Stony Brook University

Multi-Marker Linkage Disequilibrium Mapping of Quantitative Trait Loci

A Dissertation Presented

by

Soyoun Lee

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration – Statistics)

Stony Brook University

August 2015

Stony Brook University

The Graduate School

Soyoun Lee

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Song Wu – Dissertation Advisor
Assistant professor, Department of Applied Mathematics and Statistics

Pei Fen Kuan - Chairperson of Defense
Assistant professor, Department of Applied Mathematics and Statistics

Wei Hou – Dissertation Co-Advisor
Assistant professor, Department of Preventive Medicine

Leslie Hyman
Professor, Department of Preventive Medicine

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Multi-Marker Linkage Disequilibrium Mapping of Quantitative Trait Loci

by

Soyoun Lee

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration – Statistics)

Stony Brook University

2015

Background: Genome-wide associations have been studied for years to infer relationships between biological traits and their underlying genetic causes, known as quantitative traits loci (QTL), in a process known as QTL mapping. Single nucleotide polymorphisms (SNPs), which are commonly used genetic markers, are usually in linkage disequilibrium (LD) with each other within a small genetic region. Both single- and double-marker-based LD mapping methods have been developed by taking advantage of LD structure.

Method: In this thesis, a more general LD mapping framework with an arbitrary number of markers has been developed in order to improve LD mapping and its detection power. This method is referred to as multi-marker linkage disequilibrium mapping (mmLD). For the mmLD model estimation, novel two-phase estimation procedures were implemented. In the first phase, haplotype frequencies were estimated for known markers. In the second phase, haplotype frequencies, including the unknown QTL, were updated based on estimates from the first step.

To test hypotheses, we used the likelihood ratio test (LRT). We propose a sequential LRT method that compares likelihood values of a reduced and an alternative model, and determines the optimal degree of freedom for testing.

Results: To compare our method with other existing mapping methods, such as the single-marker LD mapping and the SKAT_C, we performed intensive simulations studies. These studies showed that our proposed mmLD method performed higher powers or equal powers to than existing mapping methods under various simulation scenarios while maintaining the correct type-I errors. The mmLD method also showed good performance for QTL mapping of the GAW17 public dataset. We conclude that the mmLD method will be useful for future association analyses.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis of Variance
DNA	deoxyribonucleic acid
EM algorithm	Expectation-Maximization algorithm
GWAS	genome-wide association study
HWE	Hardy-Weinberg Equilibrium
LD	linkage disequilibrium
MAF	minor allele frequency
mmLD	Multi-marker LD mapping
QTL	quantitative trait loci
SKAT	Sequence Kernel Association Test
SMCP	Smoothed Minimax Concave Penalty
smLD	Single-marker Association Test
SNP	single nucleotide polymorphisms
tmLD	Two-marker LD mapping

Acknowledgments

I would never have been able to finish my research without the guidance of my committee members, help from friends, and love from my family.

Firstly, I would like to express my sincere gratitude to my advisor, Dr. Wu, for his warm guidance, support and encouragement for three years. He has showed me the way how to approach and solve critical and statistical problems. Through his training, I have been able to strengthen my academic knowledge. Besides my advisor, I would like to thank my co-advisor, Dr. Hou and committee members, Dr. Kuan and Dr. Hyman, for their insightful comments and encouragement. Their feedback were very helpful to improve my dissertation. I also thank Dr. Yang, for her support.

I thank my lab mates, Jiawen, Jiayu, Yijin, Jianjin, Hao, Fei, Qiao, Yang, Ziqi, Yaqi and the others, for their help. It was a great experience to study with them together.

I also thank my dearest friends in Stony Brook, Jayon, Jeewoen, Joowon, Jihye, Joyce, and Kwangmin and his family, Eunsuk and Inhye. Special thanks to Gongjun.

To my valued friends, Miyeon, Hyanga, and Soonim, I cannot thank them enough for their encouragement and support.

Lastly, I deeply thank my loving parents, Jonghwa Lee and Gilsoon Oh, for their love and caring. I also thank my brother, Hyunwoo, and his lovely family, Eunyoung, Hwiseo, and Dongha.

This dissertation is dedicated to my parents.

Chapter 1 Introduction

1.1. Background

The human body consists of 22 pairs of autosomal chromosomes and one pair of sex chromosomes. Chromosomes contain hereditary information encoded in double-strands of deoxyribonucleic acid (DNA). Genetic regulation of development, organismal function, and reproduction is programmed into these DNA sequences. DNA sequences are usually a combination of four nucleotides, guanine (G), adenine (A), thymine (T), and cytosine (C) [1]. For diploid species, such as humans, the DNA sequences are bi-allelic. Identical bi-alleles are known as homozygous; non-identical bi-alleles are known as heterozygous. Genomic DNA sequences are nearly identical between different people except for some minor genetic variation (< 1%). This genetic variation can cause different physical traits or specific diseases. A study that examines how genetic variants are related to biological traits across complete sets of genomes is called a genome-wide association study (GWAS).

A GWAS usually investigates many genetic variants simultaneously [1-3] to infer the relationship to biological traits. For example, case-control studies, a comparison of genotypic distributions between two groups of participants (case vs. control), have been used to detect genetic variants are significantly associated with each group [2, 3]. Most GWAS analyses have focused on the single-marker association, i.e., testing genetic

markers one at a time, mainly due to easy implementation. However, analyses based on multiple markers can be more powerful and will be the focus of this study.

In this thesis, I will first briefly introduce genetic background that serves as the basis for the development of my model. Next, I will review existing models of the GWAS. Finally, I will propose and describe a novel model, mmLD, for detecting the association between DNA variants and phenotypes by considering multiple markers. Detailed simulations and real data application will be presented to demonstrate the applicability of this new method.

1.2. Chromosomes, Single Nucleotide Polymorphisms (SNPs) and Quantitative Traits

Figure 1 illustrates the structures of chromosomes and DNA sequences. Within a population, members of the same biological species might differ at a single nucleotide. These differences represent a common and important type of genetic variation, single nucleotide polymorphisms (SNPs). SNPs are typically used as genomic markers in genetic studies [2]. Most SNPs in humans are bi-allelic and contain a major (M) allele and a minor (m) allele, determined by relative frequency. The frequency of the less common allele is usually referred as the minor allele frequency (MAF). Based on a combination of two alleles, a SNP usually has three genotypes; MM, Mm, and mm.

Many SNPs have been found by the International HapMap Project [4], which aimed to develop a haplotype map of the human genome to describe common patterns of genetic variation. The International HapMap Project used a variety of sequencing techniques to search for and catalog SNPs across the world. Thus far, the project includes 11 human populations with genotypes for 16 million SNPs. Through the HapMap project, the linkage disequilibrium can be examined with genotype data [2]. The genetic variants identified thus far have roles in human health, specific diseases or responses to drugs and environmental factors [2]. SNPs with a MAFs of 5% or greater were targeted by the HapMap project and have been used in GWAS analyses [4]. It is plausible that for these types of SNPs existing in a large population are inheritable and might explain biological variations that are known to have a genetic basis.

If a SNP is involved in a physical quantitative trait, it is called a quantitative trait locus (QTL) [3]. QTLs usually consist of dichotomous alleles and have three genotypes, denoted as QQ, Qq, and qq. In practice, QTLs are not always observable. Therefore, it is very difficult to identify which SNP is the QTL for a specific quantitative trait. The process of identifying QTLs associated with a phenotype is called QTL mapping.

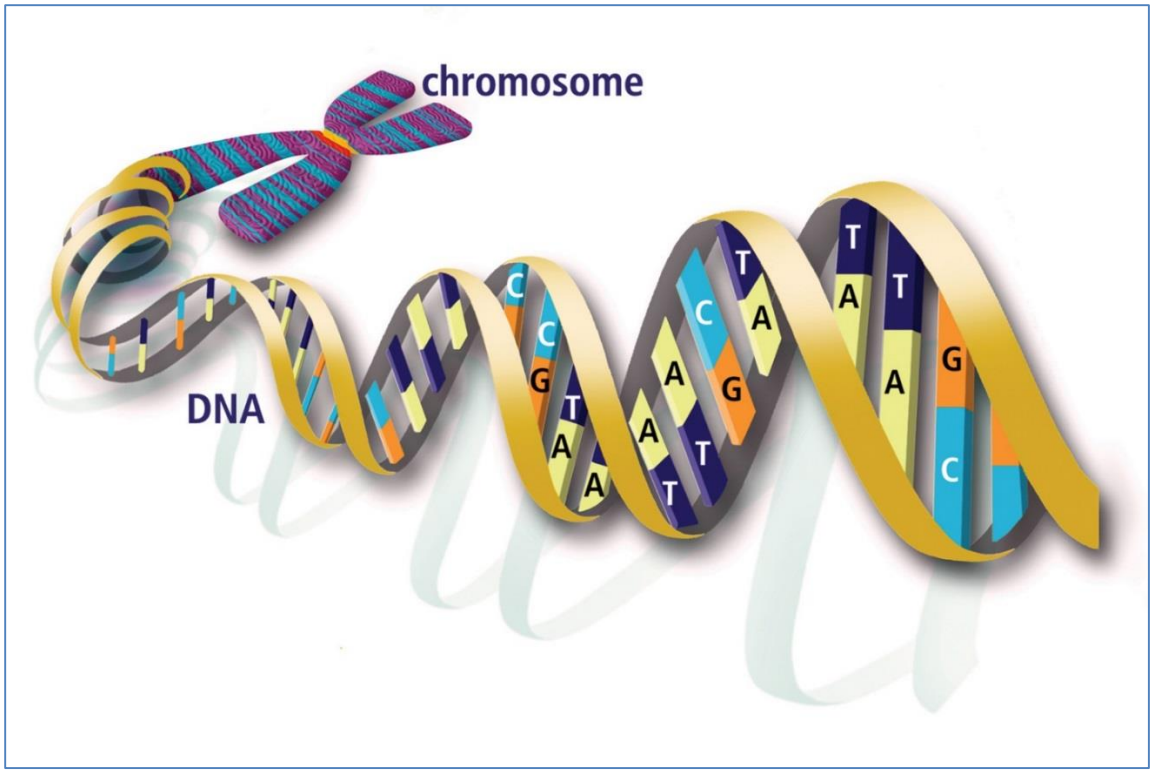

Figure 1 Chromosome and DNA sequences: Image from http://www.conservapedia.com/File:763.jpg

1.3. Haplotype under Linkage Disequilibrium

For humans, each genetic marker usually has two alleles, one inherited from the mother and the other from the father. A haplotype is a group of markers inherited from one parent that is physically located on the same chromosome. Therefore, genetic markers that are close to each other tend to be inherited together, except in case of

chromosomal crossover. Consequently, SNPs correlate with each other in a particular genetic region. This phenomenon is called linkage disequilibrium (LD).

More formally, LD describes the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP in a natural population [2]. The non-random association between alleles of the adjacent markers, measured by LD, is usually caused by several factors, such as the selection, mutation, genetic drift, system of mating, and population structure [5]. LD in a population usually decays over time. The rate of decay is dependent on multiple factors, such as the population size, number of founding chromosomes in the population, and number of generations. Different sub-populations might have different degrees and patterns of LD. Although LD was first introduced as a concept in the population genetics, it has recently been employed in QTL mapping methods. These LD-based mapping methods are particularly suitable for natural populations, such as humans, where controlled mating is implausible.

To illustrate the concept of LD, consider two bi-allelic markers, A and B. The major and minor alleles of marker A are denoted as *A* and *a*, and those of marker B as *B* and *b*. The frequencies of the two major alleles are denoted as p_A and p_B, respectively. The possible haplotypes of the two markers are *AB*, *Ab*, *aB*, and *ab* (*Figure 2*). The frequencies of these haplotypes are expressed as p_{AB}, p_{Ab}, p_{aB}, and p_{ab}. If two markers are independent of each other, i.e., they are not linked, each haplotype frequency should simply be a product of the two corresponding allele frequencies. However, if the two markers are correlated, as most adjacent markers are, the haplotype frequencies will

differ from the product of the allele frequencies. In *Table 1*, the D value denotes the deviation from the product of the corresponding allele frequencies and quantifies the degrees of the linkage disequilibrium genetically. The D value in *Table 1* represents the two-way LD between two markers. For three or more markers, there might be three-way or multi-way LDs. In these cases, representations and estimations are not simple as with two markers. *Figure 3* illustrates an example of the structure of LDs with multiple markers. The data was generated by *Forsim* software [6], a simulator used for generating gene sequences.

Table 1 Haplotype frequencies and allele probabilities of two markers

	A	a	Total
B	$p_{AB} = p_A p_B + D$	$p_{aB} = p_a p_B - D$	p_B
b	$p_{Ab} = p_A p_b - D$	$p_{ab} = p_a p_b + D$	$q_B = 1 - p_B$
Total	p_A	$q_A = 1 - p_A$	1

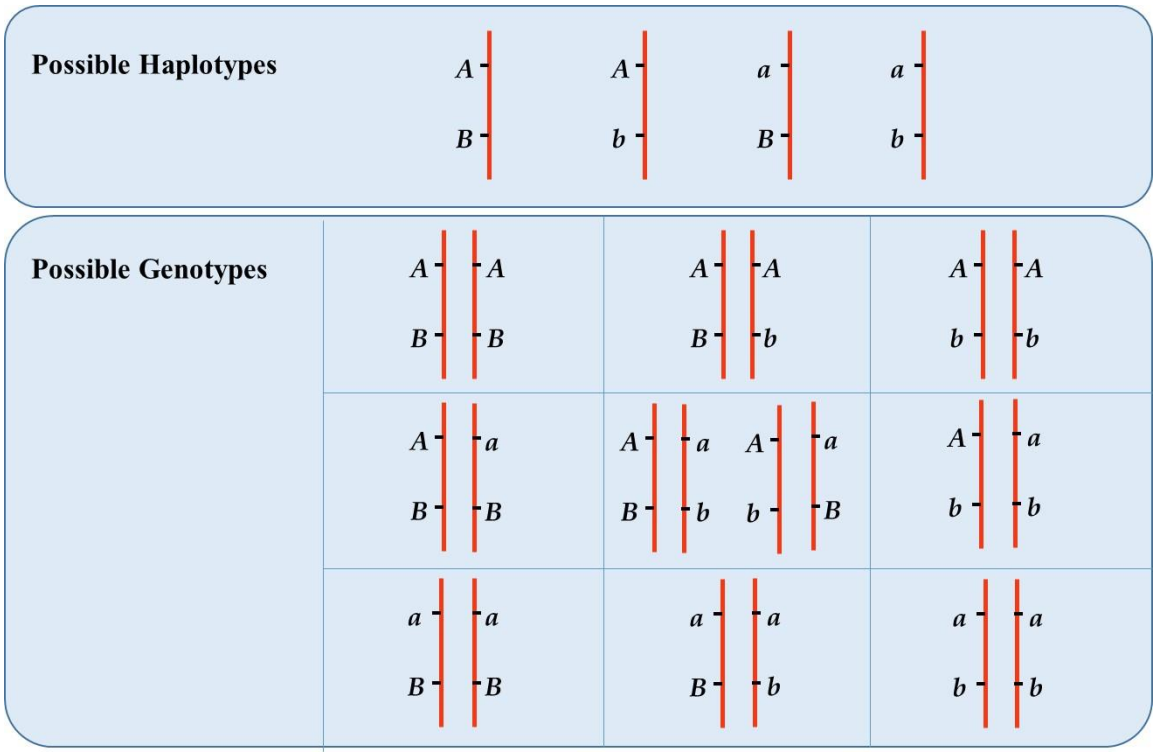

Figure 2 Possible haplotypes and genotypes for two markers

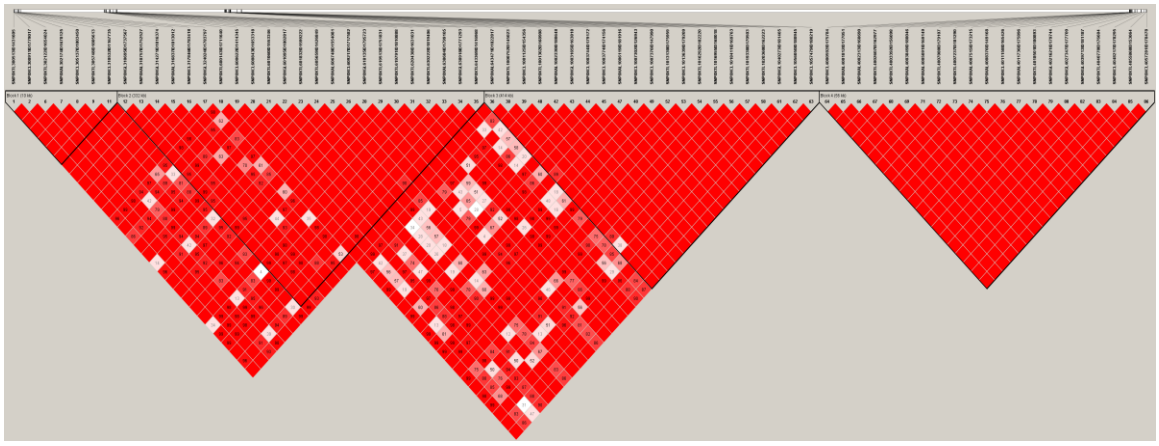

Figure 3 Example of linkage disequilibrium block generated by ForSim[6] software

1.4. Hardy-Weinberg Equilibrium, genotypic and haplotypic probabilities

In a natural population, allelic and genotypic probabilities usually remain constant through generations unless there are impacted by evolutionary influences such as mate choice, mutation, selection, genetic drift, gene flow, or meiotic drive. The Hardy-Weinberg Equilibrium (HWE) describes an ideal condition without these influences [3, 7, 8]. Under HWE, genotypic probabilities are simply the products of corresponding allele frequencies.

