methods can be described by two stages as follows.

Stage 1 (Identifying the potential risk genotypes): Let $G_j, 1 \leq j \leq J$ be the observed genotypes with genotypes counts $N_{0j}, N_{1j}, 1 \leq j \leq J$ in controls, cases, respectively. Let $N_0 = \sum_{j=1}^{J} N_{0j}$ and $N_1 = \sum_{j=1}^{J} N_{1j}$. Let $G_j = (h_{j1}, h_{j2})$ be the observed haplotype pairs that forms the genotype G_j. We also let $H = \{h_k, 1 \leq k \leq K\}$ denote the distinct haplotypes inferred from $\mathbf{G}$, where $\mathbf{G} = \{G_j, 1 \leq j \leq J\}$ with haplotype counts $n_{0k}, n_{1k}, 1 \leq k \leq K$ in the controls and cases, respectively, with total counts n_0, n_1. Then, the respective frequencies of the genotype G_j in the controls and cases can be estimated by $q_{0j} = N_{0j}/N_0$, $q_{1j} = N_{1j}/N_1$, respectively. The cases and controls samples are phased by PHASE separately to infer the haplotypes structure.

We then use PHASE again to reconstruct these haplotypes by consider them as one sample. In other words, we put the ambiguous genotypes of the cases and the controls individuals together in the same input file of PHASE. Without loosing the risk identity of each individual, PHASE can reconstruct these genotypes based on the information of the whole sample. The advantage of doing so is to allow PHASE to enrich the disease probability in the second way of reconstruction which results in significant frequencies for the risk haplotypes in the two ways. Providing there are risk haplotypes in the samples, finding the significance in the frequencies can be done by using permutation test.

We then perform the randomization method by swapping the individuals' genotypes of the latter sample between cases and controls. To do so, we choose a half of the individuals in the cases and permute their genotypes with the same number of the individuals in the controls. We then calculate the new respective fre-

quencies of the current permutation. Let $q^*_{i0j} = N^*_{i0j}/N_0$, $q^*_{i1j} = N^*_{1j}/N_1$, respectively, where $i = 1, 2, 3, ..., I$, and I is the total number of the permutations. Here, $N^*_{i0j}, N^*_{i1j}, 1 \leq i \leq I$ represent the counts of the genotype j in the new controls and the new cases, respectively, at the current permutation i. In our later applications of this method to the simulation and the real data, we choose I equal to 1000. The average frequencies of I permutations of all the genotypes in the all new controls and all the new cases can be calculated by $q^*_{0j} = (\sum_{i=1}^{I} q^*_{i0j})/I$, $q^*_{1j} = (\sum_{i=1}^{I} q^*_{i1j})/I$, respectively.

As in any permutation technique, we will check for the significance between the original frequencies and the permuted ones. If any significant different has been found, the corresponding genotype would be considered as a potential risk genotype.

Under the null hypothesis that there is no significant difference between q^*_{1j} and q_{1j} the statistic T_j of the genotype j is distributed as standard normal. T can be calculated as follows:

$$T_j = \frac{q_{1j} - q^*_{1j}}{\delta_j},$$

where

$$\delta_j = \sqrt{\frac{\sum_{i=1}^{I}(q_{1j} - q^*_{i1j})^2}{I-1}}.$$

To this end, if T_j was larger than a specific threshold, we would consider the corresponding difference $q_{1j} - q^*_{1j}$ significant. Implies that the genotypes G_j is a potential risk genotypes. We then define the set of the potential risk group as follows:

$$S = \{h : h \in \{h_{0j}, h_{1j}\}, G_j = (h_{0j}, h_{1j}))\ , 1 \leq j \leq J,\ T_j > \gamma\},$$

where γ is a pre-defined threshold and it can be chosen according to the multiple tests conducted for all the regions under study for the simulation or the real data.

Stage 2: Performing association test to refine the set S

This stage to be used to prevent inflation of type I error. At this stage, we examine the differences in the frequencies of the haplotypes in the set S in the controls and the cases to find the potential risk group. Let ℓ be the number of all different haplotypes in S. We then refine the above set in terms of their Odd Ratio tests. Let $n_{0\overline{r}} = \sum_{h_k \notin S} n_{0k}$, $n_{1\overline{r}} = \sum_{h_k \notin S} n_{1k}, 1 \leq k \leq K$ denote the cumulative frequency of all non-detected haplotypes by the first stage in controls and cases, respectively. The corrected OR statistic for the haplotype $h_\nu, 1 \leq \nu \leq \ell$ is defined

by

$$\mathrm{OR}_\nu = \frac{(n_{1\nu} + 0.5)(n_{0\bar{r}} + 0.5)}{(n_{0\nu} + 0.5)(n_{1\bar{r}} + 0.5)}.$$

Then, the risk haplotype set S_r is updated by

$$S_r = \{h_\nu \in S : \mathrm{OR}_\nu \geq \exp(c_1 \phi(n_{0\nu}, n_{1\nu}, n_{0\bar{r}}, n_{1\bar{r}}))\},$$

where

$$\phi(n_{0\nu}, n_{1\nu}, n_{0\bar{r}}, n_{1\bar{r}}) = \sqrt{1/(n_{0\nu} + 0.5) + 1/(n_{1\nu} + 0.5) + 1/(n_{0\bar{r}} + 0.5) + 1/(n_{1\bar{r}} + 0.5)}$$

and c_1 is a pre-specified constant and it can be set in the real data analysis according to Bonferroni adjustment.

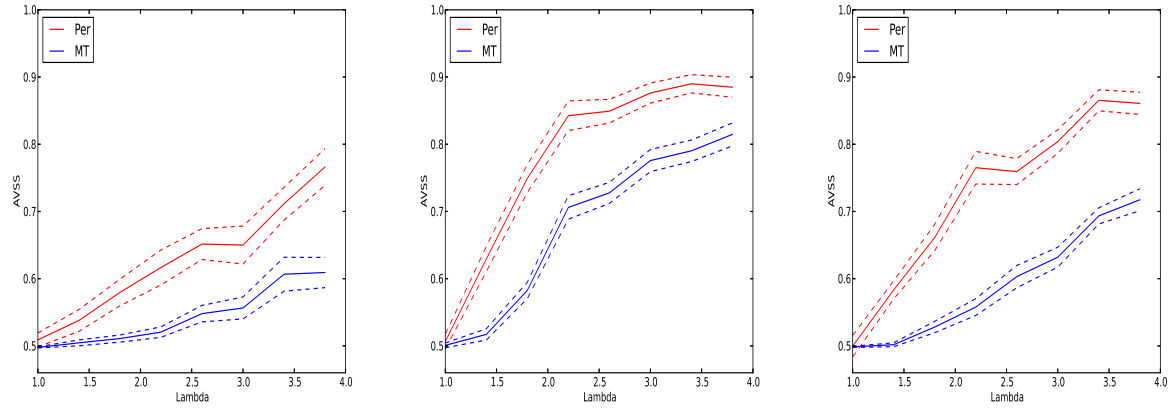
5.3 Simulation

In application of this method, we used the same data we generated according to cohort design and case-control design that we described in 3.3.2.

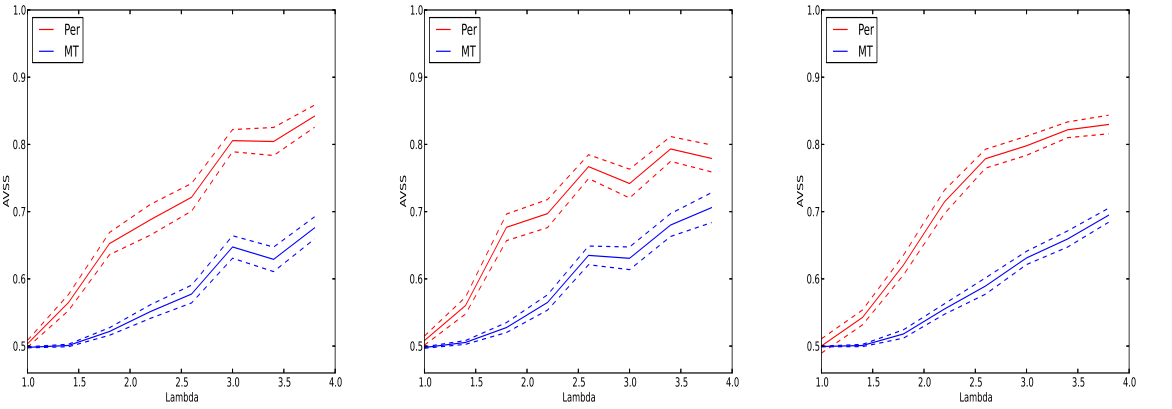
Application of permutation method to cohort design

We applied our method and the multiple testing method to these datasets. We compare them in terms of AVSS. As it is shown in the figures that our method outperformed multiple tests method. In the plots of Figure 5.1, the red and the blue solid curves, showing means of the AVSS values over 30 datasets, were plotted against the values of λ for the permutation method and the multiple testing method respectively. The two red dash curves are one standard error up and down from the red mean curves. Similarly, the two blue dash curves are one standard error up and down for blue mean curves. The plots in the columns from the top to the bottom are for the cases where there were 5, 10, and 20 risk haplotypes in the underlying haplotypes. The sample size of the cases and the controls altogether for each plot is 5000. The plots are based on different modes of inheritance. Figure 5.2 shows the results based on the same scenarios except the sample sizes were 3000. In both figures 5.1 and 5.2, we can clearly conclude that our method shows some advantages over multiple testing method. As it can be seen the methods have resulted in AVSS equal or close to 0.5 when the GRR (λ) equal to 1. This makes sense as we expected a specificity equal to 0 and sensitivity equal to 1 when $\lambda = 1$. This is also one of the conditions that determine setting the pre-specified constants in the thresholds of the final stages of our method c_1 and multiple testing method (P-value threshold

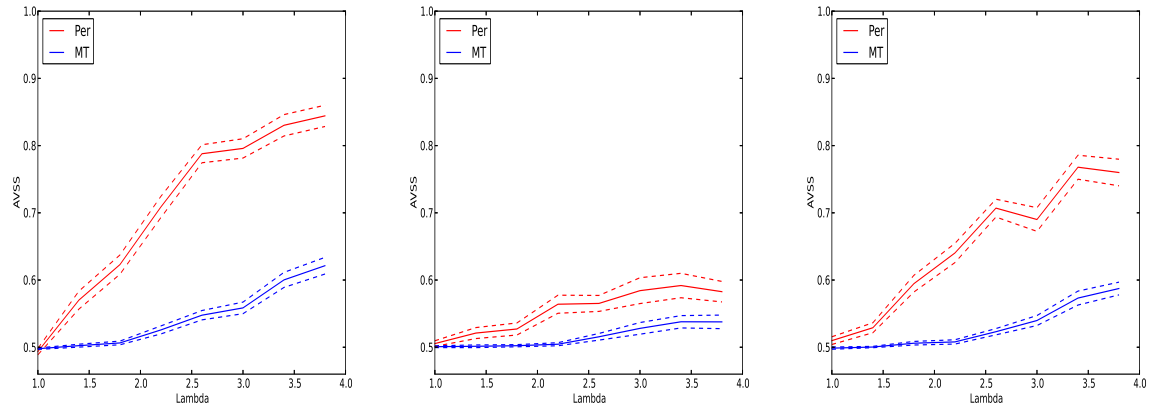
for Fisher's exact test=0.001). The figures clearly demonstrate that as λ value increased, the AVSS increased of both methods. However, our method showed that the AVSS is higher than multiple testing in about 0.05-0.10 when λ is greater than 2 in Figure 5.1 that represents figures of datasets of sample size 5000. The striking results are shown in Figure 5.2 when the datasets of size 3000 as the difference in the AVSS of our method and the multiple testing method could hit 0.20 in some figures except in the case where $m_r = 20$ and IM =dom as both methods showed less power in detecting risk haplotypes.



(a) N=5000, m_r=5, rece (b) N=5000,m_r=5,dom (c) N=5000,m_r=5,mult

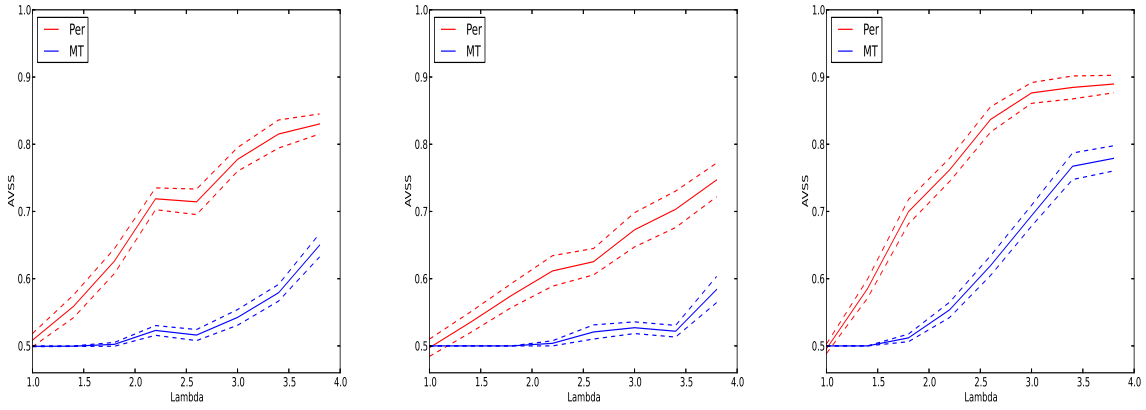


(d) N=5000,m_r=10,rece (e) N=5000,m_r=10,dom (f) N=5000,m_r=10,mult

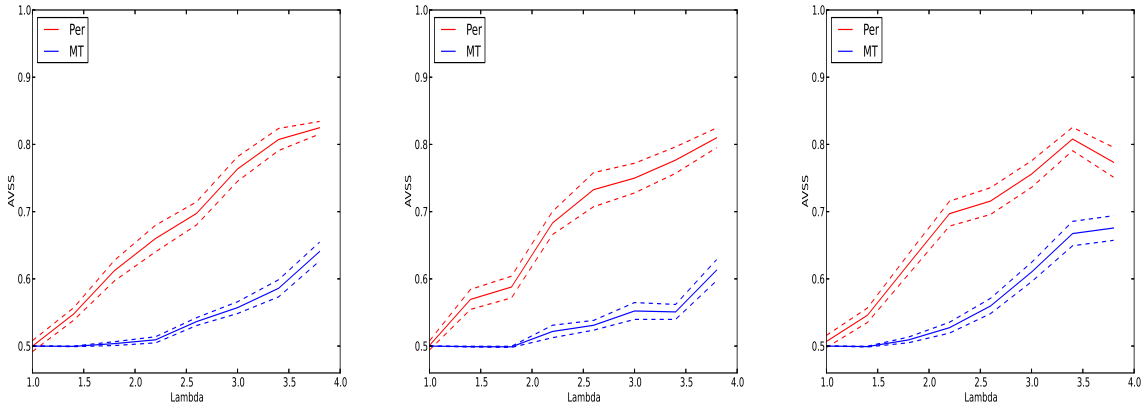


(g) N=5000,m_r=20,rece (h) N=5000,m_r=20,dom (i) N=5000,m_r=20,mult

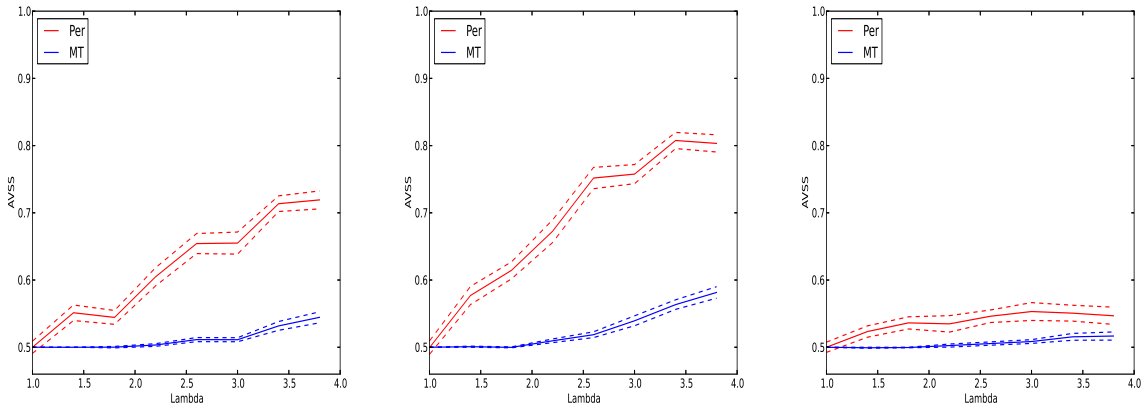
Fig. 5.1: Performances of the proposed permutation method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models based on sample sizes of 5000.



(a) N=3000,m_r=5,mult (b) N=3000,m_r=5,rece (c) N=3000,m_r=5,dom



(d) N=3000,m_r=10,mult (e) N=3000,m_r=10,rece (f) N=3000,m_r=10,dom



(g) N=3000,m_r=20,mult (h) N=3000,m_r=20,rece (i) N=3000,m_r=20,dom

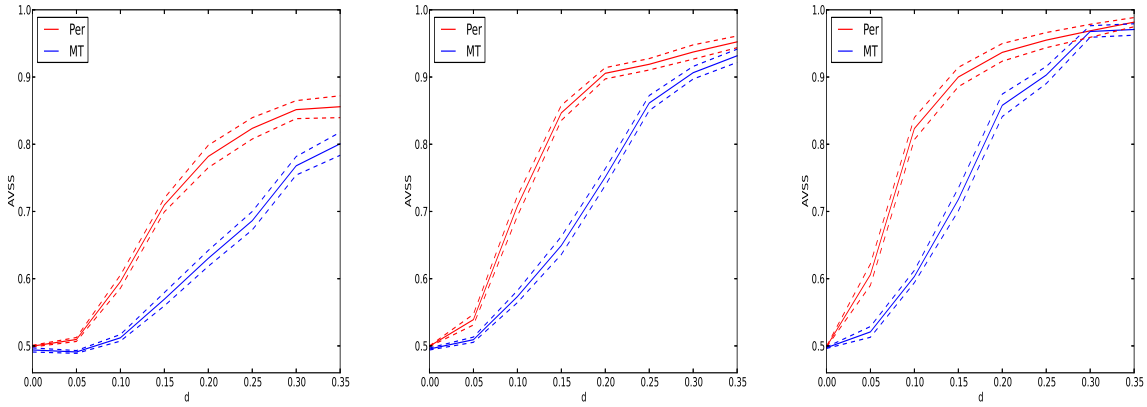
Fig. 5.2: Performances of the proposed permutation method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models based on sample sizes of 3000.

Application of permutation method to case-control design

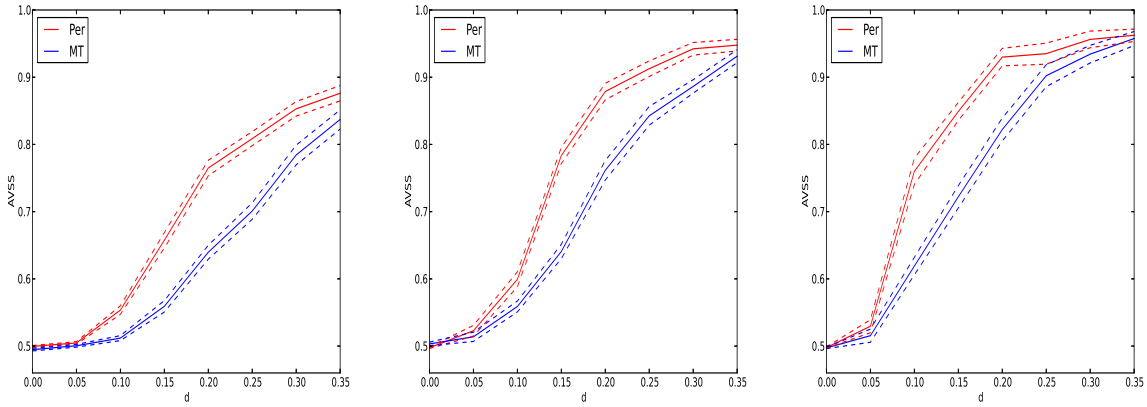
We applied the permutation method and the multiple testing method to these case-control data. The mean curves of the AVSS values with one standard error up and down were plotted against the d values in Figure 5.3. In this figure, the plots in the columns from the left to the right are for the scenarios, where the underlying number of risk haplotypes $m_r = 5, 10$, and 20. The top row stands for the cases, where $N = 5000$, while the bottom row stands for the cases, where $N = 3000$. In these plots, the red and the blue solid curves show mean curves of the AVSS values over 30 datasets as functions of $d = 0, 0.05, 0.1, 0.1,\ 0.15, 0.2, 0.25, 0.3$, and 0.35 for the hybrid mixture method and the multiple testing method respectively. The dash curves are one standard error up or down from the mean curves.

In these figures, we can see that the permutation method achieved some advantages in detecting risk haplotypes compared to multiple testing method in all conducted scenarios.

The main difference between the two methods in their performance is the sample sizes that were considered. In our method, we considered all the cases and the controls to perform the first stage of our method, whereas in the multiple testing method, only subsets of the cases and the controls were considered at the first stage, which could result in loosing some of risk haplotypes in the first place.



(a) N=5000,m_r=20 (b) N=5000,m_r=10 (c) N=5000,m_r=5



(d) N=3000,m_r=20 (e) N=3000,m_r=10 (f) N=3000,m_r=5

Fig. 5.3: Performances of the proposed permutation method and the multiple testing method on the case-control data.

5.4 Real data analysis

We applied the proposed permutation method to the GWAS genotype datasets on coronary artery disease (CAD) and hypertension (HT) obtained by Affymetrix 500K SNP chips in the WTCCC study (WTCCC, 2007). The datasets are prepared in the same way we described in 3.5. The whole genome resulted in 1983537 genotypes in total for the CAD data and 2097111 genotypes in total for the HT data. We used the total number of the genotypes to apply Bonferroni correction to the threshold of the permutation test. Given that we use a rejection level 0.05 for each genotype in all regions, the corrected rejection level for all the genotypes would be $0.05/1983537 = 2.52 \times 10^{-8}$, $0.05/2097111 = 2.38 \times 10^{-8}$ for CAD data and HT data, respectively.

Therefore, we use z-test threshold equal to 5.5 for the CAD and HT genotypes. The permutation test in the first stage of our procedure result in 1433 potential risk haplotypes in CAD data and 430 potential risk haplotypes in HT data.

Note that there were two sub-populations in controls. We applied further filtering on the regions to exclude the ones that have significant differences in the haplotypes frequencies within the two sub-control samples. The exclusion criterion was based on calculating chi-square p-value. Any region resulted in p-value less than 0.30 was excluded from the suspicious regions. This criterion was concluded from the simulated case-control samples when the risk factor d is less than 0.15 as we found out that the p-values for most of the 30 datasets are greater than 0.30, see Figure 3.6.

Note that the declared risk haplotypes at the end of the second stage should also meet this criterion. Toward this end, the chi-square p-value of the frequencies in the two sub-control samples of each potential risk haplotype should be greater than 0.30. Otherwise the haplotype would be eliminated.

At the final stage, we used odds ratio test to refine the groups of the potential risk haplotypes. We calculated the ORs for all the estimated haplotypes and thresholded them by using the bound

$$\exp(c_1\sqrt{1/(n_{0H}+0.5)+1/(n_{1H}+0.5)+1/(n_{0\bar{r}}+0.5)+1/(n_{1\bar{r}}+0.5)}),$$

with $c_1 = 4$, 3.6 for CAD data and HT data, respectively, according to the corrected rejection levels 0.05/1433 and 0.05/430 for CAD and HT, respectively. This gave the final risk-haplotype set as displayed in Tables 5.1, 5.2, 5.3 below.

In these tables, each haplotype has been assigned to a physically closest gene on the basis of the information provided in the GWAS catalog and the genetic information from the British 1958 Birth cohort *http://www2.le.ac.uk/projects /birthcohort/1958bc*. In the CAD case, we did rediscover the CAD risk gene CDKN2B and the risk haplotype $GGTGCCAG$ found by the previous study (WTCCC, 2007; Zhu et al., 2010).

Tab. 5.1: The suspicious regions for coronary artery disease of WTCCC data detected by permutation method.

CAD								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
1	3910010 – 3932838	$rs4654522 - rs10915469$	$CGACGGCC$	0.04238	0.01861	3.09933	4.5×10^{-16}	$hCG_2036596$
1	1902751 – 37450147	$rs6673253 - SNP_A$	$CAACGGAT$	0.05116	0.03019	2.33902	3.0×10^{-14}	LOC728431
1	202166400 – 202187685	$rs6692041 - rs1041311$	$AAATGGGA$	0.07815	0.05083	1.72409	4.3×10^{-09}	LOC284577
1	225406446 – 225425470	$rs4654697 - rs10916399$	$TTGTAAAA$	0.06155	0.03524	1.85056	8.1×10^{-10}	RHOU
1	227569611 – 227620956	$rs7514972 - rs9431663$	$CGCGTAGG$	0.05807	0.0297	2.06768	2.2×10^{-12}	TRIM67
1	239380743 – 239454253	$rs2491826 - rs7533316$	$AGCTCACG$	0.09857	0.07858	1.63864	7.4×10^{-08}	CEP170
1	240360846 – 240438647	$rs12083813 - rs472276$	$CAACATAG$	0.01905	0.00712	2.94026	2.1×10^{-08}	AKT3
2	3789586 – 3821960	$rs7576476 - rs12618184$	$GCTTACAG$	0.03451	0.01119	3.14706	3.1×10^{-15}	LOC442006
2	$rs2314703 - 3942429$	$SNP_A - 1841609$	$CACGCCGT$	0.02055	0.00552	3.78775	3.3×10^{-11}	LOC442006
2	49934439 – 50000082	$rs6736617 - rs17039375$	$CCAAAGGT$	0.02347	0.00757	3.09136	2.7×10^{-10}	NRXN1
2	81525887 – 81577090	$rs1011364 - rs17020239$	$GGATGTGC$	0.03758	0.0202	1.96428	1.3×10^{-07}	LOC442021
3	2557255 – 2599938	$rs6787604 - rs2619566$	$AAGGACGA$	0.07666	0.04763	1.64989	3.1×10^{-08}	CNTN4
3	14422977 – 14471151	$rs4684216 - rs9834629$	$GATGATGC$	0.01815	0.00509	3.67773	8.7×10^{-10}	SLC6A6
3	73461569 – 73510299	$rs7647311 - rs3845868$	$AGGCGCGG$	0.03876	0.01161	3.98169	6.9×10^{-23}	PDZRN3
3	197256495 – 197339533	$rs6583286 - rs9834962$	$TAGACTTA$	0.0498	0.02364	2.17213	2.5×10^{-11}	TFRC
4	3636361 – 3700212	$rs10025237 - rs16844722$	$GGGGAGGG$	0.22491	0.15492	1.62473	6.4×10^{-15}	FLJ35424
4	167440772 – 167457521	$rs9995087 - rs17047336$	$GGACGCAG$	0.03434	0.01139	3.12327	8.2×10^{-14}	TLL1
5	124765522 – 124843518	$rs4836190 - rs13187198$	$TGAAGGCA$	0.04275	0.02795	2.02205	2.0×10^{-09}	LOC644659
5	157267571 – 157303032	$rs10071157 - rs17055168$	$GTGAGCAA$	0.02135	0.00701	3.09771	9.0×10^{-10}	CLINT1
5	166764561 – 166801933	$rs6863935 - rs7724862$	$CTATGTGT$	0.09145	0.05448	1.63602	8.8×10^{-09}	ODZ2
7	77695246 – 77717237	$rs2215379 - rs4515471$	$TCTAAAAA$ $CTTGGAAA$	0.03291 0.03609	0.01786 0.01061	2.04961 3.77003	1.7×10^{-07} 7.3×10^{-19}	MAGI2
7	153371858 – 153449397	$rs6464391 - rs1861139$	$CGGGTAGA$	0.04119	0.02159	2.31998	1.7×10^{-11}	LOC653748
8	71022178 – 71086937	$rs7836791 - rs388511$	$TACAGAAG$	0.02204	0.00555	3.68611	4.1×10^{-11}	SLCO5A1
9	22088619 – 22120515	$rs2891168 - rs10965245$	$GGTGCCAG$	0.34939	0.29298	1.40724	3.2×10^{-13}	CDKN2B
9	74180343 – 74241329	$rs10114124 - rs17081046$	$GTATTTAT$	0.21608	0.13046	1.66562	4.0×10^{-17}	RORB
9	77341767 – 77366988	$rs2889774 - rs3780296$	$ATGGAAAT$	0.06672	0.042	1.69537	1.2×10^{-07}	GNA14
9	119506057 – 119537035	$rs2191675 - rs10984648$	$GTTGGCTA$	0.08762	0.03361	2.8056	1.8×10^{-28}	CDK5RAP2
9	135269746 – 135320703	$rs731533 - rs7870302$	$TGTCTCCC$	0.03175	0.01296	2.57076	9.3×10^{-11}	OLFM1
10	11879196 – 11924252	$rs6602535 - rs11257355$	$TCTGCCGG$	0.1694	0.12811	1.57916	1.3×10^{-12}	C10orf47
10	14795325 – 14817082	$rs2688827 - rs12246518$	$ATGACCGC$	0.34815	0.32333	1.71018	4.1×10^{-09}	FAM107B
11	8165969 – 8200374	$rs4758310 - rs11041816$	$ATAATGGG$ $GCTGTAGA$	0.36298 0.05243	0.3164 0.02741	1.34831 2.24619	2.8×10^{-08} 7.5×10^{-12}	LOC644497
11	36361306 – 36410807	$rs330255 - rs331485$	$GCGATTAA$	0.0309	0.00779	4.20172	5.6×10^{-18}	FLJ14213
11	69213458 – 69295251	$rs1192923 - rs3168175$	$TCGTGGCA$	0.10225	0.05587	2.24141	5.7×10^{-21}	FGF4
11	83230307 – 83256927	$rs1878266 - rs1878264$	$TATATTCA$	0.03571	0.01807	2.24283	6.3×10^{-09}	CCDC90B
11	99383206 – 99391536	$rs3911286 - rs10501939$	$TTAGATAT$	0.03303	0.01472	2.21561	9.3×10^{-09}	CNTN5
11	112952870 – 113015533	$rs4936278 - rs12577253$	$CCTCGTGC$	0.05824	0.03474	1.75496	1.9×10^{-08}	DRD2
11	129102667 – 129124330	$rs532427 - rs691197$	$ACCGCGGA$	0.08519	0.05612	1.73953	2.1×10^{-11}	TMEM45B
11	133079508 – 133113640	$rs4937817 - rs4937826$	$CCGGCCCG$ $GTAGCCCG$ $GTAGTGCC$	0.05747 0.04001 0.04216	0.04018 0.02779 0.02425	1.89429 1.90705 2.30133	5.6×10^{-10} 9.3×10^{-08} 8.2×10^{-12}	LOC646522
12	5619429 – 5628923	$rs11063791 - rs454704$	$TACATAAA$	0.02897	0.0124	2.50152	8.0×10^{-10}	TMEM16B
12	112703139 – 112738033	$rs11066758 - rs7137339$	$ACGGTCAC$	0.02681	0.01286	3.14709	1.5×10^{-12}	RBM19
12	116500495 – 116514298	$rs10850852 - rs1400593$	$CTCTCTTT$	0.14523	0.12089	3.21401	8.3×10^{-21}	NOS1