The allelic probabilities of major and minor allele for one marker can be denoted as p_0 and $p_1 = 1 - p_0$ respectively. The genotypic probabilities are the binomial expansion of the square of the sum of p_0 and p_1 as shown in *Table 2*. With two markers, there are four haplotype frequencies, p_{00}, p_{01}, p_{10}, and p_{11}, and nine possible genotypes. Based on *Figure 2*, the joint genotypic probabilities and their relationships to the haplotypic probabilities are summarized in *Table 3*. *Table 4* shows the genotypic and haplotypic probabilities of three markers with 27 possible genotypes. In general, for k markers, the number of genotypes is set to 3^k, while the number of haplotypes is 2^k.

As shown above, the relationship between genotypic and haplotypic probabilities can become very complicated when the number of markers is large. However, this calculation is essential for extending LD mapping to an arbitrary number of markers. In Chapter 3, I will propose an algorithm that can efficiently derive genotypic probabilities from haplotypic frequencies.

Table 2 Genotypic probabilities of one marker under Hardy-Weinberg Equilibrium

	AA	Aa	aa
Genotypic probabilities	p_0^2	$2p_0p_1 = 2p_0(1-p_0)$	$p_1^2 = (1-p_0)^2$

Table 3 Genotypic probabilities of two markers under Hardy-Weinberg Equilibrium and numbers n denote probabilities of each genotype

Genotypic probabilities	BB	Bb	bb
AA	p_{00}^2 (n_{00})	$2p_{00}p_{01}$ (n_{01})	p_{01}^2 (n_{02})
Aa	$2p_{00}p_{10}$ (n_{10})	$2p_{00}p_{11} + 2p_{01}p_{10}$ (n_{11})	$2p_{01}p_{11}$ (n_{12})
aa	p_{10}^2 (n_{20})	$2p_{10}p_{11}$ (n_{21})	p_{11}^2 (n_{22})

Table 4 Genotypic probabilities of three markers under Hardy-Weinberg Equilibrium[3]

Genotypic probabilities	CC	Cc	cc
AABB	p_{000}^2	$2p_{000}p_{001}$	p_{001}^2
AABb	$2p_{000}p_{010}$	$2p_{000}p_{011} + 2p_{010}p_{001}$	$2p_{001}p_{011}$
AAbb	p_{010}^2	$2p_{010}p_{011}$	p_{011}^2
AaBB	$2p_{000}p_{100}$	$2p_{000}p_{101} + 2p_{100}p_{001}$	$2p_{001}p_{101}$
AaBb	$2p_{000}p_{110} + 2p_{010}p_{100}$	$2p_{000}p_{111} + 2p_{010}p_{101} + 2p_{110}p_{001} + 2p_{100}p_{011}$	$2p_{001}p_{111} + 2p_{011}p_{101}$
Aabb	$2p_{010}p_{110}$	$2p_{010}p_{111} + 2p_{110}p_{011}$	$2p_{011}p_{111}$

aaBB	$p_{100}{}^2$	$2p_{100}p_{101}$	$p_{101}{}^2$
aaBb	$2p_{100}p_{110}$	$2p_{100}p_{111} + 2p_{110}p_{101}$	$2p_{101}p_{111}$
aabb	$p_{110}{}^2$	$2p_{110}p_{111}$	$p_{111}{}^2$

1.5. Estimating haplotypic frequency

Although many SNPs have been found by International HapMap projects [9], their genotypes cannot fully reveal the genetic structure of each individual because one genotypic pattern can arise from several possible haplotypic combinations. For example, the genotype AaBb, which is heterozygous at two markers, might come from two different sets of haplotypes, $AB|ab$ or $Ab|aB$ (*Figure 2*) [3, 10]. Therefore, haplotypes can be unknown even when genotypes are fully known. To address this issue, haplotypes are usually inferred by haplotype-estimation methods [11, 12].

Estimating haplotypic probabilities from genotypes has been studied for many years and several representative methods have been proposed, such as the expectation-maximization algorithm [11] and the Bayesian approach [12]. When heterozygous genotypes exist at more than one locus, it is difficult to estimate correct haplotype frequencies [11]. I will introduce the haplotype frequency estimation method with EM algorithm proposed by Excoffier and Slatkin (1995) [11], which will also be used in my model to be proposed in Chapter3.

Suppose there are m genotypes with corresponding frequencies P_i, $i = 1, \dots, m$. For a sample of n subjects, their probabilities can be expressed as a multinomial probability function:

$$\mathrm{P}(\text{sample}|P_1, P_2, \dots, P_m) = \frac{n!}{\prod_{i=1}^{n} n_i!} \prod \prod_{i=1}^{m} p_i^{n_i}$$

where n_i, $i = 1, \dots, m$ denotes the observed count for the m genotypes.

As previously discussed, haplotypes from samples with heterozygotes cannot be directly phased out and are referred to unphased genotypes. The number of unphased genotypes (c_j) is a function of the number of heterozygous loci s_j,

$c_j = 2^{s_j - 1}$, $\quad s_j > 0$ and $c_j = 1$, $\quad s_j = 0$

Under the assumption of HWE and random mating, the probability P_j of the j-th unphased genotype is given by the sum of probabilities of each of the possible c_j genotypes,

$$P_j = \sum_{i=1}^{c_j} P(\text{genotype } i) = \sum_{i=1}^{c_j} P(h_k h_l)$$

$$P(h_k h_l) = \begin{cases} p_k^2 & \text{when } k = l \\ 2p_k p_l & \text{when } k \neq l \end{cases}$$

where $P(h_k h_l)$ is the probability of the i-th genotype made up of haplotypes k and l and p_i denotes the frequency of the i-th haplotype. The likelihood of the haplotype frequencies is then given as follows:

$$\mathrm{L}(p_1, p_2, \ldots, p_h) = a_1 \prod_{j=1}^{m} \left(\sum_{i=1}^{c_j} P(h_{ik} h_{il}) \right)^{n_j}$$

where $p_h = 1 - p_1 - p_2 - \cdots - p_{h-1}$ and a_1 is a constant incorporating the multinomial coefficient. To obtain the maximum likelihood estimates for haplotypic probabilities, this likelihood can be solved by the EM algorithm, which is given in detail below.

E-step:

$$P(h_k h_l)^{(g)} = \frac{m_j}{m} \frac{P(h_k h_l)_j^{(g)}}{P_j^{(g)}}$$

$$P(h_k h_l)^{(g)} = \begin{cases} p_k^{(g)2} & \text{when } k = l \\ 2p_k^{(g)} p_l^{(g)} & \text{when } k \neq l \end{cases}$$

where $p_k^{(g)}$ and $p_l^{(g)}$ are the g-th iteration of frequencies of haplotype k and l.

M-step:

$$P_j^{(g)} = \sum_{i=1}^{c_j} P(\text{genotype } i)^{(g)} = \sum_{i=1}^{c_j} P(h_k h_l)^{(g)}$$

$$P(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{P(h_k h_l)^{(g)}}{P_j^{(g)}}$$

$$\widehat{p_t}^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j (h_k h_l)^{(g)}$$

where δ_{it} is an indicator variable equal to the number of times haplotype t is present in genotype i. EM algorithm iterates E and M steps until the difference of previous and current likelihoods converges into a small quantity.

This algorithm can be explained by an example with two markers, as shown in *Figure 2* and *Table3*. For two markers, the likelihood of haplotypic frequency is given as follows:

$$\begin{aligned}
&\mathrm{L}(p_1, p_2, \dots, p_h|\text{genotypes}) \\
&\quad \propto 2n_{00}\log(p_{00}) + n_{01}\log(2p_{00}p_{01}) + 2n_{02}\log(p_{01}) + n_{10}\log(2p_{00}p_{10}) \\
&\quad + n_{11}\log(2p_{00}p_{11} + 2p_{01}p_{10}) + n_{12}\log(2p_{01}p_{11}) + 2n_{20}\log(p_{10}) \\
&\quad + n_{21}\log(2p_{10}p_{11}) + 2n_{22}\log(p_{11})
\end{aligned}$$

Then, the corresponding EM algorithm can be explicitly written as:

E-step:

$$\phi = \frac{p_{00}p_{11}}{p_{00}p_{11} + p_{01}p_{10}}$$

where ϕ denotes the probability that the genotype $AaBb$ comes from haplotype $AB|ab$, and $1 - \phi$ denotes the probability from haplotype $Ab|Ab$.

M-step:

$$\hat{p}_{00} = \frac{1}{2n}(2n_{00} + \phi n_{11} + n_{01} + n_{10})$$

$$\hat{p}_{10} = \frac{1}{2n}(2n_{20} + n_{10} + (1-\phi)n_{11} + n_{21})$$

$$\hat{p}_{01} = \frac{1}{2n}(2n_{02} + n_{01} + (1-\phi)n_{11} + n_{12})$$

$$\hat{p}_{11} = \frac{1}{2n}(2n_{22} + \phi n_{11} + n_{12} + n_{21})$$

1.6. Heritability value

In genetics, heritability refers to how much of the variation in a trait is caused by genetic variants in a natural population. Other causes of the variation could be environmental factors or genetic drift [13]. Statistically, heritability indicates the proportion of phenotypic variance that is attributable to genetic variance. Therefore, it can be measured by estimating the relative contributions of genetic and non-genetic differences to the total phenotypic variation. The model of heritability can be defined as follows:

Phenotype (*P*) = Genotype (*G*) + Environment (*E*)

The variance of phenotype can be shown as Var(*P*) = Var(*G*) + Var(*E*) + 2Cov(*G*,*E*). In a designed experiment, Cov(G,E) can be controlled. If it is defined as 0, then the heritability is shown as follows:

$$H^2 = \frac{\mathrm{Var}(G)}{\mathrm{Var}(P)}$$

Therefore, the range of H^2 is from 0 to 1. If H^2 is 0, the genetic variant does not have any contribution to the phenotype. If H^2 = 1, the variation of the phenotype is caused only by the genetic variant.

In this study, the heritability value will be used to design various simulations for the null and alternative hypothesis.

Chapter 2 Review of previous GWAS methods

In this chapter I will review some GWAS methods, later used for model comparison. First, I will introduce single-marker based tests—single-marker LD mapping and adjusted and minimum p-value single- marker LD mapping. Second, I will introduce multi-marker based tests—two-marker LD mapping, the Sequence Kernel Association Test, and the Smoothed Minimax Concave Penalty.

2.1. Genome-wide Association Study

Methods used in GWAS analyses can be broadly classified by phenotypic features. There are two major types of phenotypic traits, quantitative (continuous) and qualitative (dichotomous case-control). Quantitative traits are measurable phenotypes, such as weight, height, and blood pressure, which are usually assumed normally distributed. Qualitative traits are typically binary (case vs. control) in nature, such as with or without a specific disease. Different GWAS methods can be applied to detect interesting associations according to phenotypic feature.

GWAS methods can also be grouped by the number of markers considered in each model, such as single-marker or multi-marker based test. As the aim of this thesis is to develop a novel method based on multiple markers, I will review some important single- and multi-marker based methods that are most relevant to this study.

2.2. Single-marker based tests

A single-marker test is used to examine the association between phenotypes and a single marker using statistical methods. Quantitative traits are commonly analyzed with ANOVA or linear regression based approaches [2, 14]. ANOVA compares three means of phenotype independently by the genotypes of a single-marker. Linear regression assumes linearity for three means of phenotype. In both methods, the null hypothesis is that there is no difference between the three mean values of phenotype. Both methods also assume that quantitative phenotypes are normally distributed and that the variance of the phenotypes within each group is constant [2]. For dichotomous case-control phenotypes, a 2 by 3 contingent table is formed, and independence can be tested with the Pearson χ^2 test or Fisher exact test [14]. The null hypothesis is that no association exists between the binary phenotype and three genotypes of each marker.

LD mapping, another approach that has recently emerged for QTL mapping, uses LD information to link phenotype to genotypes [15]. The idea of LD mapping is that a phenotype-related QTL is linked to a small group of genetic sequences that can be directly incorporated into the mapping model. Most LD mapping has focused on a single-marker association [15]. As our new method is also based on LD mapping, I will introduce the single-marker LD mapping method in detail in what follows [15-17].

2.2.1. Single-marker Linkage Disequilibrium (smLD) mapping

Suppose *n* samples with the quantitative phenotype and corresponding one dichotomous marker. The null hypothesis is that there is no linkage disequilibrium between one marker and the QTL (H_0: $D_{\mathcal{M},Q} = 0$) and the alternative hypothesis is the opposite. Assume that the phenotype is normally distributed with three different means and a common variance by three genotypes. In addition, there are six parameters to be estimated under the null hypothesis; $p_{\mathcal{M}}, p_Q, \mu_j, \sigma^2, j$ = 0, 1, 2, and one more parameter of the linkage disequilibrium (D_{1Q}) is added under the alternative hypothesis. The likelihood function can be written as:

$$L\left(p_Q, D_{\mathcal{M},Q}, \mu_0, \mu_1, \mu_2, \sigma^2 \mid y, \mathcal{M}\right) = \prod_{i=1}^{n} \sum_{j=0}^{2} \pi_{j|i} f_i(y_i|\mu_0, \mu_1, \mu_2, \sigma^2)$$

In order to obtain the maximum likelihood estimates (MLEs) of the parameters, the EM algorithm can be applied. The deviance of likelihoods between null and alternative hypothesis approximately follows χ^2-distribution with 1 degree of freedom because one parameter of the LD value is added under the alternative hypothesis. Therefore, when the deviance is rejected, it means the marker is linked with the QTL significantly.

2.2.2. Adjusted and minimum p-value smLD

Although the smLD is an efficient method to map a QTL based on one SNP, the multiple-testing issue may occur when multiple markers are considered in a GWAS. Therefore, to address this issue, Bonferroni correction can be applied, and it is called the

adjusted smLD. Bonferroni correction is a conservative adjustment for multiple-testing by controlling family-wise type I error.

As an alternative to the adjusted smLD, the minimum p-value smLD can also be applied, which rejects the null hypothesis when at least one of multiple testings is rejected. The minimum p-value smLD is likely to reject the null hypothesis too easily. Due to this, we expect arise of a high power, while it leads to an inflated type I error as well. Although the minimum p-value smLD cannot be used as the reliable method for the multiple-testing, it might be useful for the power comparison. In this study, the minimum p-value smLD is applied for the power comparison to verify the efficiency of our newly proposed mmLD method.

2.3. Multi-marker based tests

Although these single-marker methods are straightforward, they usually suffer from limited powers. Recently, more and more researches have shifted the focus to the association test with multiple markers. However, the multiple-marker association is usually not a simple extension from the single-marker association test [2], and is statistically and computationally more challenging. In the following, I will introduce some recent development on association methods using multiple markers.

2.3.1. Two-marker LD mapping

The two-marker LD mapping (tmLD) is built on the smLD and was first proposed by Yang et al (2014) [18]. In the two-marker LD mapping framework, it assumes a dichotomous QTL of alleles major Q and minor q, which is causal but unobserved. Also, it considers two neighboring markers $\mathcal{M}_1$ and $\mathcal{M}_2$, and assumes that the phenotype affected by the QTL follows a mixture Gaussian distribution with three different means and the same variance by the genotypes of the QTL.

Because the tmLD examines the association of two known markers and an unobserved QTL, there are four LDs as follows; D_{12}, D_{1Q}, D_{2Q}, and D_{12Q}. The first LD D_{12} can be estimated directly by the two markers which are already observed. However, the other three LDs (D_{1Q}, D_{2Q}, D_{12Q}) between two markers and the QTL should be tested for whether they have zero quantities or not. Therefore, the null hypothesis is that the QTL is not associated with two adjacent SNP markers (H_0: $D_{1Q} = D_{2Q} = D_{12Q} = 0$) and the alternative is the opposite. The likelihood function can be written as:

$$L(\Omega_p, \Omega_q \mid y, \mathcal{M}_1, \mathcal{M}_2) = \prod_{i=1}^{n} \sum_{j=0}^{2} \pi_{j|i} f_i(y_i|\Omega_q)$$

where Ω_p denotes the parameters of haplotype frequencies of two known markers and the QTL and Ω_q does the phenotypic parameters; μ_j, σ^2, j = 0, 1, 2.

The tmLD uses the EM algorithm to obtain MLEs for the parameters. The computational algorithms are given in what follows:

Step 1: Give initial values to the unknown parameters (Ω_p, Ω_q)

Step 2: E-Step – Calculate the posterior probabilities for each subject i carrying a particular QTL genotype j using the equation $\Pi_{j|i} = \frac{\pi_{j|i} f(y_i|\Omega_q)}{\sum_{j=0}^{2} \pi_{j|i} f(y_i|\Omega_q)}$.

Step 3: M step – Solve the log-likelihood equations for each parameter based on the observed data and $\Pi_{j|i}$ to obtain its estimate. To estimate the quantitative genetic parameters (Ω_q), their expressions in closed forms can be derived based on the estimation equations. For the estimates of the population genetic parameters (Ω_p), another inner layer of EM algorithm can be employed.

Step 4: Repeat the E and M steps until the estimates converge to stable values. The estimates at convergence are the MLEs of parameters.

The tmLD uses the likelihood ratio test with 3 degree of freedom. If the deviance of likelihoods between the null and alternative hypothesis is rejected, it means that the QTL is located on two adjacent markers and linked with them. The tmLD shows better power performances than those of single-marker association test [18]. Therefore, it is reasonable to assume that the LD mapping including more than two-markers is likely to have the better powers for mapping QTL than the smLD.

However, in order to use multiple markers, some limitations of the tmLD must be solved first. The first is to derive genotypic probabilities from estimated haplotype frequencies under Hardy-Weinberg Equilibrium for three or more markers because its calculation becomes complicated. The second issue is how to determine the degree of

freedom of LRT for the multiple-markers. Therefore, these challenges are the motivation of my research.

2.3.2. Sequence Kernel Association Test

The sequence kernel association test (SKAT) is another multi-marker based methods proposed by Wu (2011) and Lee (2012) [19, 20]. It is a regression approach that tests for the association between SNPs in a small panel and either case-control or continuous phenotypes [19]. The SKAT uses a multiple regression model to directly regress the phenotype on genetic variants (SNPs) in a region and on covariates. It assumes that n individuals are sampled with observed K SNPs. Covariates such as the demographic variables and top principal components of genetic variation are allowed for controlling population stratification. For i-th subject, y_i denotes the phenotype and $\mathbb{X}_i = (X_{i1}, X_{im}, \dots, X_{im})$ and $\mathbb{G}_i = (G_{i1}, G_{im}, \dots, G_{iK})$ do the covariates and genotypes, respectively. The model for the continuous phenotypes is as follows:

$$y_i = \alpha_0 + \alpha' \mathbb{X}_i + \beta' \mathbb{G}_i + \varepsilon_i$$

where α_0 is an intercept term and $\alpha = [\alpha_1, \dots, \alpha_m]'$ is the vector of regression coefficient for the m covariates and, $\beta = [\beta_1, \dots, \beta_K]'$ is the vector of regression coefficients for the K observed SNPs, ε_i is an error term with zero mean and σ^2 variance. The null hypothesis is that the coefficients of variants (SNPs) are zero; $H_0: \beta = 0$. However, since the standard likelihood ratio test has little power, the SKAT tests assume that each β_k

follows an arbitrary distribution with zero mean and $w_k\tau$ variance, where τ is a variance component and w_k is a pre-specified weight for variant k. Therefore, it uses a variance-component score test in the corresponding mixed model.

$$\mathrm{Q} = (\mathrm{y} - \hat{\mu})'\mathbb{K}(\mathrm{y} - \hat{\mu})$$

where $\mathbb{K} = \mathbb{G}\mathbb{W}\mathbb{G}'$, $\mathbb{G}$ is an $n \times k$ matrix with the (i,k)-th element being the genotype of variant k of subject i, $\mathbb{W} = \mathrm{diag}(w, \dots, w_K)$ contains the weights of the K SNPs, $\hat{\mu} = \widehat{\alpha_0} + \mathrm{X}\hat{\alpha}$ is the predicted mean of y and $\widehat{\alpha_0}$ and $\hat{\alpha}$ are estimates of covariates $\mathbb{X}$ under the null hypothesis. The power of the SKAT depends on choices of weights (w_k). Under H_0, Q follows a mixture of χ^2-distributions. The SKAT is computationally efficient and it yields p-value easily with simple analytic formulae, and the features of the SKAT are exploration of local correlation structure, incorporation of flexible weights to boost power and allowance for epistatic variant effects.

This SKAT for continuous phenotypes without covariates is applicable for the power comparison with the mmLD method because it is able to deal with multiple markers to detect their relationships with continuous phenotypes.

2.3.3. Smoothed Minimax Concave Penalty

The smoothed minimax concave penalty (SMCP) method was proposed by Liu (2013) [21]. The SMCP is a penalized regression method for identifying important SNPs in GWAS. It is a combination of the minimax concave penalty (MCP) proposed by

Zhang [22] and a penalty, consisting of the squared differences of the absolute effects of adjacent markers. The MCP promotes sparsity in the model and selects significant SNPs and the penalty for the squared differences of absolute effects which take into account the natural ordering of SNPs and adaptively incorporates the LD structure between adjacent SNPs [21]. The MCP is defined as

$$\rho(t;\ \lambda_1, \gamma) = \ \lambda_1 \int_0^{|t|} \left(1 - \frac{x}{\gamma \lambda_1}\right)_+ dx.$$

where λ_1 is a penalty parameter, and γ is a regularization parameter that minimizes the maximum concavity.

The SMCP assumes that K SNPs and β_k are the effects of the k-th SNP in the model which describes the relationship between phenotype and markers, and the SNPs are ordered by their physical locations on a chromosome. Neighboring SNPs in high LD are expected to have similar strength of association with the phenotype. The penalty encourages smoothness in $|\beta|$s at adjacent markers.

$$\frac{\lambda_2}{2} \sum_{k=1}^{K-1} \varsigma_k (|\beta_k| - |\beta_{k+1}|)^2$$

where the weight ς_k is a measure of LD between the k-th and k+1-th SNPs and it promotes $|\beta_k|$ and $|\beta_{k+1}|$ to be similar to an extent inversely proportional to the LD strength between the corresponding SNPs. Neighboring SNPs in weak LD are allowed to have the larger difference in smoothness in $|\beta|$s than if they are in stronger LD.

The SMCP is applicable for scanning a dense set of SNPs by incorporating the LD information. Although it is an efficient tool to deal with a large number of SNPs, it is

not comparable with the mmLD because it does not give the overall p-value but the p-value of each SNP. Meanwhile, the mmLD gives only the overall p-value for the multiple markers. Therefore, it was not plausible to compare their powers in this study.

Chapter 3 Multiple-marker LD mapping

In this chapter, I will propose a novel method in details. It is organized as follows: I will introduce a novel model mmLD and explain how it works for the QTL mapping. In addition, I will show the results of various simulations to verify the applicability of the mmLD.

3.1. Method

3.1.1. Setting multi-marker LD (mmLD) mapping

In the mmLD mapping framework, we assume a dichotomous quantitative trait locus (QTL, $\mathcal{Q}$) of alleles Q and q that is causal and is unobserved, and the allele frequencies of Q and q are expressed as p_Q and $p_q = 1 - p_Q$. Suppose that the QTL is genetically linked to a group of genotyped SNP markers, $\mathcal{M}_\mathrm{i}$ ($\mathrm{i} = 1, \ldots, \mathrm{k}, \mathrm{k} > 2$) that are from a LD block, and each marker $\mathcal{M}_\mathrm{i}$ has two alleles $\mathrm{M_i}$ and $\mathrm{m_i}$ with corresponding frequencies of p_i and $1 - p_i$. Then, the k markers may form 2^k possible haplotypes, and form 2^{k+1} possible joint haplotypes together with the unknown QTL. Let $\mathrm{p}_{\mathcal{M}_1 \cdots \mathcal{M}_\mathrm{k}}$ denotes the frequency of haplotypes formed by k markers, and $\mathrm{p}_{\mathcal{M}_1 \cdots \mathcal{M}_\mathrm{k}, \mathcal{Q}}$ the frequency of the joint haplotype formed by the k markers and the QTL. The LD between the QTL and the group of markers can be described by the following equation:

$$p_{\mathcal{M}_1 \cdots \mathcal{M}_k, \mathcal{Q}} = p_{\mathcal{M}_1 \cdots \mathcal{M}_k} p_Q + D_{\mathcal{M}_1 \cdots \mathcal{M}_k, \mathcal{Q}} \qquad \textit{Eqn 1}$$

In general, there are $2^k - 1$, $D_{\mathcal{M}_1 \cdots \mathcal{M}_k, \mathcal{Q}}$ values, one for each marker haplotype. If all $D_{\mathcal{M}_1 \cdots \mathcal{M}_k, \mathcal{Q}}$s are zeros, it indicates that the QTL and the marker group are independent; otherwise, the QTL and the marker group are in LD.

3.1.2. Mixture Gaussian Model of mmLD

Suppose that there is a random sample of size n, For subject i $(i = 1, \ldots, n)$, k markers $(\mathcal{M}_{i1} \cdots \mathcal{M}_{ik})$ have been genotyped and a continuous phenotypic trait (y_i) has also been obtained. Assume these samples are drawn from a natural population under HWE, and the continuous trait is directly affected by the QTL. Then, the relationship between the observed phenotypes and their expected means, which are determined by the genotypes of the QTL, can be described as follows:

$$y_i = \sum_{j=0}^{2} \xi_{ij} \mu_j + e_i, \quad i = 1, \ldots, n \qquad Eqn\ 2$$

where ξ_{ij} is an indicator variable defined as 1 if subject i has a QTL genotype j (2 for QQ, 1 for Qq, and 0 for qq), μ_j is the expected phenotypic mean for a QTL genotype j, and e_i is the error term that is assumed to follow a Gaussian normal distribution with zero mean and variance σ^2. Based on the conditional probability of subject i carrying a certain QTL genotype j given its markers, $\pi_{j|i} = P(\mathcal{Q} = j|\mathcal{M}_{i1} \cdots \mathcal{M}_{ik})$ or $P(\xi_{ij} = 1)$, the likelihood of the phenotype and multiple markers can be constructed by the following mixture model:

$$L\left(\Theta_p,\Theta_q;y,M_1,...,M_k\right)=\prod_{i=1}^{n}\left[\sum_{j=0}^{2}\pi_{j|i}\,f_j\left(y_i;\Theta_q\right)\right]=\prod_{i=1}^{n}\left[\sum_{j=0}^{2}\pi_{j|i}\,\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right)\right]$$

Eqn 3

where Θ_p is a vector of the population genetic parameters that describe haplotype frequencies including the k known markers and one putative QTL, $\Theta_q = \{\mu_0, \mu_1, \mu_2, \sigma^2\}$ is a vector of parameters of the phenotypic traits. We assume that these phenotypic values are normally distributed. The calculation $\pi_{j|i}$ becomes complicated as k, the number of known markers, increases. This issue will be addressed in the next section.

3.1.3. Calculation of joint and conditional genotypic probabilities with *k* markers

For small k (2 or 3), the conditional genotypic probabilities can be easily expressed and calculated by the haplotype frequencies in *Table 2* and *Table 3* [7, 8]. However, since the numbers of haplotypes (2^k) and genotypes (3^k) increase exponentially with the arbitrary number of genetic loci, the calculation of genotypic probabilities becomes difficult for large k. In general, genotypes are classified into homozygote or heterozygote [24]. The homozygote denotes the identical bi-alleles at each locus and the heterozygote does bi-alleles with different alleles at a specific locus [25]. If a subject has homozygotes at all genetic loci, its joint genotypic probabilities can be simply calculated with the square of a haplotype frequencies. However, if at least one of loci contains the heterozygote, the calculation of joint genotypic probabilities becomes

complicated as one genotype may come from combinations of different haplotypes as shown in *Figure 4*.

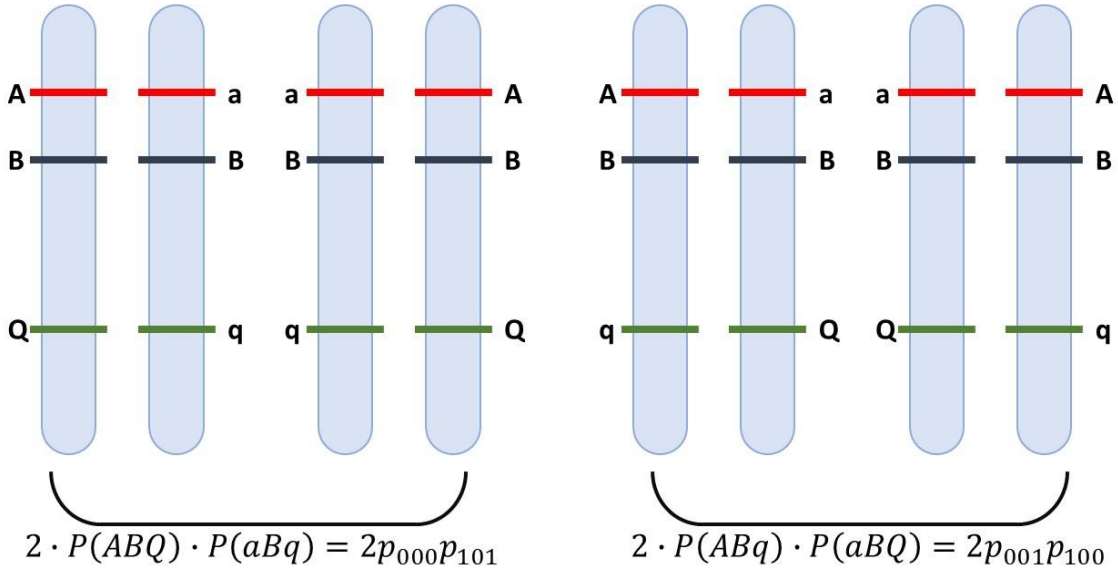

Figure 4 Example of genotypic probabilities of heterozygote

For the purpose of the simplicity, for a specific locus, its two alleles are denoted as 0 and 1 and correspondingly, its three genotypes can be denoted as 0, 1, and 2, as simply the summation of the two alleles. Below we describe a general algorithm for the calculation of joint genotypic probabilities for the combinations of genotypes at k markers. Here, G_k denotes the genotype of the k markers.

Step 1. Determine zygotic status (homozygote vs heterozygote) for genotypes at all loci.

Step 2. If a subject has homozygote genotypes at all k markers like G_{00000} and G_{22222}, its genotype frequency is the square of the corresponding haplotype frequency. For example,

if there are five markers (k = 5) and its genotypes are given like G_{00202} and G_{02002}, then their genotypic probabilities can be calculated to $P(G_{00202}) = p_{00101}{}^2$, and $P(G_{02002}) = p_{01001}{}^2$, respectively.

Step 3. If at least one of loci has the heterozygote,

Sub-step 3.1. Consider the haplotypic frequencies which have the correct major or minor allele correspond to the homozygote loci.

Sub-step 3.2. Combine the haplotypic frequencies of which one has major allele and the other has minor allele at the heterozygote loci.

Sub-step 3.4. Genotype probability is the sum of two times the combination of haplotypic frequencies. For example, if there are five loci and its genotype is given like G_{11202}, then its genotypic probability is $P(G_{11202}) = 2p_{00101}p_{11101} + 2p_{10101}p_{01101}$.

Table 5 Joint genotypic probabilities of k marker

Genotypic probabilities	M_kM_k	M_km_k	m_km_k
$M_1M_1\ M_2M_2 \ldots M_{k-1}M_{k-1}$	$p_{00\ldots0}{}^2$	$2p_{00\ldots0}p_{00\ldots1}$	$p_{00\ldots1}{}^2$
$M_1M_1\ M_2M_2 \ldots M_{k-1}m_{k-1}$	$2p_{00\ldots0}p_{0\ldots10}$	$2p_{00\ldots0}p_{0\ldots11} + 2p_{0\ldots10}p_{00\ldots1}$	$2p_{00\ldots1}p_{0\ldots11}$
$M_1M_1\ M_2M_2 \ldots m_{k-1}m_{k-1}$	$p_{0\ldots10}{}^2$	$2p_{0\ldots10}p_{0\ldots11}$	$p_{0\ldots11}{}^2$
...	...	...	...

$m_1m_1\ m_2m_2 \ldots m_{k-1}m_{k-1}$	$p_{11\ldots0}{}^2$	$2p_{11\ldots0}p_{11\ldots1}$	$p_{11\ldots1}{}^2$

With this algorithm, *Table 5* shows the example of joint genotypic probabilities for k genetic markers. If one is QTL (the column variables), the conditional probabilities of QTL given markers $(\pi_{j|i})$ can then be derived, correspondingly.

3.1.4. Conditional joint genotypic probabilities

In the mmLD procedure, the mmLD conducts two-phase estimations, separately. The first-phase is to estimate haplotype frequencies with k known markers and the second-phase is to update haplotype frequencies with both k known markers and the QTL. Here, estimations of the second phase depend on the estimates of haplotype frequencies from the first phase. In other words, the 2^{k+1} number of haplotype frequencies in the second-phase are derived from the 2^k number of estimates of haplotype frequencies from the first-phase. Thus, this dependence of haplotype frequencies corresponds to the dependence of genotype probabilities.

Each posterior probability ($\pi_{j|i} = P(Q = j|\mathcal{M}_{i1} \cdots \mathcal{M}_{ik})$ of the QTL is expressed as the conditional joint genotypic probabilities in *Eqn 3*; the genotypic probability of the second phase over the genotypic probability of the first phase. Also, the summation of three posterior probabilities should be 1 for each subject $i = 1, 2, \ldots, n$.

$$\sum_{j=0,1,2} \pi_{j|i} = 1$$

Table 6 shows the example of the conditional joint probabilities of one known marker and one QTL.

Table 6 Table of conditional probabilities of 1 SNP and 1 QTL for several subjects

Subject	Genotypes of one SNP	Genotypic probability of one marker	Conditional probabilities $(\pi_{j\|i})$ for unknown QTL genotypes		
			QQ (j = 0)	Qq (j = 1)	qq (j = 2)
1	AA	$p_0{}^2$	$\frac{p_{00}{}^2}{p_0{}^2}$	$\frac{2p_{00}p_{01}}{p_0{}^2}$	$\frac{p_{01}{}^2}{p_0{}^2}$
2	Aa	$2p_0p_1$	$\frac{2p_{00}p_{10}}{2p_0p_1}$	$\frac{2p_{00}p_{11}+2p_{01}p_{10}}{2p_0p_1}$	$\frac{2p_{01}p_{11}}{2p_0p_1}$
3	Aa	$2p_0p_1$	$\frac{2p_{00}p_{10}}{2p_0p_1}$	$\frac{2p_{00}p_{11}+2p_{01}p_{10}}{2p_0p_1}$	$\frac{2p_{01}p_{11}}{2p_0p_1}$
4	aa	$p_1{}^2$	$\frac{p_{10}{}^2}{p_1{}^2}$	$\frac{2p_{10}p_{11}}{p_1{}^2}$	$\frac{p_{11}{}^2}{p_1{}^2}$
…	….	…	…	…	…
n	aa	$p_1{}^2$	$\frac{p_{10}{}^2}{p_1{}^2}$	$\frac{2p_{10}p_{11}}{p_1{}^2}$	$\frac{p_{11}{}^2}{p_1{}^2}$

3.1.5. Expectation-Maximization (EM) algorithm

Maximum likelihood estimates (MLEs) of parameters can be obtained by maximizing the log-likelihood function (*Eqn 3*). The EM algorithm can be very efficient for parameter estimations in the mixture model. In this study, I propose a two-phase EM

algorithm. The first phase is to estimate haplotype frequencies based on known genotypes of k markers and the second phase is to update haplotype frequencies with unknown QTL and phenotypic parameters. Therefore, two-phase estimations are performed, separately in the mmLD procedure.

Since the mmLD assumes that the QTL is unknown, estimating haplotype frequencies for k known markers and the QTL in the second phase entirely depends on the estimates of haplotype frequencies for k known markers from the first phase. For the first phase, estimating haplotype frequencies for known genotypic probabilities has been studied with various methods as explained in Chapter 1. Thus, in this study, the EM algorithm, which is one of popular methods proposed by Excoffier and Slatkin (1995) [11], is implemented in the first phase estimation[26-28] . The "*haplo-stat*" is one of well-known software for the estimation of haplotype frequencies based on EM algorithm [26]. The detailed algorithmic working flow is described as follows:

Step 1: Estimate the 2^k haplotype frequencies for k known markers from given genotype data (1st-phase EM). The R-package "*haplo.stat*" is applied [26].