Tab. 5.2: Continuation of Table 5.1.

CAD								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
			CTCTCTTC	0.28034	0.26232	2.85847	1.5×10^{-19}	
13	108372995 – 108432811	$rs4773010 - rs3842945$	*AGAGACCC*	0.27486	0.19222	1.59282	1.3×10^{-21}	MYO16
14	25140850 – 25159405	$rs8020556 - rs1951062$	*AGTAAACT*	0.09084	0.02999	3.36068	1.2×10^{-37}	LOC401767
14	53221435 – 53244046	$rs1563719 - rs210351$	*AGATAGGT*	0.15385	0.10566	1.56278	1.2×10^{-12}	BMP4
14	65343491 – 65401760	$rs3924222 - rs12896836$	*TATAACTC*	0.0462	0.01904	2.70766	1.1×10^{-16}	FUT8
15	20624103 – 21246055	$rs7166056 - rs8024346$	*GTGACGTG*	0.08093	0.04109	1.90364	2.7×10^{-12}	NIPA1
15	21729952 – 21760003	$rs4778264 - rs9796712$	*TGATAGGG*	0.03064	0.00783	3.91789	2.2×10^{-16}	MAGEL2
15	37962389 – 38014169	$rs11633436 - rs534757$	*TTACAACC*	0.07798	0.03763	2.31448	1.1×10^{-18}	GPR176
16	55207138 – 55253047	$rs8055724 - rs12447986$	*TTCTCCTC*	0.03044	0.01113	2.89551	1.5×10^{-09}	MT1L
16	79852394 – 79892297	$rs6564863 - rs11639552$	*TTCGTTAT*	0.02663	0.01053	3.15992	2.7×10^{-10}	BCMO1
17	27921023 – 27963104	$rs225215 - rs17780520$	*GGGTTAAC*	0.0205	0.00465	4.05617	2.7×10^{-11}	MYO1D
17	74629176 – 74682195	$rs2612793 - rs8072667$	*CGAGGTTG*	0.06276	0.03471	1.82966	4.4×10^{-09}	FLJ21865
18	8212591 – 8279839	$rs10468776 - rs11876033$	*GGGACAAG*	0.02689	0.00982	2.94852	1.7×10^{-11}	PTPRM
18	2291328 – 22715430	$rs3974646 - SNP_A$	*TGCGGAGT*	0.05382	0.02751	1.98739	2.3×10^{-10}	AQP4
18	32033296 – 32083366	$rs8095718 - rs8082899$	*CAAAACCA*	0.0592	0.04484	1.65827	1.7×10^{-07}	MOCOS
19	6641966 – 6717213	$rs3745566 - rs7248911$	*TAAGCTAC*	0.02312	0.00521	4.97801	1.0×10^{-14}	C3
19	15365766 – 15477256	$rs7257156 - rs6512039$	*AAGCGCGG*	0.08169	0.05278	1.69741	1.1×10^{-09}	AKAP8L
19	17595848 – 17649789	$rs10419511 - rs7252308$	*TTGGTATG*	0.04657	0.01971	2.8095	1.1×10^{-17}	UNC13A
19	18225800 – 18277972	$rs10417536 - rs4808781$	*CTCCGCAA*	0.04034	0.02211	1.94095	6.7×10^{-08}	LOC729966
19	52946204 – 53026777	$rs10402957 - rs4427918$	*CATTCAGC*	0.0741	0.04321	1.81613	4.1×10^{-10}	GLTSCR2
20	5604763 – 5643174	$rs8118780 - rs805726$	*CCGTAGTA* *CTTTAGTA* *CTTTAGTG*	0.05455 0.01801 0.01698	0.03836 0.00794 0.00777	1.76976 2.81211 2.7096	1.3×10^{-08} 2.7×10^{-08} 1.6×10^{-07}	C20orf196
20	6055964 – 6078025	$rs6117090 - rs3897509$	*AGGCCGCA* *AAGCCGAA*	0.09945 0.03039	0.05857 0.01269	1.89101 2.66015	9.9×10^{-13} 1.2×10^{-09}	C20orf42
20	51996013 – 52017348	$rs12480336 - rs6013853$	*CACCGATC*	0.02844	0.01511	2.17303	1.5×10^{-07}	BCAS1
20	55607831 – 55637003	$rs17498081 - rs17414380$	*CAATGTCC*	0.02768	0.01127	2.6821	1.2×10^{-09}	TMEPAI
22	16871076 – 16895136	$rs8142200 - rs975826$	*TCGGGAGG*	0.03219	0.00253	12.43113	5.4×10^{-28}	LOC729269
22	35324014 – 35335429	$rs7410412 - rs12160203$	*GCCTAGGG*	0.1967	0.14314	1.46774	4.7×10^{-11}	CACNG2

Tab. 5.3: The suspicious regions for hypertension of WTCCC data detected by permutation method.

HT								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
2	39199834 – 39248354	$rs6758330 - rs10184046$	*CGCCAAAA*	0.03665	0.00147	26.83195	1.3×10^{-31}	SOS1
4	17856580 – 17878437	$rs11941617 - rs1503880$	*GTATTTGT*	0.0584	0.00019	236.45945	1.2×10^{-73}	LCORL
6	107236669 – 107248636	$rs3121432 - rs2354550$	*TGATTGTC*	0.07759	0.00247	35.82646	6.5×10^{-82}	QRSL1
10	30990752 – 31024312	$rs16931828 - rs7078126$	*AGTGTTGC* *AACGTTGT* *AGCTCTGC* *GGCGCCGC*	0.47318 0.06589 0.24167 0.10573	0.47676 0.00314 0.24983 0.10377	1.45455 29.93248 1.41785 1.49364	1.0×10^{-08} 3.1×10^{-79} 1.2×10^{-06} 4.1×10^{-06}	LOC645954
11	55290776 – 55324792	$rs11825590 - rs17501618$	*GCCTGTGT*	0.04351	0.00947	4.47895	4.1×10^{-22}	OR5D14
11	121093256 – 121139818	$rs92061 - rs4936651$	*AATGCTGG*	0.86672	0.79508	2.49843	1.4×10^{-30}	SORL1
18	73486971 – 73493301	$rs1553419 - rs4890980$	*TTGGGTTC*	0.03825	0.00893	4.49948	2.9×10^{-21}	LOC728864

5.5 Discussion and conclusion

One of the important issues in population genetics is population substructure. Therefore, we proposed a permutation method to overcome such issues and successfully detect the risk-haplotypes. In this method, the haplotypes have been reconstructed in two different ways. The first one is by reconstruct the haplotypes of the cases and the controls samples separately. The second one is by reconstruct both samples altogether to stratified the samples and overcome the issue of substructure. We then permute the disease status in the sample, where the cases and the controls were reconstructed altogether.

Our framework has been applied to the simulation and the real data. In the simulation, we compare the results to the multiple testing method of (Zhu et al. 2010). The figures show that our method outperformed the multiple testing method in all scenarios of cohort and case-control designs. The both methods showed less AVSS in detecting risk haplotypes in the case of $m_r = 20$ of the cohort design or case-control design discarding the sample size and the mode of inheritance. The reason behind that is the fact that if the number of the risk haplotypes in the sample is close or equal to the number of the non-risk haplotypes, large number of non-risk haplotypes could be detected as risk ones or large number of risk haplotypes could be detected as non-risk. This due to the fact that the genotypes involving risk and non-risk haplotypes would be more than the others under HWE. This can disinflate specificity or sensitivity and decrease AVSS.

6. CLUSTERING-BASED LOGISTIC REGRESSION

6.1 Introduction

Genome wide association studies have played an important role in identifying genetic risk variants for complex disease. Many of which have employed statistical methods to examine the association of a disease and genetic factors. Several successes have been achieved by some studies to identify risk SNPs, genotypes and haplotypes. Some of which have proposed clustering methods to classify the variants into several groups depending on the degree of association between these variants and diseases (Binder et al., 2012; Huang et al., 2011).

One popular way of clustering methods is fitting the logistic model to case-control samples to detect the risk variants. The popularity of the logistic regression became remarkable as, (1) the log-odds ratios can easily interpret the coefficients in the model, (2) in the case-control samples, the odds ratios for the disease probabilities can be estimable (Breslow et al., 2000).

Many studies were employing the logistic regression by considering SNPs, which are close or far from each other in their physical distance, as covariates for the model and classifying them into risk and non-risk variants (Kang et al., 2008). However, logistic regression models are very sensitive to the high density of SNPs and the linkage between them within candidate genes (Byng et al., 2003). Other than that, confounding variables can result in false association with disease, specially in the case-control studies (Breslow et al., 1988). However, there are some limitations in fitting a logistic model for SNPs data when it comes to examining the covariates' effect on the response variables. In fact, as the number of SNPs considered for the study increased, the number of the covariates in the model would be increased. That might lead to many rare covariates in the model, given small samples sizes. Therefore, many researchers in favour of considering the reconstructed haplotypes or the genotypes in fitting the logistic regression in case-control studies.

Conducting haplotypes-based studies in fitting the logistic model can be more

superior as the biological relationships between disease and haplotypes can be more interpreted than in SNPs studies (Huang et al., 2011). However, some difficulties are still common in such studies such as the haplotypes structure, high dimensionality and the high number of degrees of freedom (Igo et al., 2009). Thus, fitting the logistic regression to the genotypes can be alternative solution for consideration of the haplotypes. The only problems that we need to account are the rare variants and the high dimensionality. For example, but not limited to, applying the standard logistic regression can be disadvantageous in estimation of the model's parameters and lead to no significant association between the response variables and the exposure variables. One possible reason is that we using the standard way of scoring the exposure variable by 1 if they are present in an individual or by 0 if they are not. Yet, the model will equally be affected by the exposure variables whose scores are same. As a result, no significant coefficient can be found when calculating the maximum likelihood estimation.

More to the limitations, the study design is also real challenge as far as the standard logistic regression is concerned. Therefore, undertaking cohort design or case-control design can be tricky in the way that we give weights to the model's covariates. In the cohort design, for instance, the unknown parameters in the model can be estimated by the maximum likelihood function in terms of the parameters vector. Yet, the estimated parameters can be used to estimate the conditional probability of the disease status variable for each exposure variable. However, the maximum likelihood function needs to be extended when the data come from a stratified simple random sample (Chambless and Boyle, 1985). On the other hand, in the case-control design, when the sampling is conditionally performed on the outcome variables, odds ratios can easily obtained and adjusted from the estimated slop coefficients (Breslow, 1996). Furthermore, unlike cohort studies, sample stratification can fix the binary outcome variables in a case-control study.

Therefore, we proposed a new method to handle the above issues by using the logistic regression on the genotypes data. In this method, we use a new way of scoring the exposure variable base on their probabilities, given cases and controls. We also model all the genotypes as they represent one exposure variable.

6.2 Method

6.2.1 Clustering-based logistic method(CL)

Let $G_j, 1 \leq j \leq m$ be individuals' genotypes with counts $N_{0j}, N_{1j}, 1 \leq j \leq m$ in controls, cases, respectively. Let $N_0 = \sum_{j=1}^{m} N_{0j}$ and $N_1 = \sum_{j=1}^{m} N_{1j}$. Let $H_j = \{(h_{j11}, h_{j12}), (h_{j21}, h_{j22}), ..., (h_{j\epsilon 1}, h_{j\epsilon 2})\}$ be the observed haplotype pairs that are consistent with G_j. If that is the case, the counts $N_{0j}, N_{1j}, 1 \leq j \leq m$ represent the frequencies of one haplotype pair in H_j in controls, cases, respectively. We also let $h_k, 1 \leq k \leq K$ denote the distinct haplotypes inferred from $\mathbf{G}$, where $\mathbf{G} = \{G_j, 1 \leq j \leq m\}$ with haplotype counts $n_{0k}, n_{1k}, 1 \leq k \leq K$ in the controls, cases, respectively, with total counts $2N_0$, $2N_1$. We also let $y_i, 1 \leq i \leq N$ denote the disease status of N individuals, where $N = N_0 + N_1$. and y_i is equal 1 if an individual is affected and 0 otherwise.

We developed the following approach to detect the potential risk genotypes in the sample following by the detection of the potential risk haplotypes.

Stage 1 (Identifying the potential risk genotypes):

Let R be the set of the collapsed rare genotypes and it can be define as following:

$$R = \{G_j |\ N_{0j} = 0 \ or \ N_{1j} = 0 \ or \ \frac{N_{0j} + N_{1j}}{\sum_{j=1}^{J}(N_{0j} + N_{1j})} \leq 0.005 \ , \ 1 \leq j \leq m\},$$

with collapsed counts N_{0J} in controls and N_{1J} in cases, respectively. The remaining genotypes will be indexed by j, where $1 \leq j \leq J-1$.