Step 2: Initialize parameters ($\Theta^{(0)}$) of 2^{k+1} haplotype frequencies ($\Theta_p^{(0)}$) which include k known markers and one QTL, and phenotypic parameters ($\Theta_q^{(0)}$).

Step 3: Derive the 3^{k+1} joint conditional genotypic probabilities based on initialized or updated haplotype frequencies of k known markers and one QTL.

Step 4: **E-step** Calculate posterior probabilities for each subject i carrying the particular QTL genotypes j.

For each iteration (r = 1, 2, …), log-likelihood function is

$$\log L = \ell = \sum_{i=1}^{n}\sum_{j=0}^{2}\left[z_{ij}\log\left(\pi_{j|i}\right) + z_{ij}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + z_{ij}\left(-\frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right)\right] \quad \textit{Eqn 4}$$

Posterior probabilities: $E\left(z_{ij}\right) = \Pi_{j|i}^{(r)} = \dfrac{\pi_{j|i}^{(r-1)} f_j\left(y_i \mid \Theta_q^{(r-1)}\right)}{\sum_{j=0}^{2}\pi_{j|i}^{(r-1)} f_j\left(y_i \mid \Theta_q^{(r-1)}\right)}$ *Eqn 5*

, where i = 1, 2, …, n, j = 0, 1 or 2 and $\pi_{j|i}^{(r-1)}$ is the prior probabilities.

$$E(\ell) = E\left(\sum_{i=1}^{n}\sum_{j=0}^{2}\left[z_{ij}\log\left(\pi_{j|i}\right) + z_{ij}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + z_{ij}\left(-\frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right)\right]\right)$$

$$= \sum_{i=1}^{n}\sum_{j=0}^{2}\Pi_{j|i}^{(r+1)}\log\left(\pi_{j|i}\right) + \sum_{i=1}^{n}\sum_{j=0}^{2}\Pi_{j|i}^{(r+1)}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right]$$

$$= E_{z_{ij}}\left(\ell\left(\Theta_p\right); y, \Theta_p^{(r)}, \Theta_q^{(r)}\right) + E_{z_{ij}}\left(\ell\left(\Theta_q\right); y, \Theta_p^{(r)}, \Theta_q^{(r)}\right) \quad \textit{Eqn 6}$$

Step 5: M-step Solve the log-likelihood equations for each parameter based on observed data and $\Pi_{j|i}$ to obtain its estimate. The expectation of log-likelihood equations can be divided into two parts—phenotypic and haplotypic parameters in (6).

Therefore, both parts must be maximized, separately. To estimate the phenotypic parameters (Ω_q), their expressions in closed forms can be derived based on the estimation equations. For the estimates for the haplotypic parameters (Ω_p), another inner layer of EM algorithm can be employed.

M1. to maximize $E_{z_{ij}}\left(\ell\left(\Theta_q\right); y, \Theta_p{}^{(r)}, \Theta_q{}^{(r)}\right) = \sum_{i=1}^{n}\sum_{j=0}^{2}\Pi_{j|i}{}^{(r+1)}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right]$

1) Estimates of means

$$-\frac{\partial\ell}{\partial\mu_j} = \sum_{i=1}^{n}\left[2\Pi_{j|i}{}^{(r+1)}\left(\frac{\left(y_i-\mu_j\right)}{2\sigma^2}\right)\right] = 0$$

$$=> \qquad \hat{\mu}_j{}^{(r+1)} = \frac{\sum_{i=1}^{n}\Pi_{j|i}{}^{(r+1)} y_i}{\sum_{i=1}^{n}\Pi_{j|i}{}^{(r+1)}} \quad \text{for } j = 0, 1, \text{ or } 2 \qquad \textit{Eqn 7}$$

2) Estimates of variance

$$\frac{\partial\ell}{\partial\sigma^2} = \sum_{i=1}^{n}\sum_{j=0}^{2}\left[-\frac{\Pi_{j|i}{}^{(r+1)}}{2\sigma^2} + \Pi_{j|i}{}^{(r+1)}\left(\frac{\left(y_i-\mu_j\right)^2}{2\left(\sigma^2\right)^2}\right)\right] = 0$$

$$=> \qquad \hat{\sigma}^{2(r+1)} = \frac{\sum_{i=1}^{n}\sum_{j=0}^{2}\Pi_{j|i}{}^{(r+1)}\left(y_i-\hat{\mu}_j{}^{(r+1)}\right)^2}{\sum_{i=1}^{n}\sum_{j=0}^{2}\Pi_{j|i}{}^{(r+1)}} \qquad \textit{Eqn 8}$$

M2. to maximize

$$E_{z_{ij}}\left(\ell\left(\Theta_p\right); y, {\Theta_p}^{(r)}, {\Theta_q}^{(r)}\right) = \sum_{i=1}^{n}\sum_{j=0}^{2} \Pi_{j|i}^{(r+1)} \log\left(\pi_{j|i}\right) \qquad \textit{Eqn 9}$$

Table 7 Example of maximization of haplotype frequencies including the QTL

Genotypes of known markers	Unknown QTL		
	QQ (0)	Qq (1)	qq (2)
$\underbrace{000....0}_{k}$	$N_{\underbrace{000...0}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...0}_{k}\right)\pi_{i0}$	$N_{\underbrace{000...1}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...0}_{k}\right)\pi_{i1}$	$N_{\underbrace{000...2}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...0}_{k}\right)\pi_{i2}$
$\underbrace{000....1}_{k}$	$N_{\underbrace{00...10}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...1}_{k}\right)\pi_{i0}$	$N_{\underbrace{00...11}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...1}_{k}\right)\pi_{i1}$	$N_{\underbrace{00...12}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...1}_{k}\right)\pi_{i2}$
$\underbrace{000....2}_{k}$	$N_{\underbrace{00...20}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...2}_{k}\right)\pi_{i0}$	$N_{\underbrace{00...21}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...2}_{k}\right)\pi_{i1}$	$N_{\underbrace{00...22}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{000...2}_{k}\right)\pi_{i2}$
….	….	….	….
$\underbrace{222....2}_{k}$	$N_{\underbrace{22...20}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{222...2}_{k}\right)\pi_{i0}$	$N_{\underbrace{22...21}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{222...2}_{k}\right)\pi_{i1}$	$N_{\underbrace{22...22}_{k+1}} = \sum_{i=1}^{n} \mathrm{I}\left(\underbrace{222...2}_{k}\right)\pi_{i2}$

The sum of posterior probabilities of each subject is the expectation of genotypic probabilities with k known markers and the QTL. *Table 7* shows the example of the sums of posterior probabilities. Based on this sums, the second-phase estimation is performed for the 2^{k+1} haplotypic probabilities. In this step, EH (Estimating Haplotype) software [10] is applied for another inner layer of EM algorithm.

Step 6: Iterate E and M-steps while the log-likelihood converges to the maximum.

3.1.6. Hypothesis Testing

The goal of this study is to test that the unknown QTL is in linkage disequilibrium with a group of k known markers. Significant LDs infer that k known markers and the unknown QTL are physically close, which could provide the guidance for subsequent biological validations. The hypothesis for the mmLD mapping can be formulated as follows:

H₀: A QTL and known markers are independent, i.e. all $D_{\mathcal{M}_1 \cdots \mathcal{M}_k, \mathcal{Q}} = 0$

H₁: At least one of equality of *H₀* is not true.

Under the null hypothesis (*H₀*), the conditional genotypic probabilities of the QTL are constant throughout subjects regardless of the marker genotypes they carry. So the parameter set under *H₀* is

$$\Theta_{H_0} = \{p_1, \dots, p_k, D_{12}, \dots, D_{1\dots k}, p_Q, \mu_0, \mu_1, \mu_2, \sigma^2\}$$

where $p_1, \dots, p_k$ indicate the allelic frequencies of *k* known markers, $D_{12}, \dots, D_{1\dots k}$ denote LDs between *k* known markers, $\mu_0, \mu_1, \mu_2, \sigma^2$ denote three means and the variance of phenotypic parameters, and p_Q represents the allelic frequency of the QTL. Under the alternative hypothesis, the parameters of LDs ($D_{1Q}, \dots, D_{1\dots kQ}$) between *k* known markers and the QTL are added. The EM algorithm to maximize the likelihood under *H₀* is similar to that under *H₁* in the previous section. Then, a likelihood ratio test statistics (LRT) can be constructed as follows:

$$LRT = -2\ell_{H_0} + 2\ell_{H_1} \quad Eqn\ 10$$

Under H_0, the LRT asymptotically follows a χ^2 - distribution with degrees of freedom to be the difference in numbers of parameters between null (H_0) and alternative hypothesis (H_1).

3.1.7. Estimating degree of freedom

Let's recall the parameters under null and alternative hypothesis;

$$\Theta_{H_0} = \{p_1, \dots, p_k, D_{12}, \dots, D_{1\dots k}, p_Q, \mu_0, \mu_1, \mu_2, \sigma^2\}$$

$$\Theta_{H_1} = \{p_1, \dots, p_k, D_{12}, \dots, D_{1\dots k}, p_Q, \mu_0, \mu_1, \mu_2, \sigma^2, D_{1Q}, \dots, D_{1\dots kQ}\}$$

In the likelihood ratio test, the degree of freedom represents the difference of parameters between the null and alternative hypothesis, and thus, the number of parameters of LDs ($D_{1Q}, \dots, D_{1\dots kQ}$) between k known markers and the QTL is set up as degrees of freedom for LRT.

The number of LDs of between k known markers and the QTL is equal to the number of combinations of k elements; $\sum_{i=1}^{k}\binom{k}{i} = 2^k - 1$. Thus, the expected degree of freedom should be $2^k - 1$ in a common situation. For example, when the numbers of known markers are two (k = 2) or three (k = 3), the expected degrees of freedom are 3 or 7, respectively. However, things may become more complicated when the number of markers

are large ($k > 3$). For example, some haplotypic frequencies may not be estimated from the data, or with frequency of zeros. In these cases, the degree of freedom of $2^k - 1$ will be inaccurate. Below I will provide a more systematic way of determining the degrees of freedom for our proposed test.

3.1.7.1. Correlation among markers is important but not directly related to the degree of freedom

Initially, we thought the correlations among markers may play important roles in determining the degree of freedom, which is best illustrated by a thought experiment. Suppose that there are two known markers and their correlation is exactly one, which means that they have identical genotypes, and the genetic information of two markers is essentially that of one marker. In this case, although there are two markers, the degree of freedom should not be three but one. Thus, it would be reasonable to hypothesize that the degree of freedom depends on the level of correlation between known markers. Several simulations have been conducted to check this hypothesis. However, the results of simulations did not reveal any relationship between the level of correlation and the reduction of degree of freedom. More simulated settings will be shown in the next section.

Although these trials did not give us what we expected, it did confirm that degree of freedom drops from 3 to 1 when the correlation changes from less than 1 to 1. Based on this fact, we further examined the reason of the reduction of degrees of freedom and

explored the hypothesis that the number of non-zero haplotypes is related to the test degree of freedom, which will be described in the section below.

3.1.7.2. Haplotypes with zero frequency

Table 8 shows the example of haplotype frequencies of two identical markers. There are four haplotype frequencies for these two markers, but among them, two haplotypes (p_{00}, p_{11}) have non-zero and the other two has zero frequencies. As described above, the haplotype frequencies of k known markers are estimated from the first phase, and then the estimation of haplotype frequencies in the second phase depends on the estimates of the first phase. In other words, the frequency of each haplotype from the first phase is divided into two frequencies by considering the bi-allelic QTL in the second phase. Therefore, the zero haplotype frequency of the first phase leads to two zero haplotype frequencies of the second phase. Thus, it means that zero frequency is not a parameter any more. Therefore, the degree of freedom of *Table 8* becomes 1 (= $2^2 - 2 - 1$) from 3 (= $2^2 - 1$).

In summary, if the number of known markers is two (k = 2) or three (k = 3), then degree of freedom has maximum 3 or 7 (=$2^k - 1$), respectively. However, if there are haplotypes with zero frequencies (say α of them), then the effective degree of freedom would become $2^k - \lambda - 1$, which is less than $2^k - 1$. The next section shows how the reduction of degree of freedom is related to the number of haplotypes with zero

frequencies, in the example of two known markers. The evaluation of type I error with the reduced degree of freedom will be shown later in the next section.

Table 8 Example of haplotype frequencies of identical two known markers

Haplotype frequencies of identical two known markers		Marker 1		
		A	a	Total
Marker 2	B	0.5 (p_{00})	0 (p_{01})	0.5 (p_B)
	b	0 (p_{10})	0.5 (p_{11})	0.5 $(q_B = 1 - p_B)$
	Total	0.5 (p_A)	0 $(q_A = 1 - p_A)$	1

Example of Reduced degree of freedom for haplotypes with zero frequency

Let's look at how to determine the degree of freedom with two known markers and one QTL. Haplotype frequencies of two known markers and a QTL are set up using the equations below. The red font denotes the parameters estimated in the second phase estimation.

Haplotype frequencies of Two known markers ($\mathcal{M}_1$ & $\mathcal{M}_2$) and one QTL($\mathcal{M}_3$)

$p_{000} = p_1p_2p_3 + p_1D_{23} + p_2D_{13} + p_3D_{12} + D_{123}$

$p_{011} = p_1q_2q_3 + p_1D_{23} - q_2D_{13} - q_3D_{12} + D_{123}$

$p_{101} = q_1p_2q_3 - q_1D_{23} + p_2D_{13} - q_3D_{12} + D_{123}$

$p_{110} = q_1q_2p_3 - q_1D_{23} - q_2D_{13} + p_3D_{12} + D_{123}$

$p_{001} = p_1p_2q_3 - p_1D_{23} - p_2D_{13} + q_3D_{12} - D_{123}$

$p_{010} = p_1q_2p_3 - p_1D_{23} + q_2D_{13} - p_3D_{12} - D_{123}$

$p_{100} = q_1p_2p_3 + q_1D_{23} - p_2D_{13} - p_3D_{12} - D_{123}$

$p_{111} = q_1q_2q_3 + q_1D_{23} + q_2D_{13} + q_3D_{12} - D_{123}$

Example 1 $\mathcal{M}_1 = \mathcal{M}_2$

⇨ $p_1 = p_2$, $q_1 = q_2$, $D_{12} = p_{00} - p_1p_2 = p_1 - p_1^2 = p_1q_1$ & $p_{010} = p_{011} = p_{100} = p_{101} = 0$

⇨ $p_{010} = p_1q_2p_3 - p_1D_{23} + q_2D_{13} - p_3D_{12} - D_{123}$

$= p_1q_1p_3 - p_1D_{13} + q_1D_{13} - p_3p_1q_1 - D_{123}$ ($D_{23} = D_{13}$, so, D_{23} is known)

$= - p_1D_{13} + D_{13} - p_1D_{13} - D_{123} = 0$

⇨ $D_{123} = (1 - 2p_1) \cdot D_{13}$ (D_{123} is known)

⇨ Finally, two parameters (p_3 & D_{13}) are left.

Example 2 One haplotypic frequency of two known markers is zero. (e.g., $p_{10} = 0$)

⇨ $p_{100} = p_{101} = 0$ & $D_{12} = p_1 p_2$

⇨ $p_{100} = q_1 p_2 p_3 + q_1 D_{23} - p_2 D_{13} - p_3 D_{12} - D_{123} = 0$

⇨ $D_{123} = q_1 p_2 p_3 + q_1 D_{23} - p_2 D_{13} - p_3 D_{12}$ (so, D_{123} is known)

⇨ Finally, three parameters (p_3, D_{13} & D_{23}) are left.

To summarize, we can determine degree of freedom as follows:

1) All haplotypes of known markers have effective non-zero frequencies.

   Degree of freedom = $2^k - 1$

2) If λ number of haplotypes have zero frequencies and the other have effective non-zero frequencies.

   Degree of freedom = $2^k - \lambda - 1$

3.1.7.3. Small haplotype frequency

In practice, we found that for large k, the genotypic data are fragmented very quickly and a few haplotypes may have very small frequencies (e.g. 1e-6). These small frequencies should be handled carefully as we do not know whether they are true zero or non-zero due to poor estimations. As the empirical experience, we found that the haplotype frequency that is larger than 1 divided by the number of subjects may serve as a useful practical cutoff for non-zero frequency.

Table 9 is the example of the differences between parameters and estimated haplotype frequencies from the first phase. 2,000 subjects were sampled in this simulated setting. Among parameters of haplotype frequencies of three known markers, there are three small haplotype frequencies (p_{000}, p_{010}, and p_{101}) which are less or equal to 1 over 2000 subjects (= 0.0005). The estimate of the haplotype p_{000} is very small but it is not zero even though the true parameter is zero. In a natural population, the genotypes of k markers are known but their haplotype frequencies are unknown. Therefore, the smallest value (3e-09), which emerges in the first phase estimation, should be determined for whether it is zero or not. If all three small frequencies are determined as non-zero values, then the deviance would follow χ^2-distribution with the full degree of freedom 7 (=$2^3 - 1$). On the other hand, if all three small frequencies are regarded as zero values, then it would follow χ^2-distribution with the reduced degree of freedom 4 (=$2^3 - 3 - 1$).