We define d_j to be the difference between the estimated probabilities $\hat{P}(G_j|\text{cases})$ and $\hat{P}(G_j|\text{controls})$. Thus

$$d_j = \log \hat{P}(G_j|\text{cases}) - \log \hat{P}(G_j|\text{controls}), \ 0 \leq j \leq J$$

In our method (CL), we use two levels of collapsing the genotypes to enrich the probability of being diseased in the sample. The first one is, as we mentioned above, collapsing the rare genotypes. In the second level, we undertake a further collapsing for all the genotypes including the first level collapsed group based on d_j. The logic behind that is to let the estimated $p(x_i) = E(Y = 1|x_i)$, which is always biased in a case-control samples, depends on the perspective frequencies that are supposed to be unbiased in the case-control samples. Note that we have to

pay more attention to the case-control design as it the only design that mimics the real data of WTCCC. Therefore, to proceed the second collapsing, we use K-means method, which is proposed by (Lloyd, 1957), to classify these genotypes into several groups based on their d_js. Before going further, let's describe K-means clustering in Python briefly.

K-means clustering:

The way we are describing here is a simple implementation of Lloyd's algorithm for performing k-means clustering in python that we used in the applications of our method. The role of K-means is to classify the genotypes into several groups based of the values of $\{d_j\}$.

Let ($z = [J/2]$, where $[\alpha]$ is the largest integer less than α, denote the initial number of clusters. Let these clusters denoted by $A_t, 1 \leq t \leq z$. Then z values will be chosen by random assignation from the set $\{d_j\}$ to represent the initial centroids, say $c_t, 1 \leq t \leq z$. The algorithm will repeatedly perform the following two steps until reaching the convergence:

(1) The clusters will be updated by adding to them all values of d_j that are closest to the centroids of these clusters in distance

$$A_t = \{d_j : \|d_j - c_t\| \leq \text{ all } \|d_j - c_\ell\|, \ell = 1, 2, ..., t-1, t+1, ..., z\}.$$

(2) Provided a set of clusters, the centroids are recalculated by averaging all points belonging to a cluster.

$$c_t = \frac{\sum_{d_j \in A_t} d_j}{|A_t|}, 1 \leq t \leq z$$

Once the convergence is achieved, the final number of clusters are not necessarily to be equal to z. Usually, it is less than or equal the initial number, which means this is another level of collapsing the genotypes to ensure that disease probability has been well-enriched in the sample. Let $\nu, \nu \leq z$ be the final number of clusters after reaching the convergence.

Assigning K-means centroids as weights to x_i:

We then assign the centroid c_t to all the individuals whose genotype is G_j if $d_j \in A_t$. Let $x_i \in \{c_1, c_2, ..., c_\nu\}$ be the covariate value corresponding to y_i. We

use the logistic regression to estimate the probability of being diseased given the centroid c_t. Let $p(x) = P(Y = 1|X = x)$. The logistic regression model can be written as

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x. \tag{6.1}$$

$p(x)$ can be calculated by

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{6.2}$$

To this end, if y_i is value that Y_i takes and c_t is the value that X_i takes, then the expression for $p(x)$ given in the equation 6.1 can be used to calculate the conditional probability of $Y_i = y_i$ given $X_i = c_t$ when $y_i = 1$. Similarly, the quantity $1 - p(c_t)$ gives the conditional probability $P(Y_i = y_i|X_i = c_t)$ when $y_i = 0$.

Hence, for those pairs (x_i, y_i), where $y_i = 1$, the contribution to the likelihood function is $p(x_i)$, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - p(x_i)$. Therefore, the contribution to the likelihood function for the pair (x_i, y_i) can be expressed as

$$p(x_i)^{y_i} \left[1 - p(x_i)\right]^{1-y_i}.$$

Assuming that the observations are assumed to be independent, we calculate the likelihood function by

$$L(\underline{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i} \left[1 - p(x_i)\right]^{1-y_i},$$

and the log likelihood function by

$$\log L(\underline{\beta}) = l(\underline{\beta}) = \sum_{i=1}^{n} \{y_i \log \left[p(x_i)\right] + (1 - y_i) \log \left[1 - p(x_i)\right]\}.$$

On equating $\frac{\partial l}{\partial \underline{\beta}}$ to zero and solving for $\underline{\beta}$, we find MLE of the parameter vector. In fact there are many computational packages available in R or Python (e.g. glm) that can be used to find MLE. We calculate $p(c_t)$,$1 \leq t \leq \nu$ by using (6.2).

The hypothesis here is that no association between the genotype G_j and disease is equivalent to $d_j = 0$. Therefore, given the estimated coefficients of the model β_0 and β_1, we can estimate $p(0)$ by (6.2). Here, $p(0)$ will be considered as non-risk background probability. We then calculate the odds ratio that correspond $c_t, 1 \leq t \leq \nu$ by

$$W_t = \frac{p(c_t)(1-p(0))}{(1-p(c_t))(1-p(0))}. \tag{6.3}$$

In light of this, we define the potential risk group which represents all the potential risk genotypes is defined as

$$S = \{h; h \in \{h_{0m}, h_{1m}\}, G_j = (h_{0m}, h_{1m}); d_j \in A_t, 1 \leq j \leq J,\ W_t > T\},$$

where T is predefined threshold.

Stage 2 (OR thresholding): In the real data and the simulation, we need to refine the subset of the potential risk haplotypes obtained at stage 1 to prevent the inflation of type I error. The reason behind that the disease prevalence of a genotype can due to only one risk haplotype under the assumptions of the modes of inheritance. We are going to refine the above selected risk haplotype set on the basis of their odds ratios. Let ℓ be the number of all different haplotypes in S. Let $n_{0\bar{r}} = \sum_{h_k \notin S} n_{0k}$, $n_{1\bar{r}} = \sum_{h_k \notin S} n_{1k}, 1 \leq k \leq K$ denote the cumulative frequency of all non-detected haplotypes by the first stage in controls and cases, respectively. The corrected OR statistic for the haplotype $h_\nu, 1 \leq \nu \leq \ell$ is defined by

$$\mathrm{OR}_\nu = \frac{(n_{1\nu} + 0.5)(n_{0\bar{r}} + 0.5)}{(n_{0\nu} + 0.5)(n_{1\bar{r}} + 0.5)}.$$

Then, the risk haplotype set S_r is defined by

$$S_r = \{h_\nu \in S : \mathrm{OR}_\nu \geq \exp(c_1 \phi(n_{0\nu}, n_{1\nu}, n_{0\bar{r}}, n_{1\bar{r}}))\},$$

where

$$\phi(n_{0\nu}, n_{1\nu}, n_{0\bar{r}}, n_{1\bar{r}}) = \sqrt{1/(n_{0\nu} + 0.5) + 1/(n_{1\nu} + 0.5) + 1/(n_{0\bar{r}} + 0.5) + 1/(n_{1\bar{r}} + 0.5)}$$

and c_1 is a pre-specified constant (in our later simulations, we set to value that result in specificity equal to 0 and sensitivity equal to 1 when there is no risk in the sample(e.g. $\lambda = 1$ in the cohort design, $d = 0$ in the case-control design) while in the real data analysis, invoking the Bonferroni adjustment to set it's value.

6.2.2 Standard multiple logistic method

The standard multiple logistic regression (SL) has been mentioned in details in the section 2.7.2. In the SL, we set the independent variables x_{ij} in the equation (2.9)

equal to 1 if the individual i has the genotype G_j, and zero otherwise. The dependent variable y_i would be 1 if the individual i is affected and zero if not. We then use GLM package to declare the potential risk genotypes based on their corresponding p-values. We noticed is always resulting in specificity equal to 0 and sensitivity equal to 1 for all scenarios of both study designs, except two cases in the case-control design where $m_r = 5, 10$. This could due to the high dimensionality of the genotypes and their rarity.

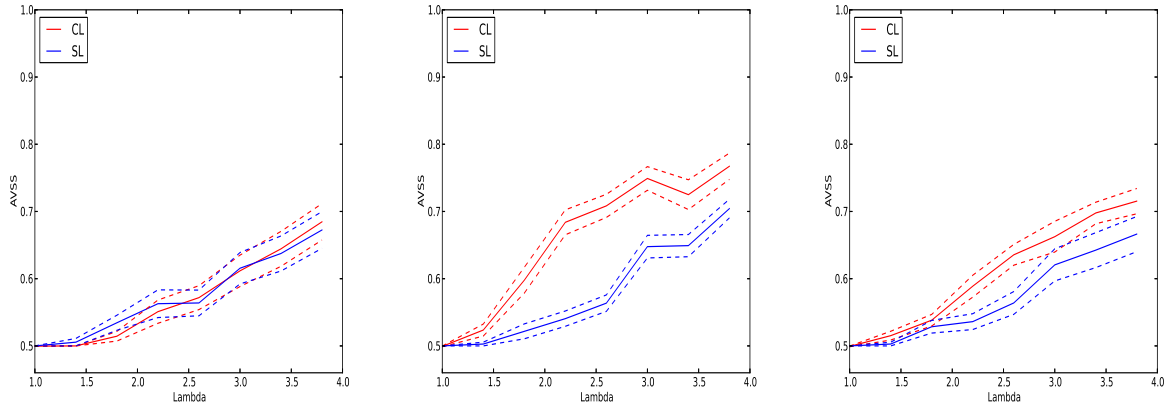
We overcome this drawback by using the same collapsed genotypes set R defined in our method. We achieved a good improvement on the performance of SL as showed in the simulation below.

6.3 Simulation

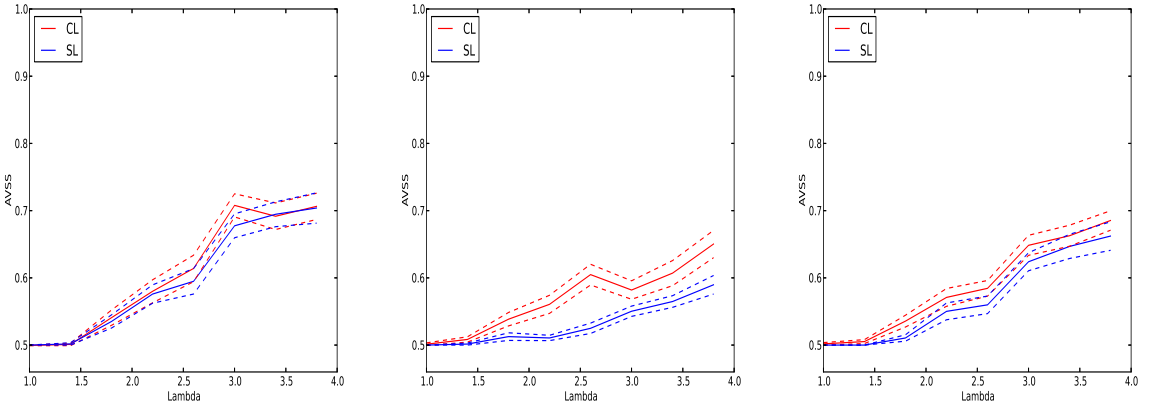
In application of this method, we used the same data we generated according to cohort design and case-control design that we described in Section 3.3.

Application of logistic regression method to cohort design

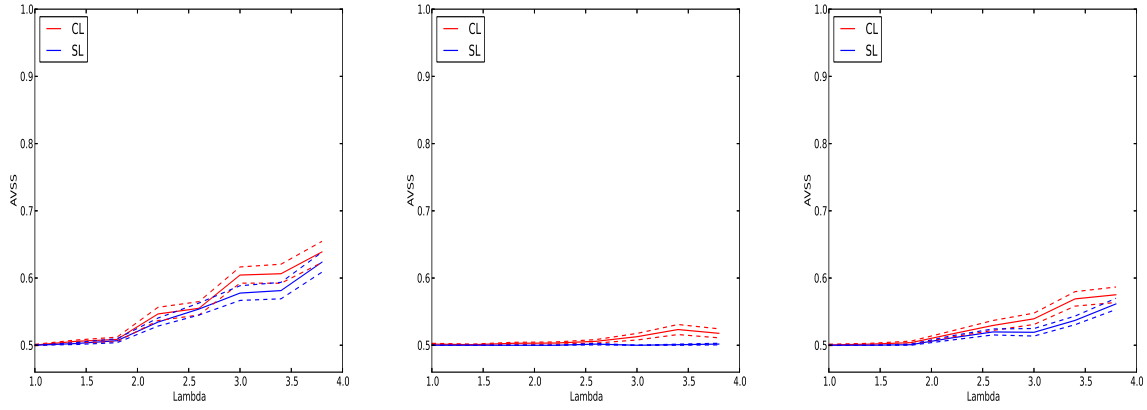
We applied our method (CL) to find the risk haplotypes of the previous datasets. We compare the results to the standard multiple logistic method (SL) in terms of AVSS. In the plots of Figure 6.1, the red and the blue solid curves, showing means of the AVSS values (i.e., the values of (specificity and sensitivity)/2)) over 30 datasets, were plotted against the values of λ for CL method and SL method respectively. The two red dash curves are one standard error up and down from the red mean curves. Similarly, the two blue dash curves are one standard error up and down for blue mean curves. The plots in the columns from the top to the bottom are for the cases where there were 5, 10, and 20 risk haplotypes in the underlying haplotypes. The sample size of the cases and the controls altogether for each plot is 5000. The plots are based on different modes of inheritance. Similarly for Figure 6.2, but the sample size considered is 3000 for the cases and the controls altogether. According to Figure 6.1 and Figure 6.2, our method outperformed the SL in the scenario where $m_r = 10, 5$ and the disease model is dominant. In the other scenarios both methods performed quite similar in terms of AVSS values.



(a) N=5000, m_r=5, rece (b) N=5000,m_r=5,dom (c) N=5000,m_r=5,mult

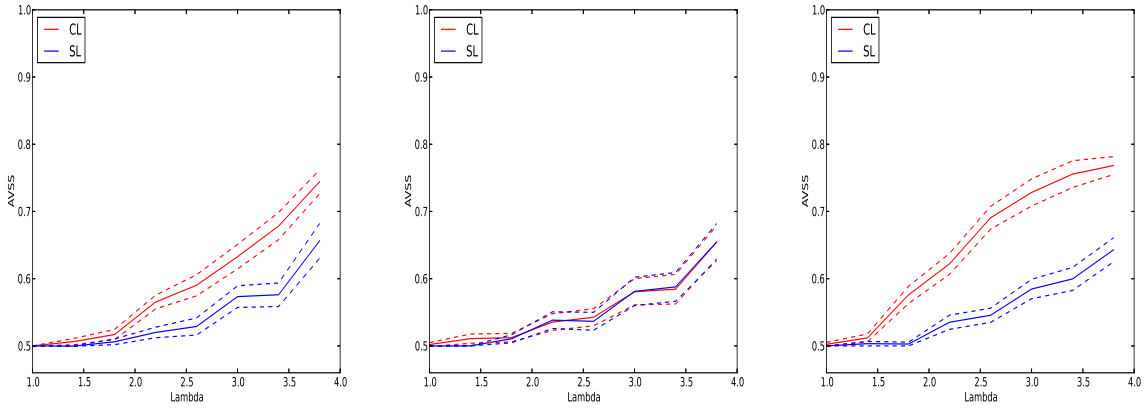


(d) N=5000,m_r=10,rece (e) N=5000,m_r=10,dom (f) N=5000,m_r=10,mult

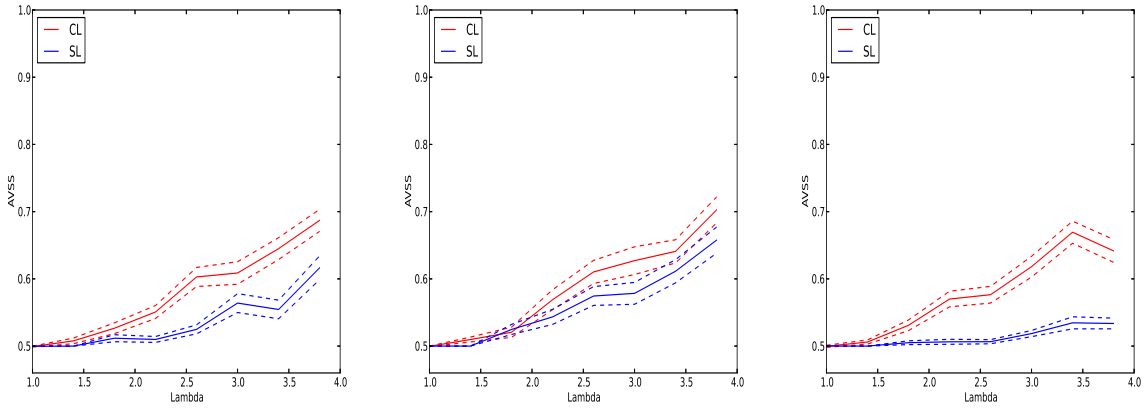


(g) N=5000,m_r=20,rece (h) N=5000,m_r=20,dom (i) N=5000,m_r=20,mult

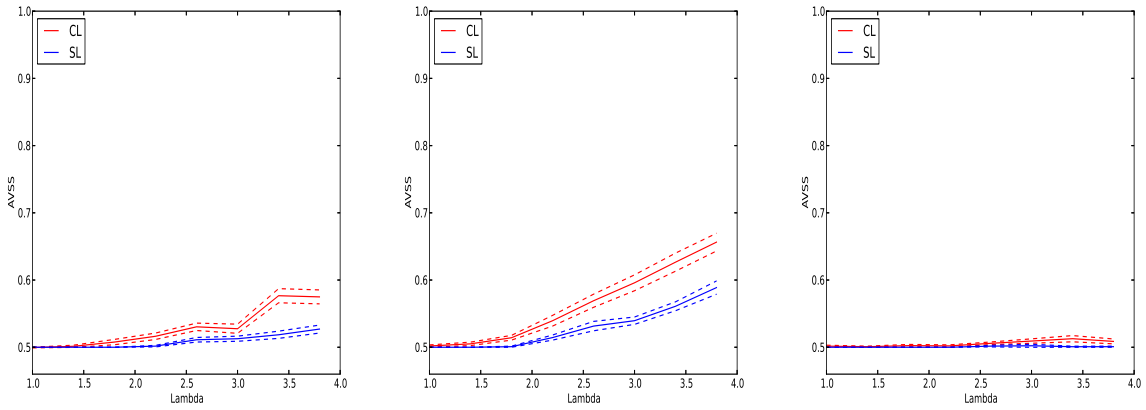
Fig. 6.1: Performances of CL method and SL method on the cohort-design data with multiplicative or dominant or recessive inheritance models with sample size $N = 5000$.



(a) N=3000,m_r=5,mult (b) N=3000,m_r=5,rece (c) N=3000,m_r=5,dom



(d) N=3000,m_r=10,mult (e) N=3000,m_r=10,rece (f) N=3000,m_r=10,dom



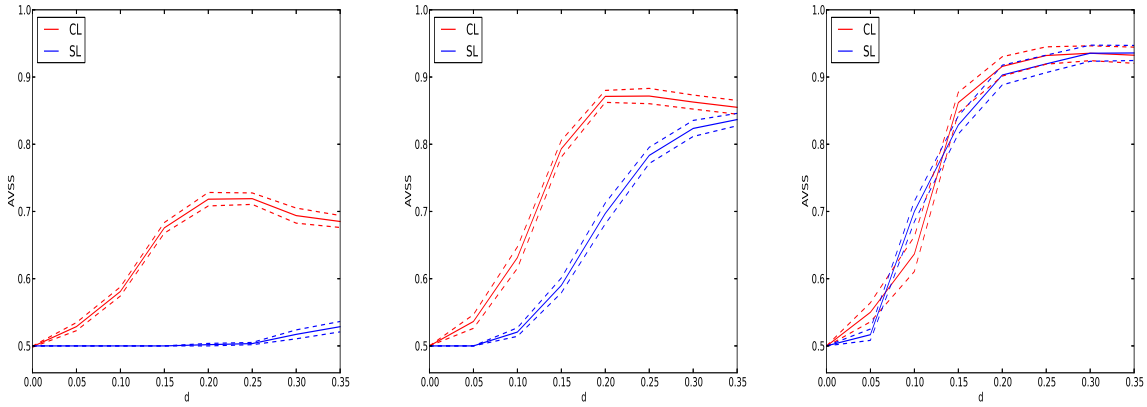
(g) N=3000,m_r=20,mult (h) N=3000,m_r=20,rece (i) N=3000,m_r=20,dom

Fig. 6.2: Performances of CL method and SL method on the cohort-design data with multiplicative or dominant or recessive inheritance models with sample size $N = 3000$.

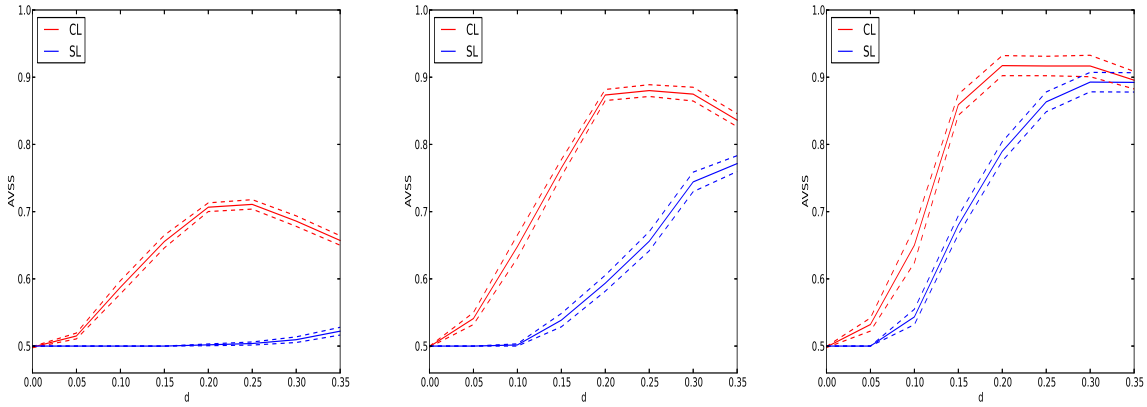
Application of logistic regression method to case-control design

We applied our method (CL) and (SL) method to the case-control samples. The mean curves of the AVSS values with one standard error up and down were plotted against the d values in Figure 6.3. In this figure, the plots in the columns from the left to the right are for the scenarios, where the underlying number of risk haplotypes $m_r = 20, 10$ and 5. The top row stands for the cases, where $(N_1, N_0) = (2000, 3000)$, while the bottom row stands for the cases, where $(N_1, N_0) = (1000, 2000)$. In these plots, the red and the blue solid curves show mean curves of the AVSS values over 30 datasets as functions of $d = 0, 0.05, 0.1, 0.1,\ 0.15, 0.2, 0.25, 0.3$, and 0.35 for CL method and SL methods, respectively.

In this figure, it can be seen that the CL method achieved some advantages in detecting risk haplotypes compared to SL method in the cases where $m_r = 20, 10$. This can be seen in Figure 6.3(a),(b),(d) and (e) and in Figure 6.3(f). However in Figures 6.3(c), both CL and SL ended up with quite similar results.



(a) N1=2000, N0=3000,m_r=20 (b) N1=2000, N0=3000,m_r=10 (c) N1=2000, N0=3000,m_r=5



(d) N1=1000, N0=2000,m_r=20 (e) N1=1000, N0=2000,m_r=10 (f) N1=1000, N0=2000,m_r=5

Fig. 6.3: Performances of CL method and SL method on the case-control data.

6.4 Real data analysis

We applied the proposed CL method to the GWAS genotype datasets on coronary artery disease (CAD) and hypertension (HT) obtained by Affymetrix 500K SNP chips in the WTCCC study (WTCCC, 2007). The datasets are prepared in the same way we described in Section 3.5. As mentioned in the previous chapters, the whole genome result in 1983537 genotypes in CAD data and 2097111 genotypes in HT data. Applying the first stage of the method to the CAD and HT data led to 258086 potential risk haplotypes out of 1448586 in CAD data and 265804 potential risk haplotypes out of 1463838 in HT data. We then calculated the OR tests on these haplotypes at Stage 2. At Stage 2, according to the Bonferroni adjustment,

the individual significance level was set at the levels of $0.05/258086 = 1.93 \times 10^{-7}$ and $0.05/265804 = 1.88 \times 10^{-7}$ for the CAD and the HT respectively.

These individual significance levels were then used to determine the thresholding level c_1 in the multiple OR thresholding, which is $c_1 = 5$. Note that there were two sub-populations in controls. We applied further filtering on the regions to exclude the ones that have significant differences in the haplotypes frequencies within the two sub-control samples. The exclusion criterion was based on calculating chi-square p-value. Any region resulted in p-value less than 0.30 was excluded from the suspicious regions. This criterion was concluded from the simulated case-control samples when the risk factor d is less than 0.15 as we found out that the p-values for most of the 30 datasets are greater than 0.30, see Figure 3.6. Note that the declared risk haplotypes at the end of the second stage should also meet this criterion. Toward this end, the chi-square p-value of the frequencies in the two sub-control samples of each potential risk haplotype should be greater than 0.30. Otherwise the haplotype would be eliminated.

Finally, we calculated the ORs for all the estimated haplotypes and thresholded them by using the bound

$$\exp(c_1\sqrt{1/(n_{0H}+0.5)+1/(n_{1H}+0.5)+1/(n_{0\bar{r}}+0.5)+1/(n_{1\bar{r}}+0.5)})$$

with $c_1 = 5$. This gave the final risk-haplotype set as displayed in Tables 6.1, 6.2, 6.3 below. In the tables, each haplotype has been assigned to a physically closest gene on the basis of the information provided the GWAS catalog and the genetic information from the British 1958 Birth cohort. See Welter et al. (2014) and the web page at
http://www2.le.ac.uk/projects /birthcohort/1958bc. In the CAD case, we did rediscover the CAD risk gene CDKN2B and the risk haplotype “GGTGCCAG" found by the previous study (WTCCC, 2007; Zhu et al., 2010). We also tested the inheritance modes for these risk haplotypes in Chapter 3.

Tab. 6.1: The suspicious regions for coronary artery disease of WTCCC data detected by the CL method.