Table 9 The difference between parameters and estimates of haplotype frequencies from the first phase in a simulated sample

Parameter	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}
	0	0.2905	0.0003	0.0988	0.182	0.0005	0.1038	0.3243
Estimates from the first phase	3e-09	0.2930	0.0004	0.0961	0.1794	0.0006	0.1063	0.3243

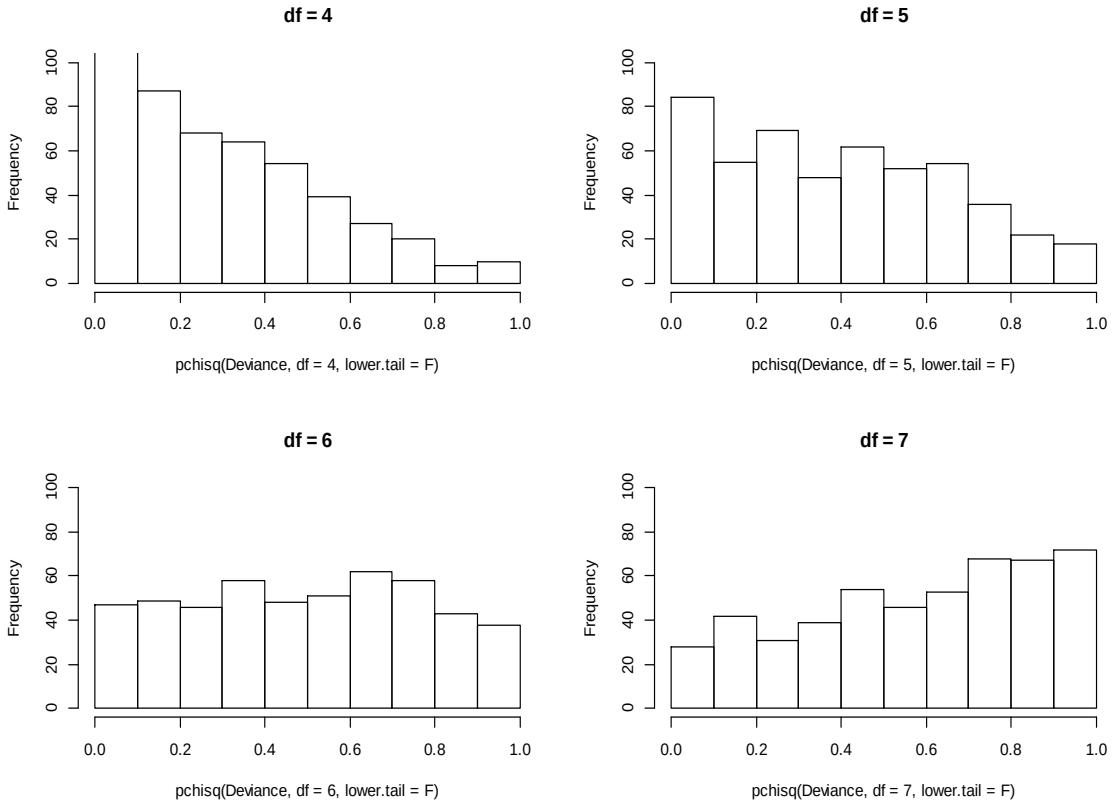

Figure 5 Type I error evaluation of the different degrees of freedom

Figure 5 shows the histograms of p-value under the null hypothesis for different degrees of freedom from 4 to 7. When they were treated as zero quantities and removed from degree of freedom, the type I error was inflated; on the other hand, when they were treated as non-zero quantities, the test was too conservative.

In order to determine the effectiveness of small haplotype frequencies, several ideas were tried such as the arithmetic or weighted average of smallest and largest degree of freedom. These simulations will be shown in the next simulation section. However, some of these trials did work well and others did not control type I error. Therefore, these ideas cannot provide stable outputs for the proper degree of freedom and they all lack some theoretical foundations.

3.1.7.4. Sequential Likelihood Ratio Test

To solve the issue with small haplotype frequencies, I propose a novel sequential likelihood ratio test procedure for the control of type I errors. The idea is analogous to the backward model selection strategy, i.e. haplotypes with small frequencies will be tested sequentially to examine if they are significantly different from zero or not, one at a time. Specifically, let L_{Full} be the likelihood under alternative hypothesis by including all non-zero haplotype frequencies, and $L_{Reduced}$ be the likelihood of setting the smallest haplotype frequency to be zero. If the LRT between L_{Full} and $L_{Reduced}$ is not significantly different under a χ^2-distribution with 1 degree of freedom, that is $-2L_{Reduced} + 2L_{Full} < \chi^2_{0.95,1}$, then it means there is no evidence that the smallest haplotype frequency is effective and thus it

should be removed from degree of freedom calculation. The procedure will be repeated until frequency of any haplotype cannot be set to zero. Once it stops, we can find the proper degrees of freedom with the control of type I error. Again, it will be tested by various settings of simulations and be discussed in the next section.

3.2. Simulations

Extensive Monte Carlo simulation experiments have been performed to examine the statistical properties of the proposed mmLD mapping method.

3.2.1. Simulated settings

Let's consider a sample of n subjects randomly chosen from a human population that is under Hardy-Weinberg Equilibrium. For the i-th subject, suppose its phenotypic value y_i is controlled by an underlying QTL, which is located at a LD block and is in linkage with a group of k markers ($k \geq 2$). The marker and QTL genotypes were first simulated based on pre-specified haplotype frequencies, and then the phenotypic values were generated based on QTL genotypes according to *Eqn 2*. The variances in phenotypic values were determined by different heritability values (H^2)[29], which quantifies the genetic contribution from the QTL to the overall trait. Specifically, $H^2 = 0$ implies that the three means of QTL genotype groups are the same, implying no QTL effect. QTL information has been removed from mmLD mapping to mimic the real scenario that QTL may be ungenotyped. Each simulated setting was performed 200, 500 or 1000 times for the evaluation of type I error and power.

3.2.2. Type I error Evaluation

In the hypothesis testing, type I error means the probability of incorrectly rejecting the null hypothesis (H_0) when the null hypothesis is true.

3.2.2.1. Relationship between degree of freedom and correlation of known markers

In this section, outputs of simulations for the trials, which were to examine the relationship between degrees of freedom and correlations of known markers, are shown.

Figure 6 and *Table 10* show the one of outputs for simulated setting that has high correlations (r $=$ 0.5, 0.9, 0.95, and 1) between two known markers. 1,000 subjects were sampled from a natural population. The major allele probabilities of both two known markers were set to be 0.5 and the linkage disequilibrium between known two markers (D_{12}) varied from 0 to 0.25. If D_{12} is zero, then their correlation is zero. Meanwhile, if D_{12} is 0.25, then their correlation is one. The phenotypic values follow the mixture Gaussian normal distributions with different three means and the same variance; $\mu_0 = 20, \mu_1 = 40, \mu_2 = 60$ and $\sigma^2 = 49$.

As shown in *Table 10*, the proper degree of freedom of each simulation is $2^2 - 1 = 3$, except the last setting (D_{12} = 0.25, r = 1), in which the proper degree of freedom was 1. Particularly, in the fourth simulation, even though it had high correlation (D_{12}=0.2375, r=0.95), its proper degree of freedom was still around 3.

Therefore, the results of simulations did not reveal any relationship between the level of correlation and the reduction of degree of freedom. More simulated settings are shown below (*Table 11 – Table 13*, *Figure 6 - Figure 9*).

Table 10 Mean and Variance of Deviance for the evaluation of the proper degree of freedom

σ^2=49	D_{12} = 0 / r = 0	D_{12} = 0.125 / r = 0.5	D_{12} = 0.225 / r = 0.9	D_{12} = 0.2375 / r = 0.95	D_{12} = 0.25 / r = 1
Mean	3.1	2.9	3.4	3.3	0.9
Variance	6.5	5.1	7.2	7.4	1.5

Table 11 Mean and Variance of Deviance for the evaluation of the proper degree of freedom with $\sigma^2 = 9$

$\sigma^2 = 9$	D_{12} = 0 / r = 0	D_{12} = 0.125 / r = 0.5	D_{12} = 0.225 / r = 0.9	D_{12} = 0.2375 / r = 0.95	D_{12} = 0.25 / r = 1
Mean	3.1	3.2	3.0	3.0	1.2
Variance	6.9	7.6	5.8	5.9	2.3

Table 12 Mean and Variance of Deviance for the evaluation of the proper degree of freedom with $\sigma^2 = 144$

$\sigma^2 = 144$	D_{12} = 0 / r = 0	D_{12} = 0.125 / r = 0.5	D_{12} = 0.225 / r = 0.9	D_{12} = 0.2375 / r = 0.95	D_{12} = 0.25 / r = 1
Mean	3.2	3.2	3.1	3.3	1.1
Variance	6.0	6.3	6.7	6.5	2.4

Table 13 Mean and Variance of Deviance for the evaluation of the proper degree of freedom with $\sigma^2 = 400$

$\sigma^2 = 400$	D_{12} = 0 / r = 0	D_{12} = 0.125 / r = 0.5	D_{12} = 0.225 / r = 0.9	D_{12} = 0.2375 / r = 0.95	D_{12} = 0.25 / r = 1
Mean	3.5	3.3	3.5	3.5	1.3
Variance	7.8	7.5	8.4	7.9	3.6

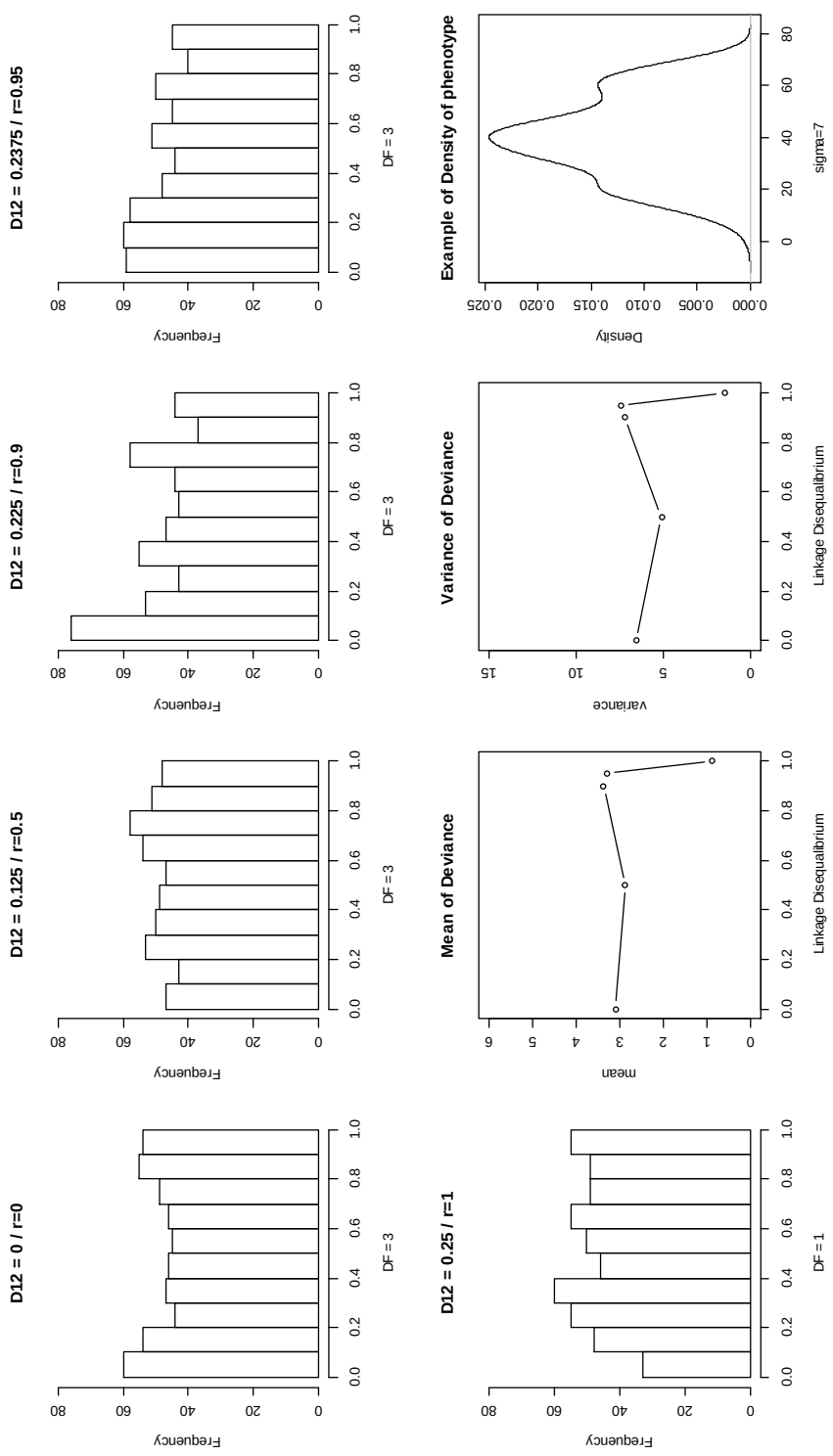

Figure 6 Output of type I error for the high correlation between known two markers with $\sigma^2 = 49$

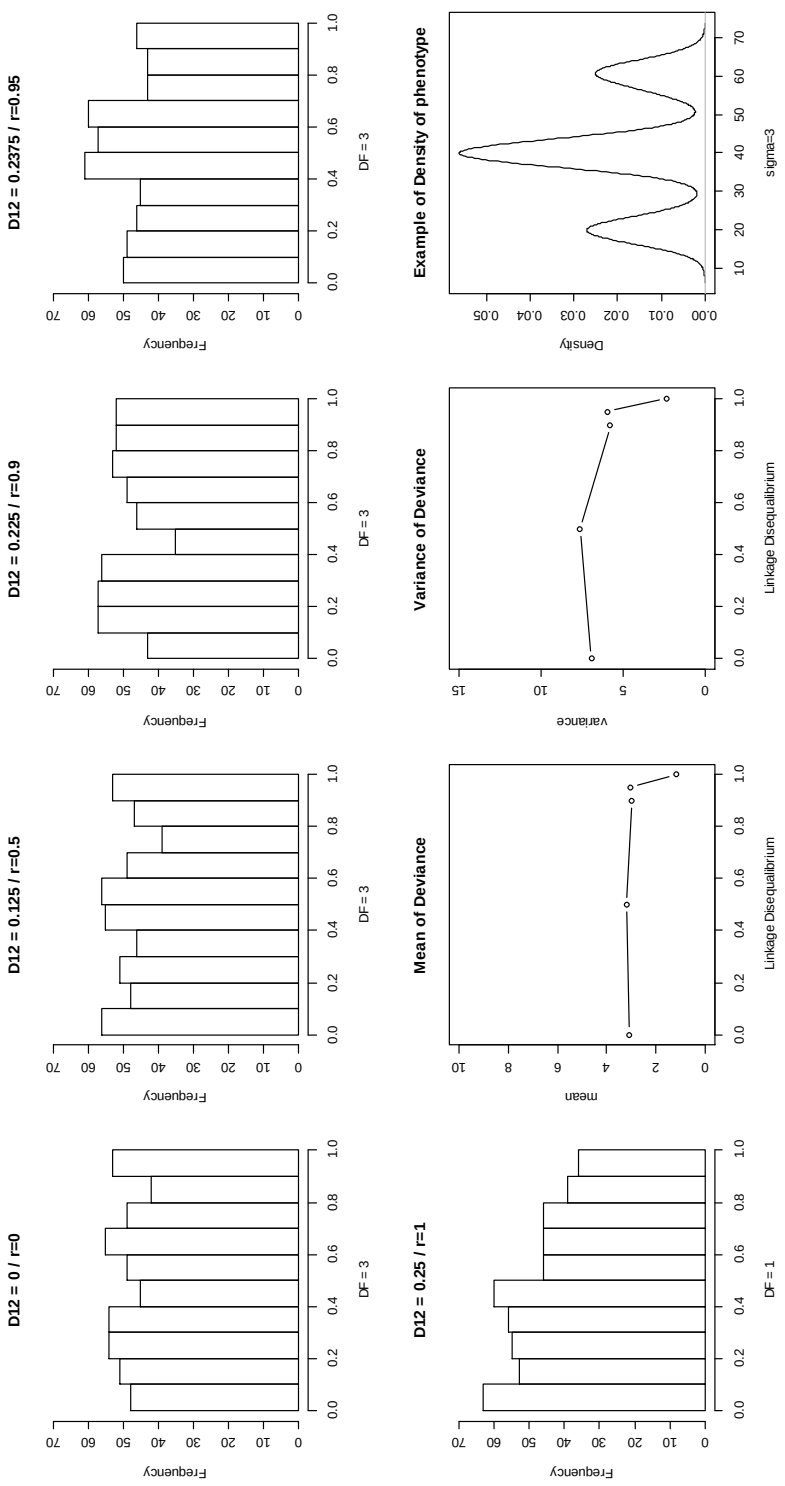

Figure 7 Output of type I error for the high correlation between known markers with $\sigma^2 = 9$