CAD								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
1	$rs10752991 - 183669960$	$SNP_A - 1976175$	$AAGCGGAC$	0.04276	0.02186	2.25033	7.9×10^{-10}	PLA2G4A
1	$241638336 - 241689464$	$rs4658439 - rs6428904$	$GGATCCGG$	0.01864	0.00735	4.00348	1.6×10^{-09}	LOC729761
2	$3789586 - 3821960$	$rs7576476 - rs12618184$	$GCTTACAG$	0.03451	0.01119	3.08154	4.3×10^{-14}	LOC442006
2	$49934439 - 50000082$	$rs6736617 - rs17039375$	$CCAAAGGT$	0.02347	0.00757	3.15172	1.5×10^{-10}	NRXN1
3	$73461569 - 73510299$	$rs7647311 - rs3845868$	$AGGCGCGG$	0.03876	0.01161	3.60719	3.8×10^{-17}	PDZRN3
3	$192463499 - 192526004$	$rs7644510 - rs293871$	$GACGCGTA$ $AGGGTGTT$	0.04375 0.03568	0.01075 0.02508	9.25792 3.25235	8.6×10^{-31} 7.2×10^{-11}	UTS2D
3	$197256495 - 197339533$	$rs6583286 - rs9834962$	$TAGACTTA$	0.0498	0.02364	2.18614	3.1×10^{-11}	TFRC
4	$3636361 - 3700212$	$rs10025237 - rs16844722$	$GGGGAGGG$	0.22491	0.15492	1.48435	8.7×10^{-10}	FLJ35424
4	$62159638 - 62224814$	$rs335339 - rs17090501$	$AGTAGCGC$	0.61988	0.56353	1.26297	5.2×10^{-08}	LPHN3
4	$167440772 - 167457521$	$rs9995087 - rs17047336$	$GGACGCAG$	0.03434	0.01139	3.4204	9.6×10^{-13}	TLL1
4	$180659963 - 180699763$	$rs6811556 - rs17090633$	$CCCCCACT$	0.01782	0.00755	3.9912	5.4×10^{-11}	LOC391719
5	$166764561 - 166801933$	$rs6863935 - rs7724862$	$CTATGTGT$	0.09145	0.05448	1.69398	8.7×10^{-09}	ODZ2
7	$42931717 - 42940671$	$rs2024125 - rs2330742$	$AGTGTAGA$	0.09745	0.0513	1.98174	2.5×10^{-15}	HECW1
7	$121502348 - 121570029$	$rs284378 - rs1443751$	$CAGGGTCT$	0.03082	0.01538	2.37983	2.0×10^{-09}	CADPS2
7	$153371858 - 153449397$	$rs6464391 - rs1861139$	$CGGGTAGA$	0.04119	0.02159	2.02259	1.6×10^{-08}	LOC653748
8	$71022178 - 71086937$	$rs7836791 - rs388511$	$TACAGAAG$	0.02204	0.00555	5.35918	1.5×10^{-07}	SLCO5A1
9	$22088619 - 22120515$	$rs2891168 - rs10965245$	$GGTGCCAG$	0.34939	0.29298	1.363	2.4×10^{-11}	CDKN2B
9	$74180343 - 74241329$	$rs10114124 - rs17081046$	$GTATTTAT$	0.21608	0.13046	1.66562	4.0×10^{-17}	RORB
9	$119506057 - 119537035$	$rs2191675 - rs10984648$	$GTTGGCTA$ $ATCGACTA$	0.08762 0.06297	0.03361 0.04219	4.37262 2.50612	6.8×10^{-51} 1.8×10^{-19}	CDK5RAP2
10	$11879196 - 11924252$	$rs6602535 - rs11257355$	$TTTGTCGG$	0.04949	0.0145	4.39053	1.5×10^{-10}	C10orf47
10	$64409674 - 64442476$	$rs1509952 - rs2842286$	$TTTCTTAC$	0.02299	0.0073	5.79388	2.5×10^{-12}	NRBF2
11	$8165969 - 8200374$	$rs4758310 - rs11041816$	$ATAATGGG$	0.36298	0.3164	1.3306	1.1×10^{-08}	LOC644497
11	$11959140 - 11969776$	$rs17464087 - rs10741565$	$GTGGTCGT$ $CCGGTCGC$	0.16348 0.10795	0.13543 0.07913	1.41928 1.60397	1.3×10^{-08} 1.8×10^{-10}	DKK3
11	$129102667 - 129124330$	$rs532427 - rs691197$	$ACCGCGGA$	0.08519	0.05612	1.99764	1.6×10^{-07}	TMEM45B
11	$133079508 - 133113640$	$rs4937817 - rs4937826$	$GTAGTGCC$	0.04216	0.02425	2.44201	3.9×10^{-08}	LOC646522
12	$5619429 - 5628923$	$rs11063791 - rs454704$	$TACATAAA$	0.02897	0.0124	2.54881	7.2×10^{-10}	TMEM16B
12	$129086441 - 129129809$	$rs713149 - rs1027557$	$AAAGCGGT$	0.18839	0.11206	1.90461	6.3×10^{-11}	FLJ31485
13	$23708179 - 23726596$	$rs881428 - rs2760374$	$AGAAGTTT$	0.11908	0.07922	1.47796	1.2×10^{-07}	SPATA13
13	$108372995 - 108432811$	$rs4773010 - rs3842945$	$AGAGACCC$	0.27486	0.19222	1.59284	1.2×10^{-21}	MYO16
14	$25140850 - 25159405$	$rs8020556 - rs1951062$	$AGTAAACT$ $AGTACATA$ $GCTACATA$	0.09084 0.24934 0.04608	0.02999 0.2259 0.01682	3.81326 1.39192 3.44683	2.3×10^{-42} 3.4×10^{-08} 3.9×10^{-22}	LOC401767
14	$53221435 - 53244046$	$rs1563719 - rs210351$	$AGATAGGT$	0.15385	0.10566	1.48023	2.3×10^{-09}	BMP4
14	$65343491 - 65401760$	$rs3924222 - rs12896836$	$TACGTCTT$ $TATAACTC$	0.05544 0.0462	0.03704 0.01904	1.89769 3.07126	1.2×10^{-08} 1.1×10^{-17}	FUT8
15	$20624103 - 21246055$	$rs7166056 - rs8024346$	$GTGACGTG$	0.08093	0.04109	2.10848	2.4×10^{-13}	NIPA1
15	$37962389 - 38014169$	$rs11633436 - rs534757$	$TTACAACC$	0.07798	0.03763	2.34448	5.0×10^{-12}	GPR176
16	$55207138 - 55253047$	$rs8055724 - rs12447986$	$TTCTCCTC$	0.03044	0.01113	2.81792	2.8×10^{-11}	MT1L
16	$63792132 - 63847234$	$rs1862709 - rs1423798$	$CGGAACCA$ $CGGATACT$ $CGGATACA$ $CGAAAATA$	0.40562 0.21037 0.08934 0.05187	0.36993 0.19685 0.08462 0.05175	4.10819 4.00571 3.96191 3.76755	2.1×10^{-12} 3.1×10^{-11} 1.4×10^{-09} 8.0×10^{-08}	LOC283867

Tab. 6.2: The suspicious regions for coronary artery disease of WTCCC data detected by the CL method.

CAD								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
17	$27921023 - 27963104$	$rs225215 - rs17780520$	$GGGTTAAC$	0.0205	0.00465	4.41157	6.8×10^{-10}	MYO1D
17	$74629176 - 74682195$	$rs2612793 - rs8072667$	$CGAGGTTG$	0.06276	0.03471	2.03532	1.9×10^{-09}	FLJ21865
18	$8212591 - 8279839$	$rs10468776 - rs11876033$	$GGGACAAG$	0.02689	0.00982	2.62558	3.7×10^{-09}	PTPRM
18	$2291328 - 22715430$	$rs3974646 - SNP_A$	$TGCGGAGT$	0.05382	0.02751	2.23011	3.8×10^{-12}	AQP4
19	$4625799 - 4746342$	$rs11670570 - rs1044409$	$AGCAACCG$	0.05419	0.02332	2.70861	5.4×10^{-17}	DPP9
19	$6641966 - 6717213$	$rs3745566 - rs7248911$	$TAAGCTAC$	0.02312	0.00521	4.75605	1.8×10^{-14}	C3
19	$17595848 - 17649789$	$rs10419511 - rs7252308$	$TTGGTATG$	0.04657	0.01971	3.86284	3.4×10^{-09}	UNC13A
19	$52946204 - 53026777$	$rs10402957 - rs4427918$	$CATTCAGC$	0.0741	0.04321	1.78028	3.2×10^{-10}	GLTSCR2

Tab. 6.3: The suspicious regions for hypertension of WTCCC data detected by the CL method.

HT								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
1	227569611 – 227620956	$rs7514972 - rs9431663$	$CGTATAGG$	0.03377	0.00926	4.68662	9.6×10^{-13}	TRIM67
1	236986859 – 237020204	$rs12137158 - rs16840310$	$ATTTAGGG$	0.08733	0.05437	1.97004	1.2×10^{-14}	GREM2
4	3700382 – 3734797	$rs177772 - rs12641338$	$TACCGATT$	0.12978	0.08988	1.85844	1.8×10^{-12}	FLJ35424
6	134839318 – 134940216	$rs17063800 - rs228439$	$GAGGAGTT$	0.04494	0.03434	2.29235	3.9×10^{-08}	HBS1L
6	139560239 – 139612833	$rs7765885 - rs9495394$	$GCGCAACG$	0.0487	0.01774	2.82143	1.6×10^{-17}	HECA
6	139693238 – 139758634	$rs11155050 - rs9373237$	$TTGCGGCT$	0.01924	0.00686	3.63934	8.1×10^{-10}	TXLNB
7	77695246 – 77717237	$rs2215379 - rs4515471$	$TCTAAAAA$ $CTTGGAAA$	0.02943 0.02094	0.01786 0.01061	2.39311 2.86264	6.3×10^{-08} 2.6×10^{-08}	MAGI2
11	69213458 – 69295251	$rs1192923 - rs3168175$	$TTGTGGCA$	0.05532	0.02803	2.07485	6.9×10^{-10}	FGF4
11	125683058 – 125763272	$rs2096915 - rs7118117$	$CACACGAG$	0.07736	0.04727	1.73075	1.2×10^{-08}	DCPS
12	19808672 – 19824536	$rs10841340 - rs10770543$	$GTTAATTC$	0.06671	0.02861	2.71722	1.4×10^{-16}	LOC400013
13	23708179 – 23726596	$rs881428 - rs2760374$	$GAAAGCTT$ $AGAAGTTT$	0.2454 0.12142	0.19993 0.07922	1.36157 1.69995	5.1×10^{-09} 1.1×10^{-13}	SPATA13
14	21674996 – 21704333	$rs12050442 - rs1894369$	$GGGGTTAC$	0.03075	0.00968	3.21035	2.3×10^{-12}	TRA@
14	25140850 – 25159405	$rs8020556 - rs1951062$	$AGTAAACT$	0.08475	0.02999	2.82991	1.3×10^{-26}	LOC401767
14	36969639 – 37032855	$rs10132119 - rs17106785$	$CTATGACA$	0.01914	0.00402	4.70209	2.3×10^{-09}	MIPOL1
16	76334152 – 76353949	$rs11646710 - rs9935394$	$CGGTGGGC$	0.01902	0.00855	3.2779	2.4×10^{-09}	LOC729777
17	6992193 – 7158208	$rs4558460 - rs6503013$	$TCGCGTCG$	0.14256	0.10161	1.49923	1.0×10^{-09}	LLGL1
18	2291328 – 22715430	$rs3974646 - SNP_A$	$TGCGGAGT$ $TGTAATGT$	0.05186 0.22261	0.02751 0.21193	3.30581 1.84357	2.0×10^{-18} 3.8×10^{-10}	LOC440489

6.5 Discussion and conclusion

We proposed a clustering-based logistic regression approach to detect disease-risk haplotypes. In this approach, we started with fitting the model to the genotypes. The independent variable was scored by the centroids of K-means clusters of the log ratios $d_j = (\log P(G_j|\text{cases})/P(G_j|\text{controls})), 1 \leq j \leq J$. The importance of K-means here is to collapse the genotypes according to their d_j to minimize the number of independent variable scores. This approach could overcome the limitations of fitting the standard multiple logistic regressions by considering a number of independent variables less than the number of the genotypes in the sample. We applied this approach to the genotype data of the two study design. Compared to the standard method, our approach was quite similar in it's performance to SL method in all cohort cases except the one where the $m_r = 5$ with underlying dominant mode of inheritanc. In contrast, in the case-control design our method CL showed some advantages over SL method in detecting risk haplotypes in most scenarios.

7. DISCUSSIONS, CONCLUSIONS AND FUTURE WORKS

In this Section, we will give discussions on the pros and cons of the proposed methods compared to the existing methods, make conclusions for the thesis, and point out future works.

7.1 Overview of the results of our methods

In this thesis, we have developed four novel methods to detect risk haplotypes for a disease. These methods have been assessed by applying them to both simulated datasets with a wide range of scenarios and WTCCC data on CAD and HT.

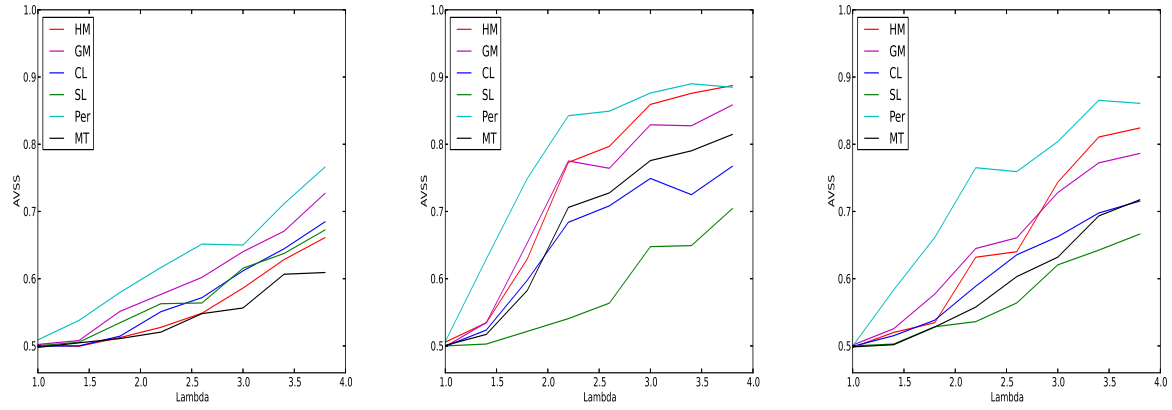
In the simulations, we generated datasets based on retrospective and prospective designs. We aimed to detect the risk haplotypes which we pretended that we did not know. We compared the proposed methods to the multiple testing method(MT) or the standard multiple regression method in terms of AVSS. Note that in the target population, disease prevalences were distributed on the basis of underlying modes of inheritance which were unknown in the population. So, we opted for *population-based designs*, where selecting a sample from the target population could be done by using the so-called cohort design and case-control design, for details see Section 2.8. In our simulations, we generated the data based on each design separately. The cohort design was easy for developing tests for mode of inheritance and for mathematical formulas derivations, whereas the case-control design was easy for finding the minimum level of the P-value between the frequencies of haplotypes between cases and controls when there was no risk in the sample.

Zhu et al. (2010) proposed multiple testing method to detect disease-risk haplotypes. In the application of their method to the simulation, they generated the cases' genotypes based on different modes of inheritance which were similar to our cohort design. In their paper, it was mentioned that the numbers of simulated cases and controls were 1900, 3000, respectively. The key requirement in their co-classification stage was that the number of cases and the number of controls were randomly chosen

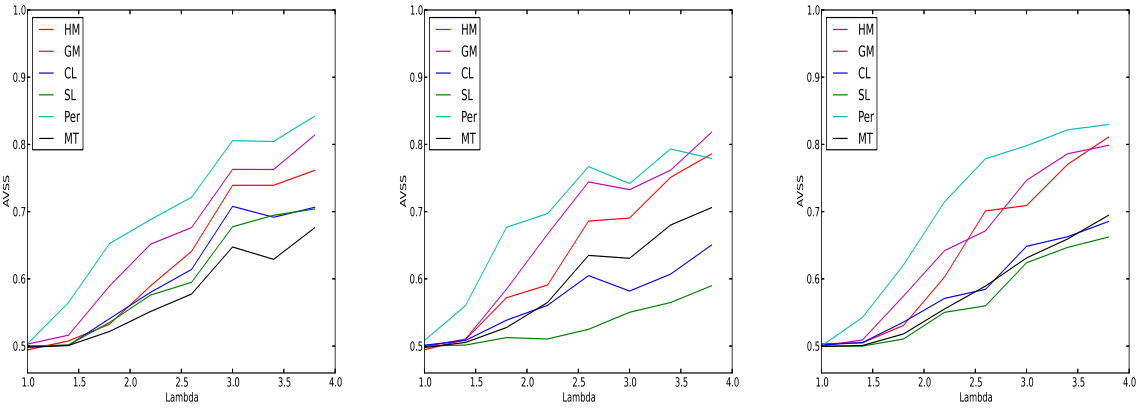
for co-classification. This was used to reduce the effects of population-substructures in the sample.

To compare the proposed methods to each other, we plotted their simulated AVSS in the various scenarios considered before. These plots were displayed in Figures 7.1, 7.2 and 7.3. It can be seen in Figures 7.1 and 7.2 that the methods HM, GM and Per outperformed the MT, CL and SL in all scenarios that we considered in the cohort design and in the case-control design in terms of their AVSS values. The permutation method had the best performance among all the six methods in the cohort design. The methods HM and GM were close to each other in their AVSS values in the cohort design, although both of them outperformed the MT. The standard multiple logistic method (SL) performed the worst in the cohort design when the sample size was 5000 and the mode of inheritance was either dominant or multiplicative as shown in Figure 7.1 (b), (c), (e), (f), (h) and (i). However, in the recessive model, it performed better than MT as demonstrated in Figure 7.1 (a), (d) and (g). The CL method showed a better result than did the MT and the SL in detecting the risk haplotypes in some scenarios. See Figures 7.1 (a), (c) and (g) and all scenarios shown in Figure 7.2 except that in (i).

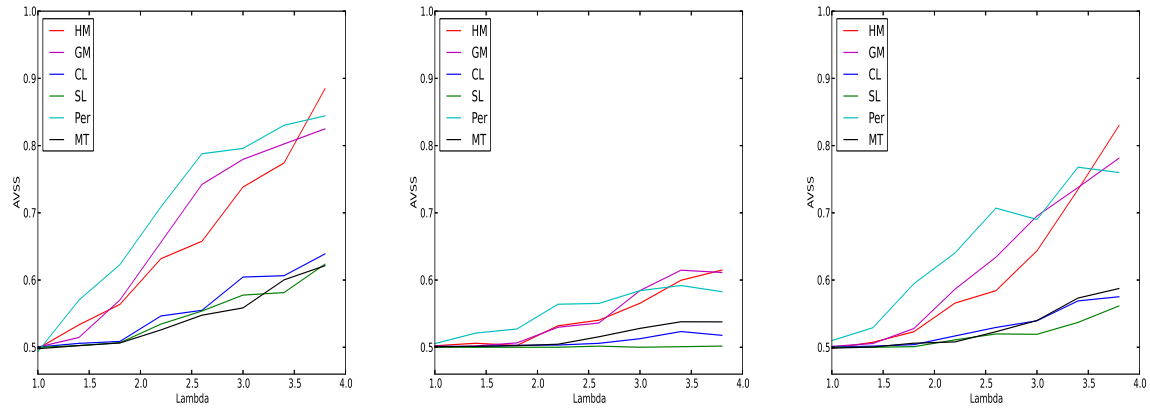
The GM and HM outperformed the rest methods in case-control design as shown in Figure 7.3. The SL method was the worst in its performance in this design. The Per and CL methods performed better than MT. However, the MT method overtook the CL method in case-control design when $d > 0.25$, where d was used to show the risk difference between risk and non-risk genotypes in the simulations. As it can be also seen in Figure 7.3, the HM method overtook the GM and Per methods when $d > 0.15$



(a) N=5000, m_r=5, rece (b) N=5000,m_r=5,dom (c) N=5000,m_r=5,mult

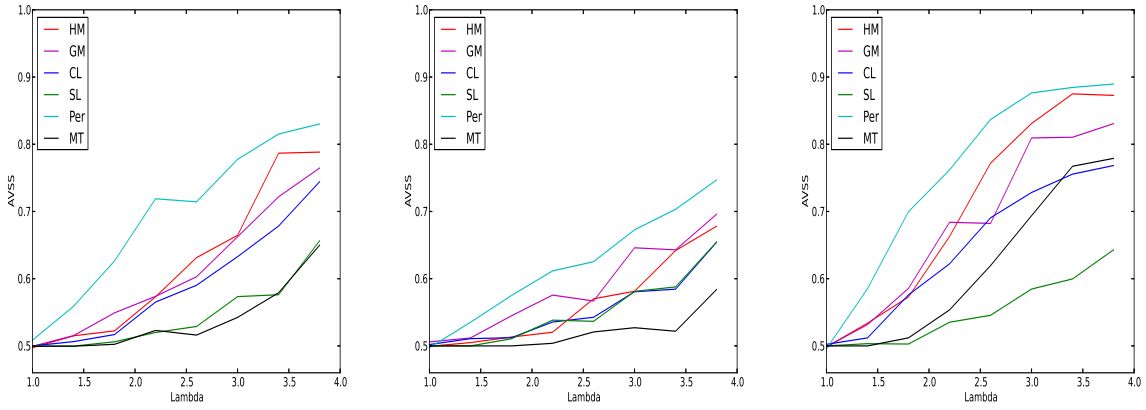


(d) N=5000,m_r=10,rece (e) N=5000,m_r=10,dom (f) N=5000,m_r=10,mult

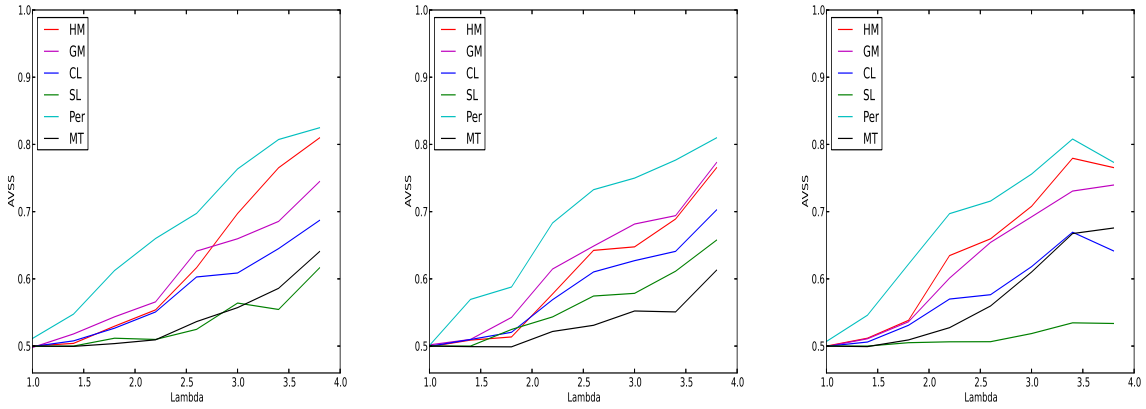


(g) N=5000,m_r=20,rece (h) N=5000,m_r=20,dom (i) N=5000,m_r=20,mult

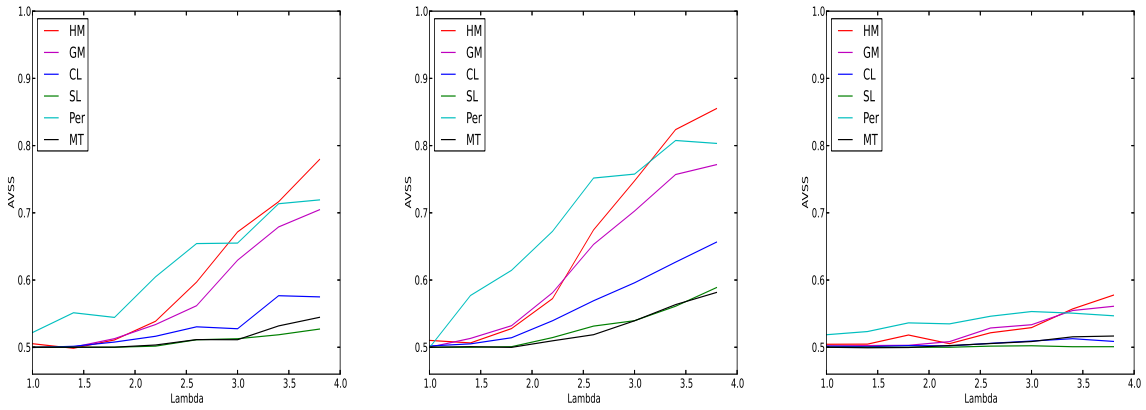
Fig. 7.1: Performances of all methods on the cohort-design data with multiplicative or dominant or recessive inheritance modes based on sample sizes of 5000. The curves show the averages of the AVSS values over 30 replicates in each scenario for the methods HM, GM, MT, Per, CL, and SL.



(a) N=3000,m_r=5,mult (b) N=3000,m_r=5,rece (c) N=3000,m_r=5,dom

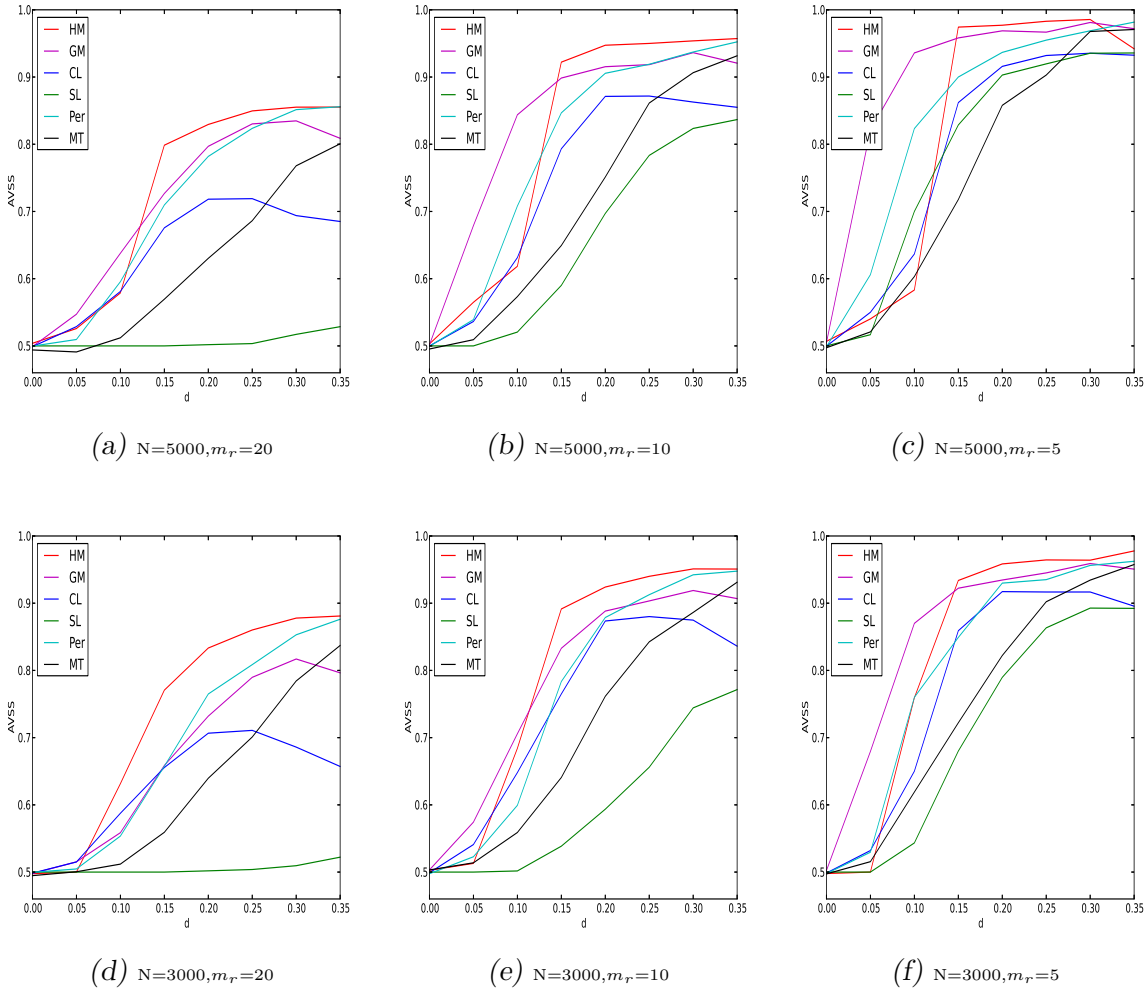


(d) N=3000,m_r=10,mult (e) N=3000,m_r=10,rece (f) N=3000,m_r=10,dom



(g) N=3000,m_r=20,mult (h) N=3000,m_r=20,rece (i) N=3000,m_r=20,dom

Fig. 7.2: Performances of all methods on the cohort-design data with multiplicative or dominant or recessive inheritance modes based on sample sizes of 3000. The curves show the averages of the AVSS values over 30 replicates in each scenario for the methods HM, GM, MT, Per, CL, and SL.



(a) N=5000,m_r=20 (b) N=5000,m_r=10 (c) N=5000,m_r=5

(d) N=3000,m_r=20 (e) N=3000,m_r=10 (f) N=3000,m_r=5

Fig. 7.3: Performances of all methods on the case-control data based on sample size 5000 or 3000. The curves show the averages of the AVSS values over 30 replicates in each scenario for the methods HM, GM, MT, Per, CL, and SL.

In summary, in the cohort design, the permutation method performed better than all other methods. This showed the benefit of considering reconstructing the haplotypes/genotypes of the case and control samples altogether to allow PHASE to stratify the sample. This of course would affect the penetrance probabilities. But it had no effects on estimating genotype frequencies in cases. In this design, we can also see that the HM and GM overtook each other when the genotype relative risk λ was increasing. Their performances were better than those of the MT, CL and SL methods. Both the MT and CL outperformed the SL, although there were no significant differences between the performances of the MT and CL. Note that the results SL was improved by using the idea of collapsing the rare genotypes as we mentioned in Section 6.2.2. However, its performance was the worst amongst the

six methods.

In the case-control design, the GM outperformed the others in all scenarios when the risk was low in the sense that $d < 0.15$. When $d > 0.15$, the HM outperformed all the others as shown in Figure 7.3. In this design, the SL method performed the worst in all scenarios except the one with the sample size 5000 and the number of risk haplotypes $m_r = 5$.

As far as real data analyses are concerned, we summarised the results of the four proposed methods in Table 7.1 and Table 7.2. It can be seen many genes have been detected as potential risk for coronary artery disease or hypertension. The differences in detecting these genes by the proposed methods could due to the difference in their performance. Take HM as an example, at the first stage of this method, we clustered the haplotypes into two groups. At the second stage, we used Bonferroni correction to adjust the significant level based on the number of the haplotypes resulted from the first stage in the whole genome. This number might be much less than the total number of the haplotypes in the whole genome. In doing so, we reduced the impact of a multiple testing problem that can inflate type II error. In the contrast, in the permutation method, we used Bonferroni correction to adjust the significant level for detecting the risk genotypes at the first stage. This resulted in detecting less haplotypes than the HM at the final stage due to the strict significant level we used. This issue would not be a problem in the simulation as the significant level we used is not very strict.

The most important finding is the genes that have been detected by more than two methods. One of these genes is CDKN2B. This gene has been detected by all the proposed methods. It has already been reported in a GWAS catalog as a risk factor for coronary artery disease. More importantly, the remaining genes need to be investigated by researchers who are specialist in Biochemistry or Medical genetics to find the biological association between them and the reported diseases in terms of their biological functions. In fact, majority of these genes have already been reported as risk factors for some diseases according to the GWAS catalog, but based on data provided by different projects worldwide.