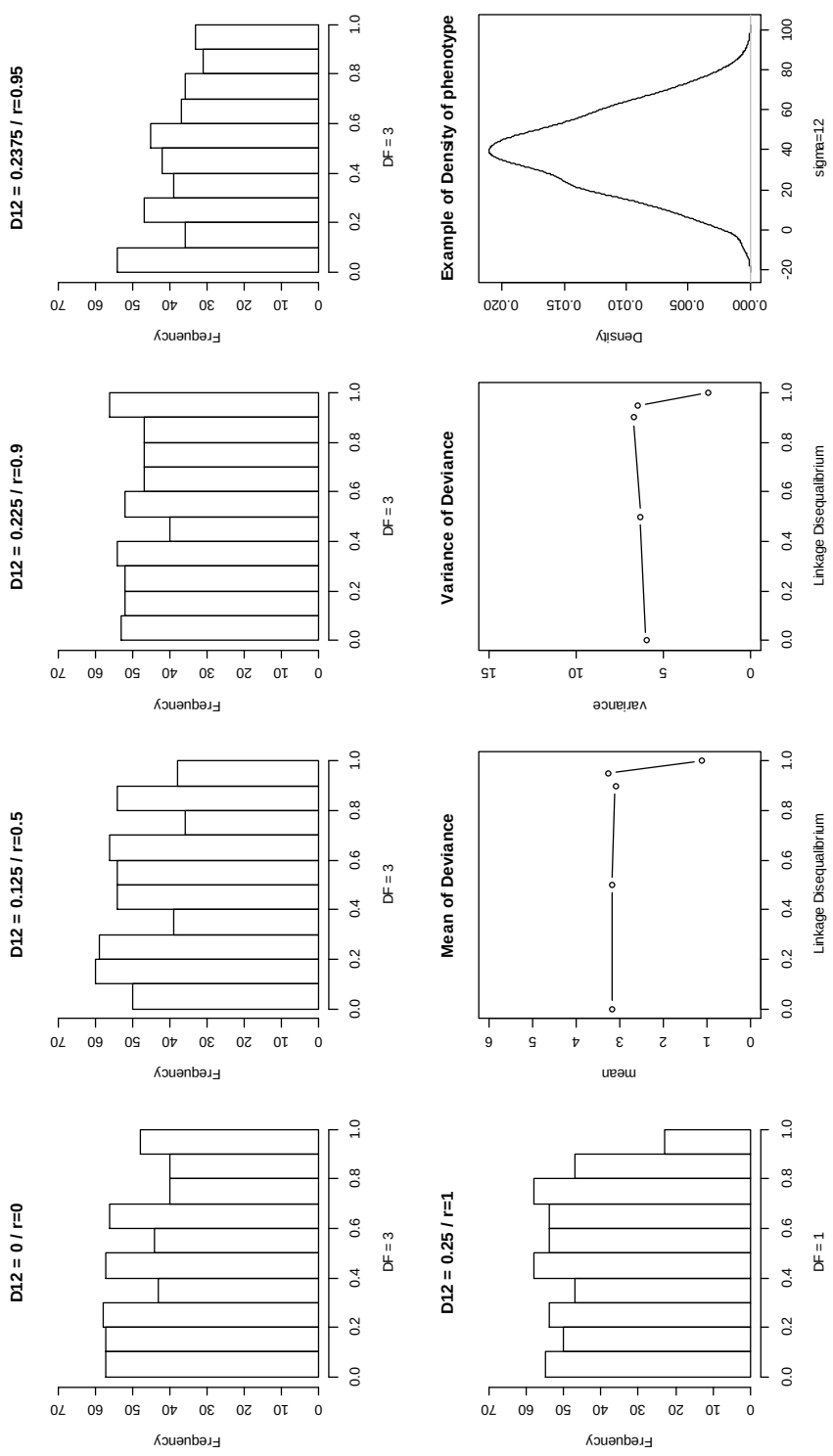

Figure 8 Output of type I error for the high correlation between known markers with $\sigma^2 = 144$

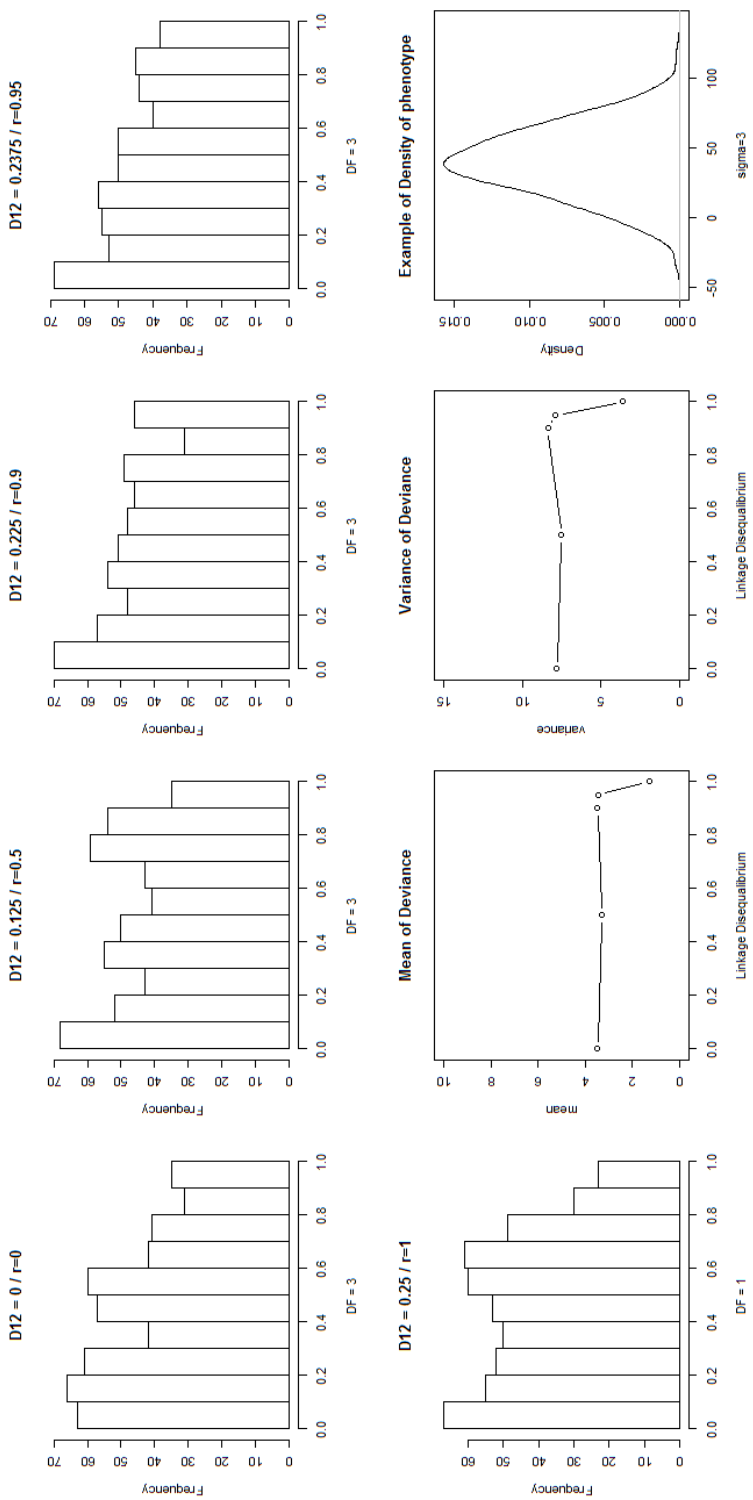

Figure 9 Output of type I error for the high correlation between known markers with $\sigma^2 = 400$

3.2.2.2. Evaluating degrees of freedom in the presence of haplotypes with zero frequencies

Based on the extensive Monte Carlo simulation, 100, 500, 1000 and 2000 subjects were selected from a natural population. Suppose that their continuous phenotypes follow a mixture Gaussian distribution with the variance of 5 ($\sigma^2 = 5$) and three different means; $\mu_0 = 20, \mu_1 = 40,$ and $\mu_2 = 60$. For type I error evaluation, the null hypothesis is set to be true. In other words, there is no linkage disequilibrium between two or three known markers and the QTL. Four different scenarios were conducted and their parameters of haplotype frequencies for two or three known markers are shown in *Table 14* and *Table 15*. Because the estimated haplotypes for two markers have two zero frequencies in *Table 14*, 2 degrees of freedom were applied to evaluate type I error. Meanwhile, in the *Table 15*, the 3rd and 4th scenarios of three markers have two and one zero frequencies respectively, 6 and 5 degrees of freedom were applied. *Figure 10* shows that the mmLD controls type I error (0.05) with the reduced degree of freedom and *Figure 11* shows the histogram of type I error evaluation.

Table 14 Parameters of haplotype frequencies of two known markers

Simulated Haplotype frequencies	p_{00}	p_{01}	p_{10}	p_{11}	**Applied Degree of freedom**
Two markers 1	0.26	0.24	0	0.50	2
Two markers 2	0	0.19	0.42	0.39	2

Table 15 Parameters of haplotype frequencies of three known markers

Simulated Haplotype frequencies	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}	**Applied Degree of freedom**
Three markers 1	0	0	0.124	0.08	0.207	0.18	0.169	0.222	5
Three markers 2	0.144	0.159	0.124	0.149	0	0.08	0.025	0.319	6

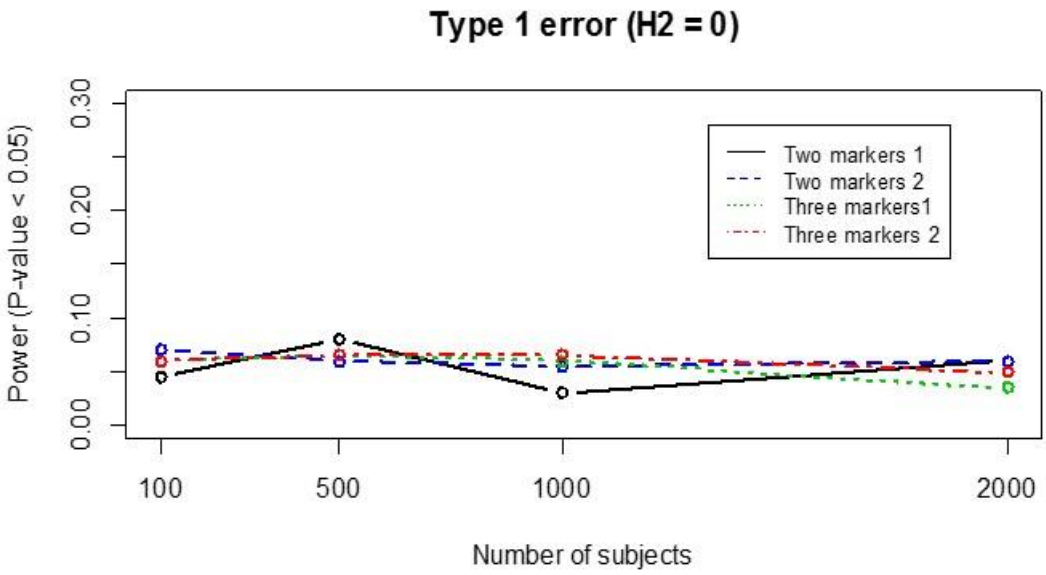

Figure 10 Type I error evaluation for reduced degree of freedom based on zero haplotype frequencies

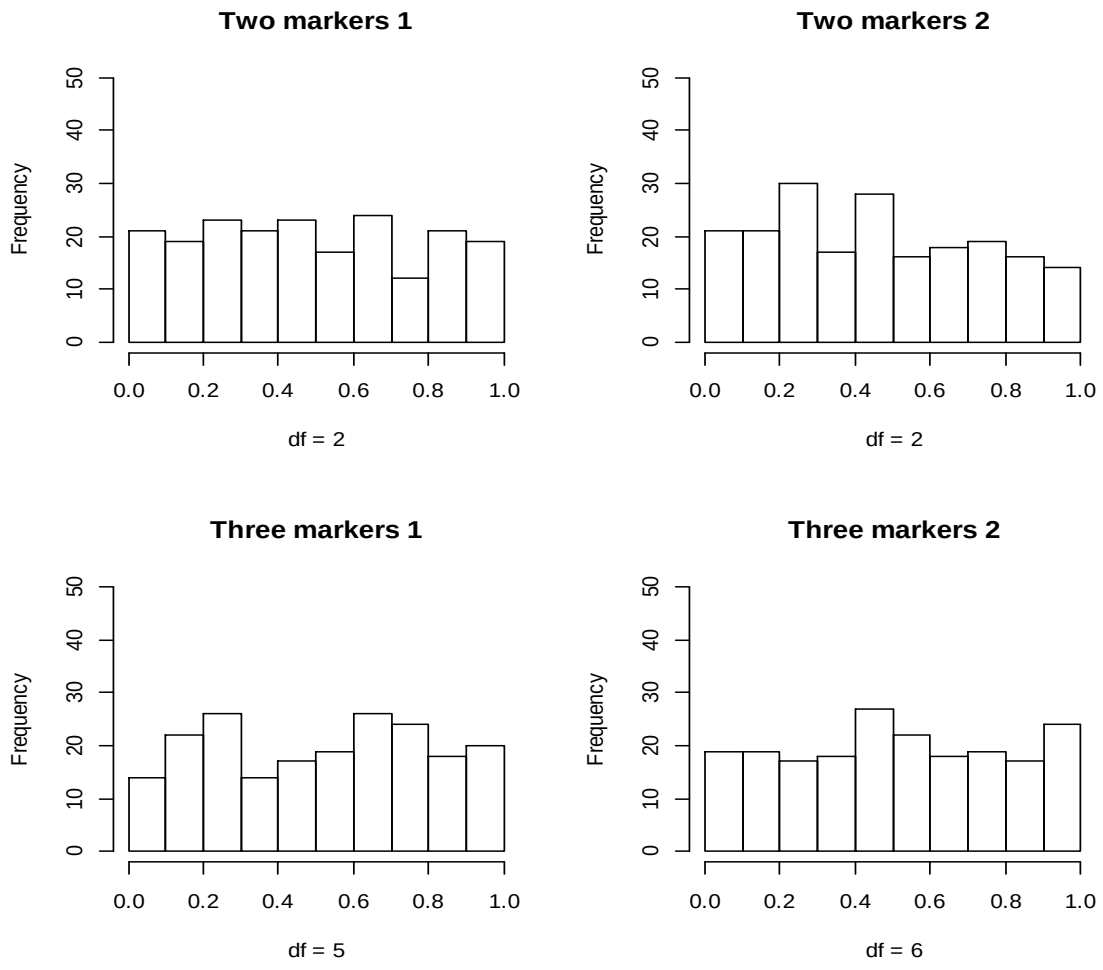

Figure 11 Histogram of Type I error evaluation for reduced degree of freedom with zero haplotype frequencies (2000 subjects)

3.2.2.3. Small haplotype frequencies

When the number of markers, k, is large, in practice, a few very small haplotype frequencies (e.g. 1e-6) may emerge. To account for these values, we initially tried to calculate an average or weighted average to find out a proper degree of freedom. As

shown in the following histograms of type I errors under various simulated settings, the ad hoc method of average did not perform well. The outputs of the arithmetic and weighted average of degree of freedom are shown from Simulation 1 to Simulation 3 below.

Nevertheless, these results are provided as our attempt to solve the problem and motivated us to come up with more powerful sequential LRT method, which was introduced in the section 3.1.7.4.

Simulation 1—Average or weighted average of degree of freedom

Table 16 Discrepancies of haplotype frequencies between parameters and estimates (Simulation 1)

Parameter	p_{0000}	p_{0001}	p_{0010}	p_{0011}	p_{0100}	p_{0101}	p_{0110}	p_{0111}
	0.0003	0.055	0.1753	0	0.0713	0.1363	0.0963	0
	p_{1000}	p_{1001}	p_{1010}	p_{1011}	p_{1100}	p_{1101}	p_{1110}	p_{1111}
	0.2058	0.0573	0	0	0.0005	0.008	0.1913	0.003
Estimates	p_{0000}	p_{0001}	p_{0010}	p_{0011}	p_{0100}	p_{0101}	p_{0110}	p_{0111}
	0.0005	0.0556	0.1753	2.1e-09	0.0716	0.1371	0.0941	2.8e-10
	p_{1000}	p_{1001}	p_{1010}	p_{1011}	p_{1100}	p_{1101}	p_{1110}	p_{1111}
	0.2054	0.0567	0	0	0.0005	0.0069	0.1931	0.0032

Table 17 Comparison of Type I error with arithmetic and weighed average of degree of freedom (Simulation 1)

p(hap) : Haplotype frequency ; ***C*** = 1 / # of subject ; ***J = 1/C*** = # of subject	d.f.	Type I error	K-S statistics with uniform dist.	P-value
$\boldsymbol{I}(p(\text{hap}) > \boldsymbol{C})$	9	0.106	0.21	< 0.0001
Average of 1 & 4	10	0.068	0.11	< 0.0001
$\boldsymbol{I}(p(\text{hap}) > \boldsymbol{C}) + \sum I(p(\text{hap}) < \boldsymbol{C}) \times p(hap) \times \boldsymbol{J}$	10.9	0.048	0.03	0.58
$\boldsymbol{I}\left(p(\text{haplotype}) > {}^{1}/_{\boldsymbol{C}}\right)$	11	0.048	0.03	0.59

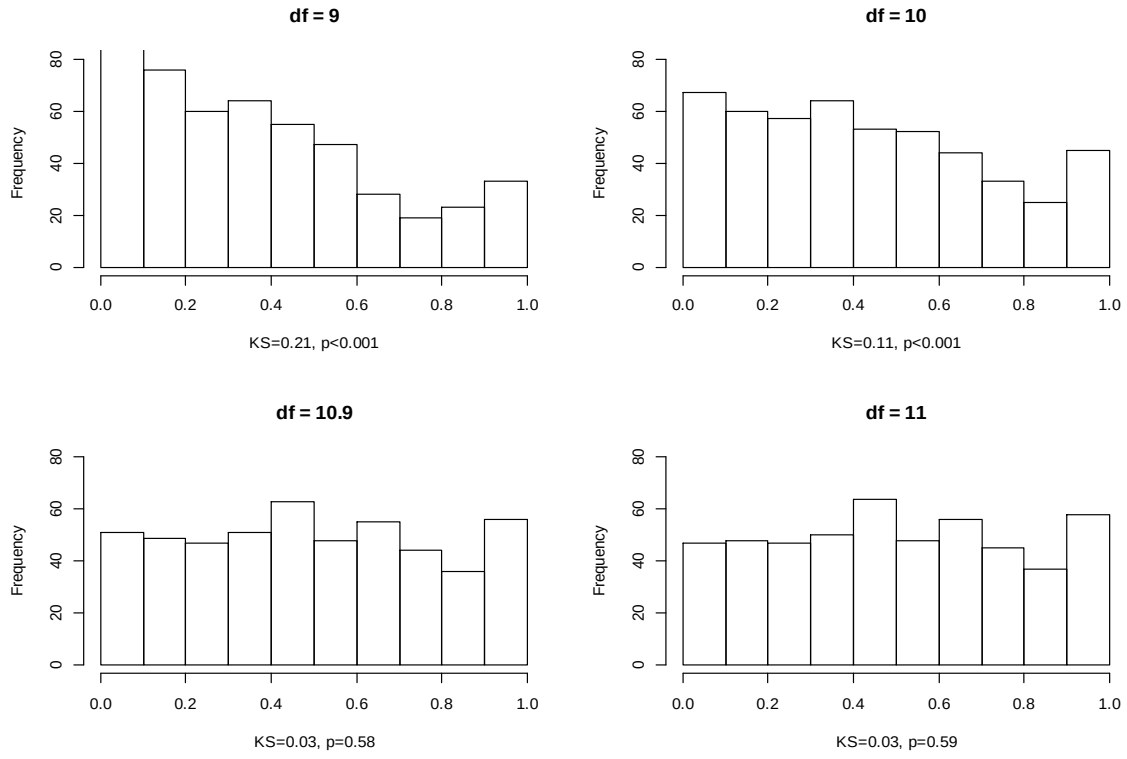

Figure 12 Distributions of Type I error of arithmetic and weighed average of degree of freedom (Simulation 1)