Tab. 7.1: Potential risk genes for coronary artery disease and detection methods

Chr	Gene	Detection method	Chr	Gene	Detection method	Chr	Gene	Detection method
1	LOC284577	HM, Per	1	RGS7	HM	1	ACTL8	GM
1	MSH4	GM	1	hCG-2036596	Per	1	LOC728431	Per
1	RHOU	Per	1	TRIM67	Per	1	CEP170	Per
1	AKT3	Per	1	PLA2G4A	CL	1	LOC729761	CL
2	NRXN1	GM, Per, CL	2	LOC442021	GM	2	LOC402120	GM
2	DNER	GM	2	FLJ43879	GM	2	PPP1R7	GM
2	LOC442006	Per, CL	2	LOC442006	Per	2	LOC442021	Per
3	ABI3BP	HM	3	ACPL2	HM	3	PLSCR5	HM
3	BHLHB2	GM	3	SLC6A6	GM, Per	3	FHIT	GM
3	SYNPR	GM	3	SUCLG2	GM	3	NFKBIZ	GM
3	ALCAM	GM	3	SLC9A9	GM	3	C3orf58	GM
3	TNIK	GM	3	UTS2D	GM, CL	3	TFRC	GM, Per, CL
3	CNTN4	Per	3	PDZRN3	Per, CL	4	LOC654254	HM
4	ZNF509	HM	4	LOC391719	HM, CL	4	FLJ35424	GM, Per, CL
4	TLL1	Per, CL	4	LPHN3	CL	5	CLINT1	HM, Per
5	LOC728682	GM	5	ODZ2	GM, Per, CL	5	LOC644659	Per
7	MAGI2	HM	7	LOC647030	HM	7	RBAK	GM
7	LOC340268	GM	7	AAA1	GM	7	HECW1	GM, CL
7	LOC653748	GM	7	MAGI2	Per	7	LOC653748	Per, CL
7	CADPS2	CL	8	BAALC	HM, GM	8	LOC648237	GM
8	PDGFRL	GM	8	WHSC1L1	GM	8	SLCO5A1	Per, CL
9	CDKN2B	HM, GM, Per, CL	9	GNA14	HM, Per	9	RAPGEF1	HM
9	BNC2	GM	9	RORB	GM, Per, CL	9	TNFSF8	GM
9	CDK5RAP2	GM, Per, CL	9	OLFM1	Per	10	NRBF2	HM, GM, CL
10	RBM20	HM	10	MKI67	HM	10	C10orf47	GM, Per, CL
10	FAM107B	Per	11	FLJ14213	HM, Per	11	LOC646522	HM, Per, CL
11	IQSEC3	HM	11	LOC644497	GM, Per, CL	11	NELL1	GM
11	FGF4	GM, Per	11	CCDC90B	GM, Per	11	CNTN5	Per
11	DRD2	Per	11	TMEM45B	Per, CL	11	DKK3	CL
12	SOX5	HM	12	KRT3	HM	12	TMEM132C	HM
12	BTG1	GM	12	TBX3	GM	12	TMEM132C	GM
12	FLJ31485	GM, CL	12	TMEM16B	Per, CL	12	RBM19	Per
12	NOS1	Per	13	MTIF3	GM	13	LOC196549	GM
13	RCBTB2	GM	13	MYO16	Per, CL	13	SPATA13	CL
14	LOC401767	GM, Per, CL	14	NPAS3	GM	14	FUT8	GM, Per, CL
14	BMP4	Per, CL	15	GPR176	HM, GM, Per, CL	15	NIPA1	GM
15	NIPA1	GM, Per, CL	15	MAGEL2	GM	15	LCTL	GM
15	C15orf26	GM	15	SLCO3A1	GM	15	RGMA	GM
15	MAGEL2	Per	16	BCMO1	HM, Per	16	A2BP1	GM
16	LONP2	GM	16	LOC643714	GM	16	MT1L	GM, Per, CL
16	LOC283867	CL	17	LOC646202	HM	17	MSI2	HM
17	TNRC6C	GM	17	FLJ21865	GM, Per, CL	17	MYO1D	Per, CL
18	RAB31	HM	18	C18orf20	HM, GM	18	FLJ44313	HM
18	PTPRM	GM, Per, CL	18	KIAA0802	GM	18	AQP4	Per, CL
18	MOCOS	Per	19	DPP9	HM, CL	19	KLK2	HM
19	VN1R4	HM	19	PVRL2	GM	19	GLTSCR2	GM, Per, CL
19	CACNG7	GM	19	C3	Per, CL	19	AKAP8L	Per
19	UNC13A	Per, CL	19	LOC729966	Per	20	FLJ33544	GM
20	HNF4A	GM	20	SLC13A3	GM	20	SALL4	GM
20	ZNF217	GM	20	PHACTR3	GM	20	C20orf196	Per
20	C20orf42	Per	20	BCAS1	Per	20	PCK1	Per
21	TPTE	GM	22	CACNG2	HM, GM, Per	22	LOC729269	GM, Per
22	SYN3	GM						

Tab. 7.2: Potential risk genes for hypertension and detection methods

Chr	Gene	Detection method	Chr	Gene	Detection method	Chr	Gene	Detection method
1	TRIM67	HM, CL	1	TSNAX	HM	1	GREM2	HM, GM, CL
2	SOS1	Per	3	TOMM70A	HM	3	ACPL2	HM
3	LOC646730	HM	4	KCNIP4	HM	4	PPARGC1A	HM
4	FLJ35424	GM, CL	4	PALLD	GM	4	LCORL	Per
5	LOC651746	HM	5	PDZD2	HM	6	HECA	HM, CL
6	TXLNB	HM, CL	6	SYNE1	GM	6	QRSL1	Per
6	HBS1L	CL	7	ABCA13	HM	7	MAGI2	HM, CL
9	GNA14	HM	10	RAB11FIP2	HM	10	LOC645954	Per
11	DCPS	HM, CL	11	FGF4	GM, CL	11	OR5D14	Per
11	SORL1	Per	12	LOC729222	HM	12	RBM19	HM
12	TBX3	HM	12	NOS1	GM	12	LOC400013	CL
13	SPATA13	HM, CL	13	ATXN8OS	HM	14	TRA@	HM, GM, CL
14	SLC25A21	HM, GM	14	LOC401767	GM, CL	14	MIPOL1	GM, CL
16	PPL	HM	16	XYLT1	HM	16	LOC729777	CL
17	RPL38	HM	17	LLGL1	CL	18	LOC728864	Per
18	LOC440489	CL	19	NFIC	HM	19	ZNF414	HM
19	UNC13A	HM, GM	19	CHST8	HM	20	ANKRD5	HM

7.2 Future work

In this thesis, we have addressed several issues on the GWAS data analysis that challenge the existing statistical methods. We considered the simple setting where we put the non disease-risk haplotypes and the disease-protective haplotypes as one category. From statistical point of view, if protective haplotypes existed in the population, they would decrease the probability of getting the disease. The proposed methods can be extended to identify the risk-protective haplotypes by increasing the number of mixture components in the models.

BIBLIOGRAPHY

[1] Balding, D.J., Bishop, M. and Cannings, C. (2007). *Handbook of Statistical Genetics*. John Wiley and Sons, Ltd.

[2] Binder, H., Müller, T., Schwender, H., Golka, K., Steffens, M., Hengstler, J. G., Ickstadt, K. and Schumacher, M. (2012). Cluster-localized sparse logistic regression for SNP data. *Statistical Applications in Genetics and Molecular Biology*, **11(4)**.

[3] Breslow, N. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, **91(433)**, 14-28.

[4] Breslow, N. and Zhao, L. (1988). Logistic-Regression for Stratified Case Control Studies. *Biometrics*, **44(3)**, 891-899.

[5] Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084-1097

[6] Burton, P. R., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447(7145)**, 661-678.

[7] Byng, M. C., Whittaker, J. C., Cuthbert, A. P., Mathew, C. G. and Lewis, C. M. (2003). SNP subset selection for genetic association studies. *Annals of Human Genetics*, **67(6)**, 543-556.

[8] Chambless, L. E. and Boyle, K. E. (1985). Maximum-Likelihood Methods for Complex Sample Data - Logistic-Regression and Discrete Proportional Hazards Models. *Communications in Statistics-Theory and Methods*, **14(6)**, 1377-1392.

[9] Chen, H., Zhu, X., Zhao, H. and Zhang, S. (2003). Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Annals of Human Genetics*, **67**, 250-264.

[10] Clark, A. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology*, **27(4)**, 321-333.

[11] David, W., Hosmer, Jr., Stanley L. (2004). *Applied logistic regression*. John Wiley & Sons.

[12] Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55(4)**, 997-1004.

[13] Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S. and Satten, G. A. (2012). A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *American Journal of Human Genetics*, **91(2)**, 215-223.

[14] Excoffier, L. and Slatkin, M. (1995). Maximum-Likelihood-Estimation of Molecular Haplotype Frequencies in a Diploid Population. *Molecular Biology and Evolution*, **12(5)**, 921-927.

[15] Feng, T., Elston, R. C. and Zhu, X. (2011). Detecting Rare and Common Variants for Complex Traits: Sibpair and Odds Ratio Weighted Sum Statistics (SPWSS, ORWSS). *Genetic Epidemiology*, **35(5)**, 398-409.

[16] Finos, L. and Salmaso, L. (2004). Nonparametric multi-focus analysis for categorical variables. *Communications in Statistics - Theory and Methods*, **33(8)**, 1931-1941.

[17] Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer series in statistics, New York.

[18] Furihata, S., Ito, T. and Kamatani, N. (2006). Test of association between haplotypes and phenotypes in case-control studies: Examination of validity of the application of an algorithm for samples from cohort or clinical trials to case-control samples using simulated and real data. *Genetics*, **174(3)**, 1505-1516.

[19] Greevenbroek, V. M., Zhang, J., Kallen, C. J., Schiffers, P., Feskens, E. and de Bruin, T. (2008). Effects of interacting networks of cardiovascular risk genes on the risk of type 2 diabetes mellitus (the CODAM study). *BMC Medical Genetics*, **9(1)**, Article 36.

[20] Hartl, D. L. and Clark, A. G. (1997). *Principles of Population Genetics*. 3rd Edition. Sinauer Associates Inc., Sunderland, MA, USA.

[21] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106(23)**, 9362-9367.

[22] Huang, Y. H., Lee, M. H., Chen, W. J. and Hsiao, C. K (2011). Using an Uncertainty-Coding matrix in bayesian regression models for Haplotype-Specific risk detection in family association studies. *Plos One*, **6(7)**.

[23] Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, **18**, 337-8.

[24] Igo, R. P., Jr., Li, J. and Goddard, K. A. B. (2009). Association Mapping by Generalized Linear Regression With Density-Based Haplotype Clustering. *Genetic Epidemiology*, **33(1)**, 16-26.

[25] Kang, G., Yue, W., Zhang, J., Huebner, M., Zhang, H., Ruan, Y., Lu, T., Ling, Y., Zuo, Y. and Zhang, D., (2008). Two-stage designs to identify the effects of SNP combinations on complex diseases. *Journal of Human Genetics*, **53(8)**, 739-746.

[26] Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Comput. Stat. & Data Ana.* , **41**, 577-590.

[27] Kwok, P. Y. (2003). *Single Nucleotide Polymorphisms*. Humana Press Inc., Totowa, NJ.

[28] Li, Y., Byrnes, A. E. and Li, M. (2010). To Identify Associations with Rare Variants, Just WHaIT Weighted Haplotype and Imputation-Based Tests. *American Journal of Human Genetics*, **87(5)**, 728-735.

[29] Li, M., Ye, C., Fu, W., Elston, R. C. and Lu, Q. (2011). Detecting Genetic Interactions for Quantitative Traits with U-Statistics. *Genet. Epidemiol.*, **35**, 457-468.

[30] Manly, B. F. J. (2007). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. (3rd ed.), London: Chapman Hall.

[31] McLachlan, G.J. and Basford, K.E. (1988). *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York.

[32] McLachlan, G. J. and Krishnan, T., (2008). *The EM Algorithm and Extensions*. Wiley, New York.

[33] McLachlan, G. J. and Peel, D., (2000a). *Finite mixture models*. Wiley, New York.

[34] Molitor, J., Marjoram, P. and Thomas, D. (2003). Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. *Genetic Epidemiology*, **25(2)**, 95-105.

[35] Molitor, J., Marjoram, P. and Thomas, D. (2003). Fine-Scale Mapping of Disease Genes with Multiple Mutations via Spatial Clustering Techniques. *American Journal of Human Genetics*, **73(6)**, 1368-1384.

[36] Morris, A. P. (2006). A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *American Journal of Human Genetics*, **79(4)**, 679-694.

[37] Morris, A. P. (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genetic Epidemiology*, **29(2)**, 91-107.

[38] Nicholas P. (2003). *Statistics for Epidemiology*. Chapman & Hall/CRC Press.

[39] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika*, **66**, 403–411.

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

[41] Robinson, R. (2010) Common disease, multiple rare (and distant) variants. *PLoS Biol* **8**, e1000293. doi:10.1371/journal.pbio.1000293.

[42] Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. and Poland, GA. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **70**, 425-434.

[43] Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78(4)**, 629-644.

[44] Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68(4)**, 978-989.

[45] Stranger, B. E., Stahl, E. A. and Raj, T. (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics*, **187(2)**, 367-383.

[46] Templeton, A. R., Boerwinkle, E. and Sing, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*, **117(2)**, 343-351.

[47] Tzeng, J. Y., Wang, C. H., Kao, J. T. and Hsiao, C. K. (2006). Regression-based association analysis with clustered haplotypes through use of genotypes. *Ameri. Jour. Hum. Genet.*, **78**, 231–42.

[48] Weir, B. S. (1996). *Genetic data analysis II*. Sinauer, Sunderland, MA.

[49] Welter D., MacArthur J., Morales J., Burdett T., Hall P., Junkins H., Klemm A., Flicek P., Manolio T., Hindorff L. and Parkinson H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, **42** (Database issue): D1001-D1006.

[50] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). Model-based clustering and data transformation for gene expression data. *Bioinformatics*, **17**, 977-987.

[51] Zakharov, S., Wong, T. Y., Aung, T., Vithana, E. N., Khor, C. C., Salim, A. and Thalamuthu, A. (2013). Combined genotype and haplotype tests for region-based association studies. *BMC Genomics*. **14**, 569.

[52] Zhang, J., Vingron, M. and Hoehe, M. R. (2005). Haplotype Reconstruction for Diploid Population. *Human Heredity*, **59**, 144-156.

[53] Zhang, J., Liang, F., Dassen, W. R., Veldman, B. A., Doevendans, P. A. and De Gunst, M. (2003) Search for haplotype interactions that influence susceptibility to type 1 diabetes, through use of unphased genotype data. *Am J Hum Genet*. **73**, 1385–401.

[54] Zhao, J. H., Curtis, D. and Sham, P. C. (2000). Model-free analysis and permutation tests for allelic associations. *Human Heredity*, **50(2)**, 133-139.

[55] Zhu, X., Feng, T., Li, Y., Lu, Q. and Elston, R.C. (2010). Detecting rare variants for complex traits using family and unrelated data. *Genet. Epidemiol.*, **34**, 171-187.

[56] Ziegler, A., König, I. R. and Pahlke, F. (2010). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. 2nd Edition, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.