Simulation 2—Average or weighted average of degree of freedom

Table 18 Discrepancies of haplotype frequencies between parameters and estimates (Simulation 2)

Parameters	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}
	0	0.2905	0.0003	0.0988	0.182	0.0005	0.1038	0.3243
Estimates	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}
	2.6e-09	0.2930	0.0004	0.0961	0.1794	0.0006	0.1063	0.3243

Table 19 Comparison of Type I error with arithmetic and weighed average of degree of freedom (Simulation 2)

p(hap) : Haplotype frequency ; ***C*** = 1 / # of subject ; ***J* = *1/C*** = # of subject	d.f.	Type I error	K-S statistics with uniform dist.	P-value
$\boldsymbol{I}(p(\text{hap}) > \boldsymbol{C})$	5	0.088	0.15	< 0.0001
Average of 1 & 4	5.5	0.062	0.09	0.0005
$\boldsymbol{I}(p(\text{hap}) > \boldsymbol{C}) + \sum I(p(\text{hap}) < \boldsymbol{C}) \times p(hap) \times \boldsymbol{J}$	5.8	0.052	0.06	0.06
$\boldsymbol{I}\left(p(\text{haplotype}) > {}^{1}/_{\boldsymbol{C}}\right)$	6	0.05	0.04	0.39

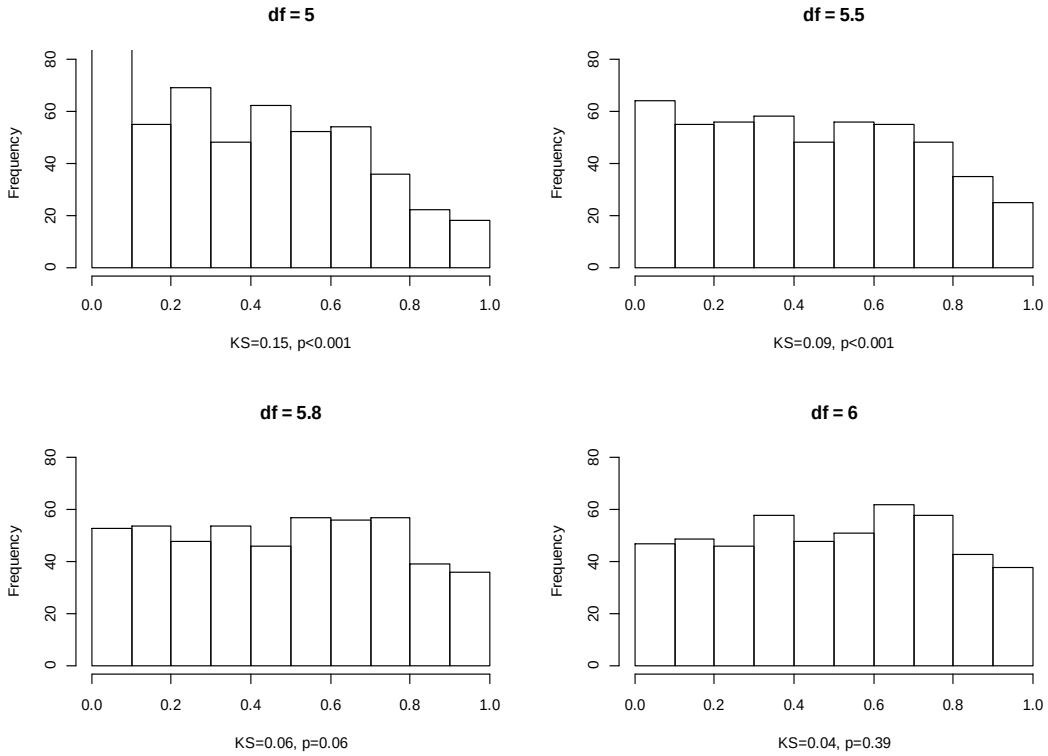

Figure 13 Distributions of Type I error of arithmetic and weighed average of degree of freedom (Simulation 2)

Simulation 3—Average or weighted average of degrees of freedom

Table 20 Discrepancies of haplotype frequencies between parameters and estimates (Simulation 3)

Parameters	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}
	0.0005	0.0005	0.001	0.2675	0.184	0.0845	0.1015	0.3605
Estimates	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}
	0.0007	0	0.0001	0.273	0.1868	0.0853	0.0985	0.3557

Table 21 Comparison of Type I error with arithmetic and weighed average of degree of freedom (Simulation 3)

p(hap) : Haplotype frequency ; ***C*** = 1 / # of subject ; ***J = 1/C*** = # of subject	d.f.	Type I error	K-S statistics with uniform dist.	P-value
$\boldsymbol{I}(p(\text{hap}) > \boldsymbol{C})$	5	0.046	0.028	0.82
Average of 1 & 4	5.5	0.024	0.09	0.0007
$\boldsymbol{I}(p(\text{hap}) > \boldsymbol{C}) + \sum I(p(\text{hap}) < \boldsymbol{C}) \times p(hap) \times \boldsymbol{J}$	5.14	0.038	0.04	0.29
$\boldsymbol{I}\left(p(\text{haplotype}) > {}^{1}/_{\boldsymbol{C}}\right)$	6	0.016	0.15	< 0.0001

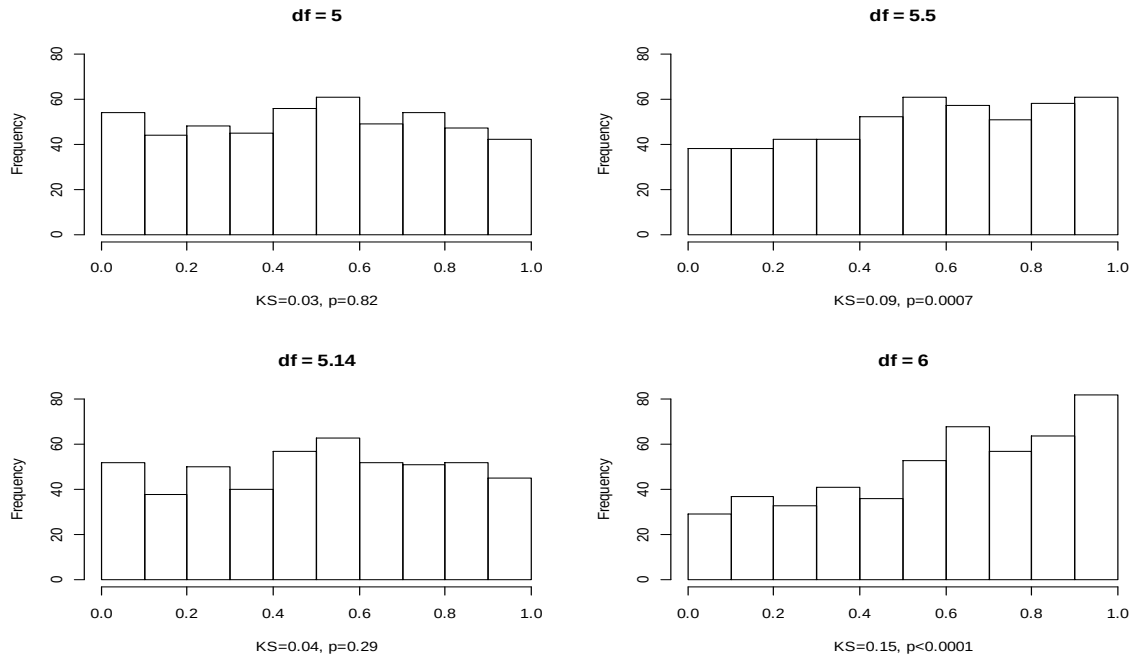

Figure 14 Distributions of Type I error of arithmetic and weighed average of degree of freedom (Simulation 3)

3.2.2.4. Sequential LRT for small haplotype frequencies

Although the idea of calculating the degrees of freedom with weighted averages did not work, the results motivated us to come up with more powerful sequential LRT method. Four different simulations were conducted to evaluate the applicability of the sequential LRT. Results show that under various scenarios, the sequential LRT procedure can control type I error well.

Simulation 1—Sequential LRT

The parameters for this simulation are given in *Table 22*, in which the frequency of one haplotype, p_{000} is set to be zero, and the frequencies of another two haplotypes (p_{010} and p_{110}) are set to be small, 0.0005 and 0.0003, respectively. The estimates for p_{000}, p_{010}, and p_{110} are 0, 0.0003 and 1.3e-7, respectively. It is clear that p_{000} is not estimable and therefore it does not account for any degree of freedom. However, if p_{010} and p_{110} are ineffective, i.e., they are treated as zero frequencies, the degree of freedom would be 4, whereas maximum degree of freedom is 6 if they are treated as non-zero frequencies.

Table 23 and *Figure 15* show type I error evaluation and the goodness of fit test (Kolmogorov-Smirnov statistics) of p-value under null hypothesis. When the degree of freedom is fixed at either 4, 5 or 6, the distributions of the p values were skewed. On the other hand, sequential LRT adaptively selects degrees of freedom based on data, which varies between 4 and 5 for the 1000 simulations. The sequential LRT procedure shows not only appropriate type I error evaluation but also uniform distribution of p-value.

Therefore, we can see that sequential LRT performs well in practical simulations, too. The reason lies in the fact that if the frequency of one haplotype is small in the general population, subjects carrying such haplotype might not be picked by sampling, in which the effective haplotype frequency in a specific sample might be truly zero. Therefore, the degrees of freedom should vary from simulation to simulation.

Table 22 Discrepancies of haplotype frequencies between parameters and estimates for the simulation 1- Sequential LRT

	Haplotype frequencies for known markers							
	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}
Parameters	0	0.3192	0.0005	0.12	0.18	0.15	0.0003	0.23
Estimates *	0	0.3242	0.0003	0.1075	0.1895	0.1496	1.3e-07	0.2289

* Estimated by one simulation of the 1st phase EM algorithm

Table 23 Comparison of Type I error between fixed degree of freedom and sequential LRT for the simulation 1- Sequential LRT

	d.f.	Type I error	Goodness of fit for uniform distribution	
			K-S statistics	p-value
Fixed d.f.	4	0.063	0.1158	< .0001
	5	0.035	0.0652	0.0004
	6	0.016	0.1926	< .0001
Sequential LRT	Data-adapted	0.051	0.0318	0.265

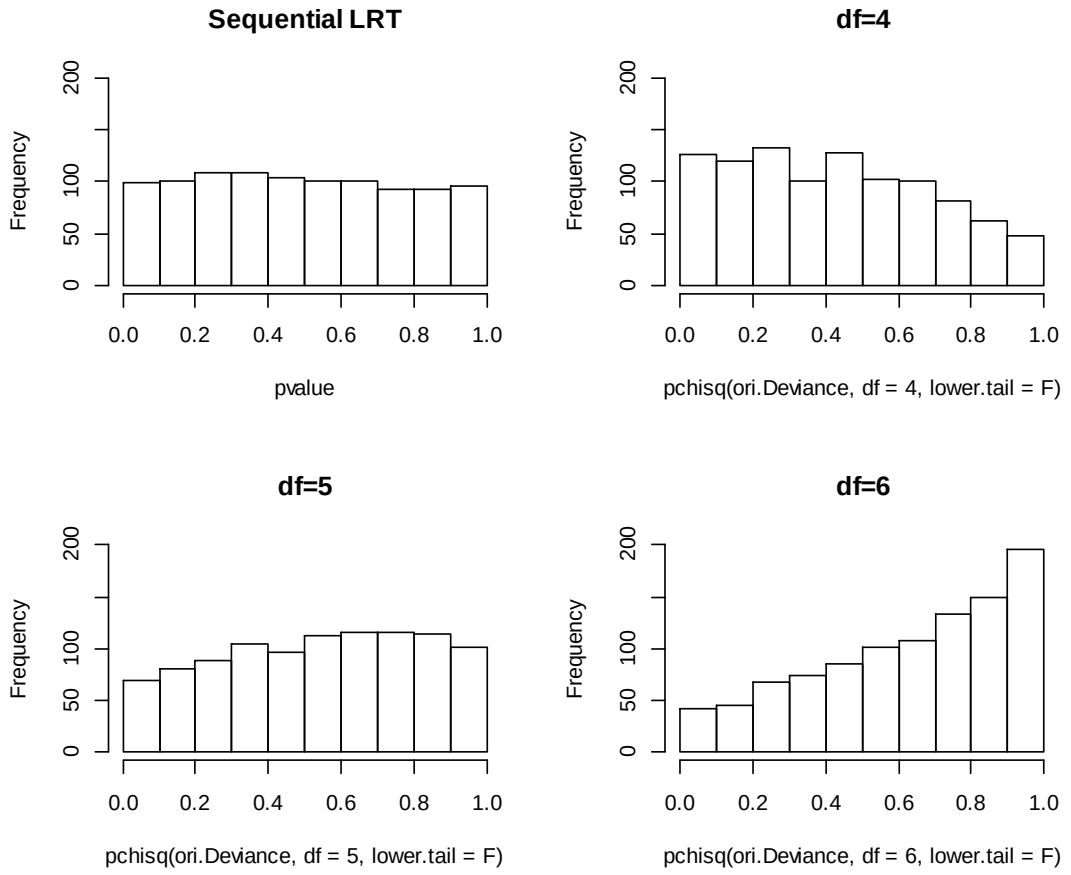

Figure 15 Distributions of Type I error of fixed degree of freedom and sequential LRT for the simulation 1- Sequential LRT

Simulation 2—Sequential LRT

The second simulation was also performed with two known markers. The values of true parameters and the estimates of haplotype frequency from the first phase are given in *Table 24*. As shown in *Table 25* and *Figure 16*, test with fixed degree of freedom of 6 can control type I error well this time. More importantly, the sequential LRT procedure can

also show good control of type I error, demonstrating its consistent performance under a different setting.

Table 24 Discrepancies of haplotype frequencies between parameters and estimates for the simulation 2- Sequential LRT

	Haplotype frequencies for known markers							
	p_{000}	p_{001}	p_{010}	p_{011}	p_{100}	p_{101}	p_{110}	p_{111}
Parameters	0	0.2905	0.00025	0.09875	0.182	0.0005	0.10375	0.32425
Estimates *	2.6e-09	0.29301	0.00039	0.0961	0.17936	0.00063	0.10625	0.32426

* Estimated by one simulation of the 1st phase EM algorithm

Table 25 Comparison of Type I error between fixed degree of freedom and sequential LRT for the simulation 2- Sequential LRT

	d.f.	Type I error	Goodness of fit for uniform distribution	
			K-S statistics	p-value
Fixed d.f.	5	0.088	0.1505	< .0001
	6	0.05	0.0404	0.3874
	7	0.024	0.1333	< .0001
Sequential LRT	Data-adapted	0.05	0.0325	0.6656

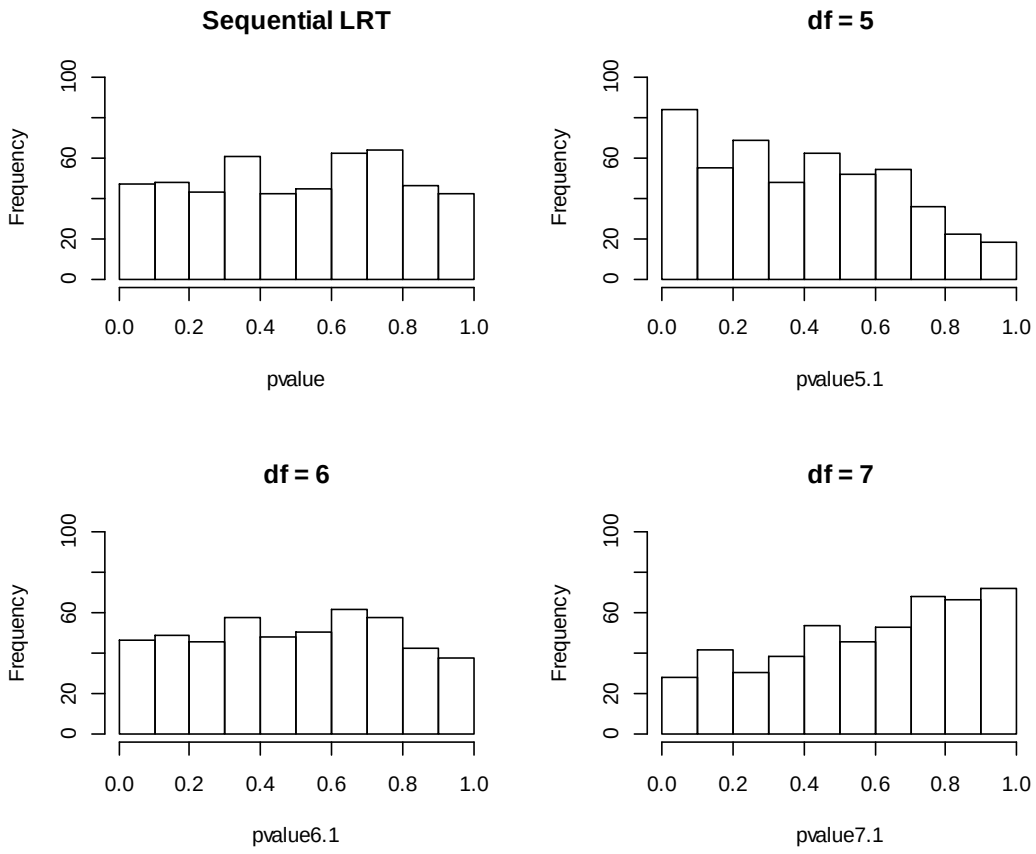

Figure 16 Distributions of Type I error of fixed degree of freedom and sequential LRT for the simulation 2- Sequential LRT

Simulation 3—Sequential LRT

In this simulation, the sequential LRT procedure is examined for three known markers. In true parameters, 4 out of 16 haplotype frequencies are set to be zero frequencies and another two are given small values (*Table 26*). Similar to previous simulations, some estimate for thesis parameters are zeros and some are very small

quantities. As shown in *Table 27* and *Figure 17*, both sequential LRT and the fixed 11 degree of freedom show good controls of type I error. Therefore, the sequential LRT works well for the three known markers, too.

Table 26 Discrepancies of haplotype frequencies between parameters and estimates for the simulation 3- Sequential LRT

	Haplotype frequencies for known markers							
Parameters	p_{0000}	p_{0001}	p_{0010}	p_{0011}	p_{0100}	p_{0101}	p_{0110}	p_{0111}
	0.00025	0.055	0.17525	0	0.07125	0.13625	0.09625	0
	p_{1000}	p_{1001}	p_{1010}	p_{1011}	p_{1100}	p_{1101}	p_{1110}	p_{1111}
	0.20575	0.05725	0	0	0.0005	0.008	0.19125	0.003
Estimates *	p_{0000}	p_{0001}	p_{0010}	p_{0011}	p_{0100}	p_{0101}	p_{0110}	p_{0111}
	0.00049	0.05561	0.1753	2.1e-09	0.07161	0.1371	0.09414	2.8e-10
	p_{1000}	p_{1001}	p_{1010}	p_{1011}	p_{1100}	p_{1101}	p_{1110}	p_{1111}
	0.20538	0.05671	0	0	0.00046	0.00688	0.19311	0.00319

* Estimated by one simulation of the 1st phase EM algorithm

Table 27 Comparison of Type I error between fixed degree of freedom and sequential LRT for the simulation 3- Sequential LRT

	d.f.	Type I error	Goodness of fit for uniform distribution	
			K-S statistics	p-value
Fixed d.f.	9	0.106	0.2066	< .0001
	10	0.068	0.1114	< .0001
	11	0.048	0.0344	0.5939
	12	0.038	0.1042	< .0001
Sequential LRT	Data-adapted	0.048	0.0425	0.3284

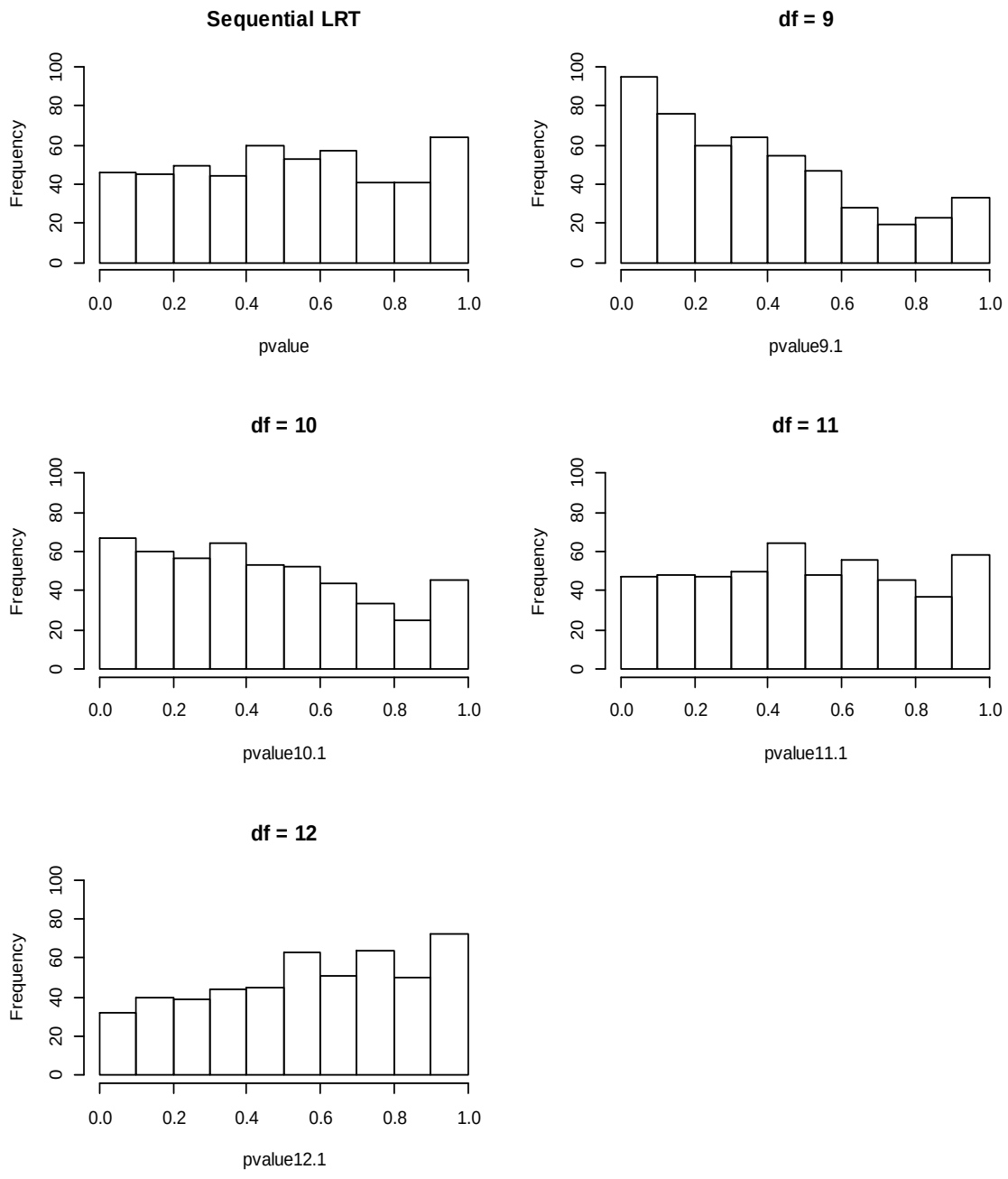

Figure 17 Distributions of Type I error of fixed degree of freedom and sequential LRT for the simulation 3- Sequential LRT

3.2.2.5. Multiple-testing issue of the sequential LRT

The sequential LRT iterates until the difference of maximum likelihoods between the full and reduced model is rejected. Since several tests will be conducted during this procedure, the multiple-testing issue may occur. To adjust this issue, we applied the conservative method that considers the maximum times of the iterations which is determined by the number of haplotype frequencies less than 1 divided by the number of subjects (Bonferroni correction).

The comparison of normal sequential LRT and multiple-adjusted sequential LRT with Bonferroni correction has been conducted. The simulated setting here is set the same as that in *Table 22*. As shown in *Table 28* and *Figure 18*, the result of adjusted sequential LRT by Bonferroni correction is similar to that of normal sequential LRT. Therefore, although multiple-testing issue is a theoretical concern, in practice it does not seem to pose problems for our proposed LRT procedure.

Table 28 Comparison of Type I error between original sequential LRT and Adjusted sequential LRT for the simulation 1

	Type I error	Goodness of fit for uniform distribution	
		K-S statistics	p-value
Normal Sequential LRT	0.051	0.0318	0.265
Adjusted Sequential LRT	0.051	0.0384	0.104

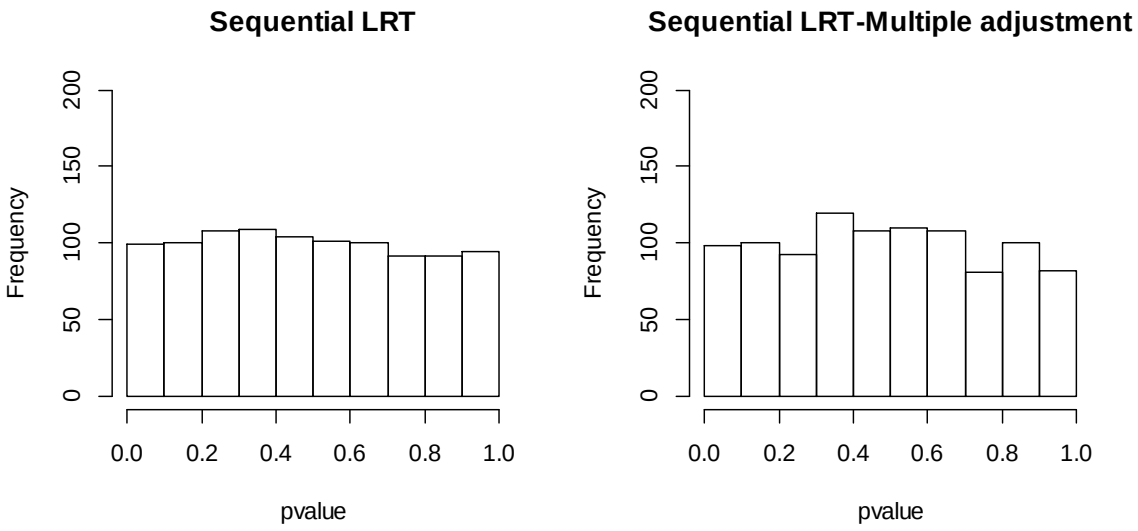

Figure 18 Distributions of Type I error of normal sequential LRT and Adjusted sequential LRT for the simulation 1

3.2.3. Power comparison between smLD, minimum p-value smLD, SKAT_C and mmLD

Next, we would like to check the power of the mmLD mapping method, which is the probability of correctly detecting the existence of a QTL when there is indeed the QTL effect. Two scenarios will be checked here: (1) The QTL is assumed to be located between adjacent *k* markers (QTL is not genotyped); or (2) already genotyped as one of the known markers (*Figure 19*). Powers were examined separately for these two scenarios. Additionally, the mmLD will be compared with other methods that can handle multiple markers, such as the adjusted single marker LD test (smLD), and SKAT_C [19, 20, 30].

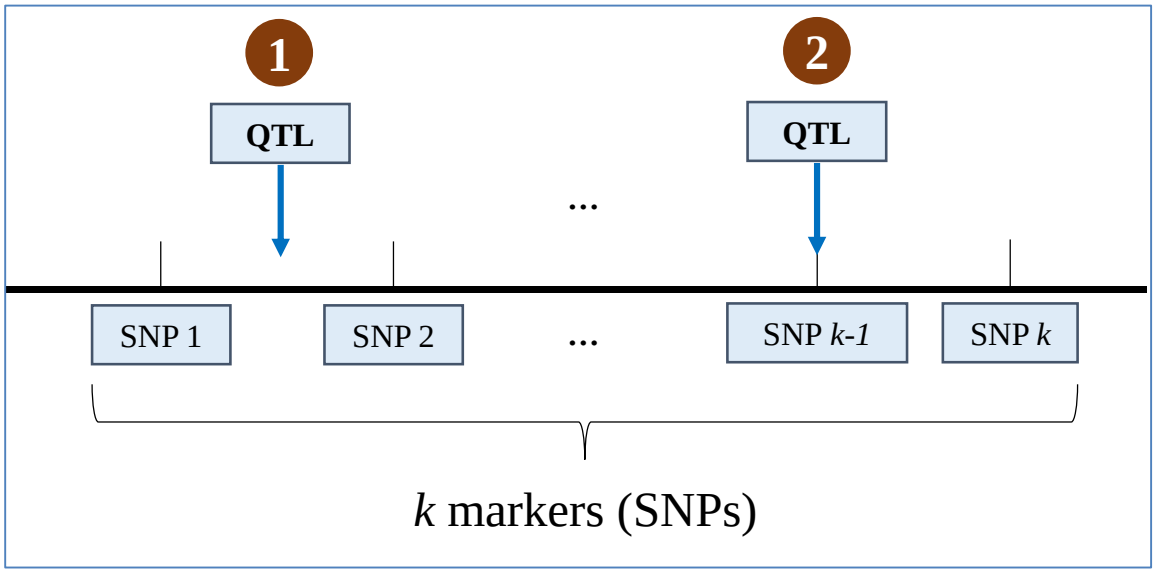

Figure 19 Example of simulated setting of k markers and one QTL

In this case, the phenotypes caused by three QTL genotypes were simulated based on *Eqn 2*, with μ_0 = 10, μ_1 = 5, and μ_2 = 0. Again, the variances in phenotypic values

were calculated based on different heritability values (H^2) [29]. Based on these parameters and designs, the power performances were conducted by different sample size (n = 100, 500, 1000, and 2000) and different heritability values (H^2 = 0.05, 0.1, and 0.2). Each simulated setting was performed 200 or 500 times for the power performance.

Scenario 1: QTL is not genotyped

Scenario 1-1. Two known markers

The two-marker LD mapping (tmLD) has already been studied [18]. Prior to extending to multiple markers (> 2), we conducted simulations to verify that the sequential LRT procedure indeed works for two markers, serving as a validation for our new framework. The LD between the two known markers is set to be 0.04 ($D_{12} = 0.04$). The other simulated settings are the same as the simulations for type I error.

Table 29 and *Figure 20* show the power comparison between the tmLD and smLD. As expected, the tmLD has higher powers under small heritability values (H^2 = 0.05, 0.1, and 0.2).

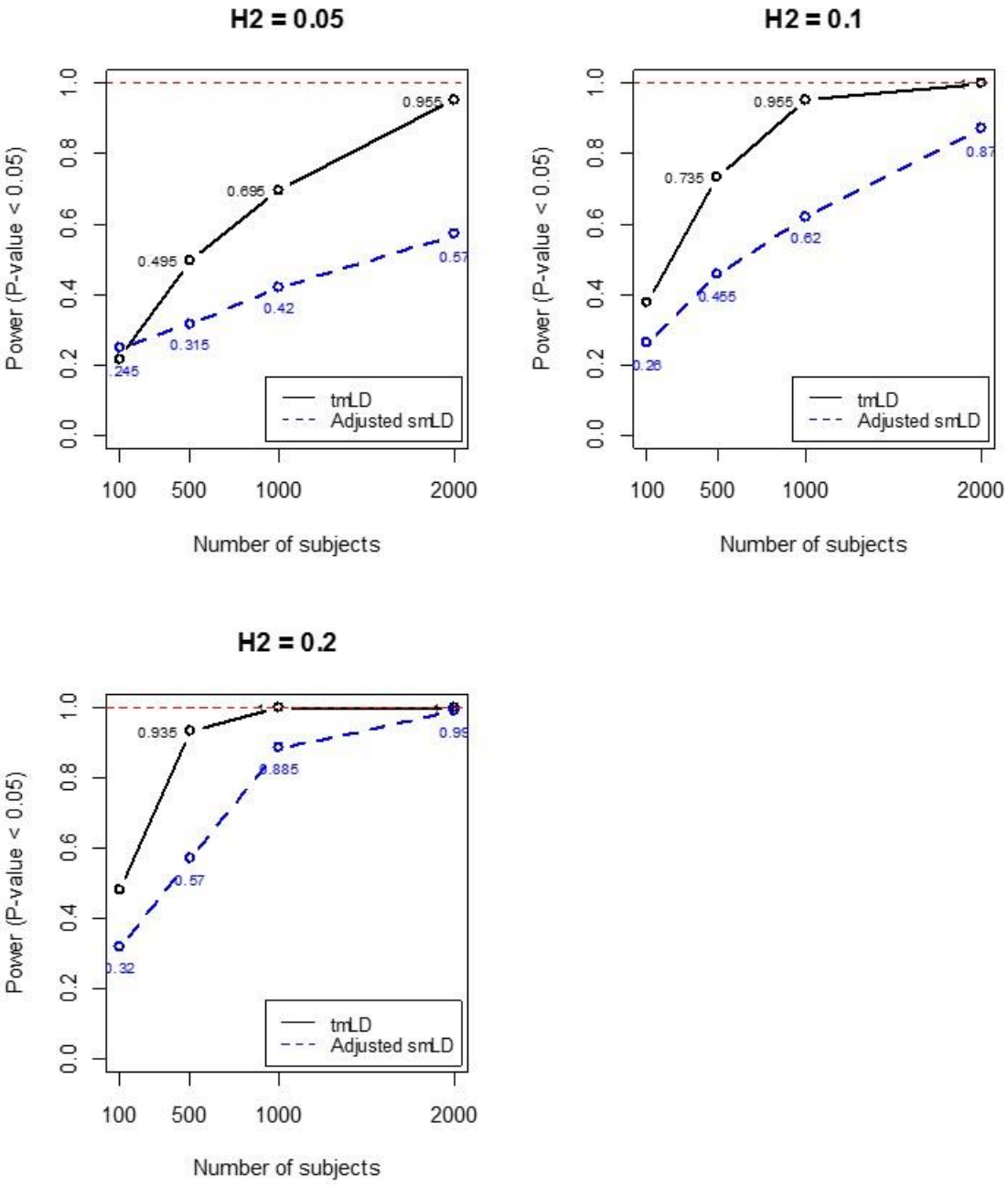

Figure 20 Power comparison of tmLD and smLD for Scenario (1); 2 known markers

Table 29 Power comparison of tmLD and smLD for Scenario (1); 2 known markers

Number of subjects	H^2= 0.05		H^2= 0.1		H^2= 0.2	
	tmLD	Adjusted smLD	tmLD	Adjusted smLD	tmLD	Adjusted smLD
100	0.215	0.245	0.375	0.260	0.480	0.320
500	0.495	0.315	0.735	0.455	0.935	0.570
1000	0.695	0.420	0.955	0.620	1	0.885
2000	0.955	0.570	1	0.870	1	1

Scenario 1-2. Three known markers

In this simulation, three known markers are considered here and the parameters are given in *Table 30*. As shown in *Figure 21* and *Table 31*, power comparison of different models with a sequence of sample sizes has been conducted. The black solid line indicates the power of the mmLD and the pink dot, blue dot, and red dot lines indicate the powers of the smLD, minimum p-value smLD, and SKAT_C, respectively. The mmLD demonstrated much higher power than those of the SKAT_C, smLD or minimum p-value smLD. *Table 32* shows the estimates for the true values. It indicates the mean and standard errors of the estimates. As shown in *Table 32*, estimates become close to true values as the larger sample size and higher heritability value are applied.

Table 30 Parameters of haplotype frequencies for Scenario (1) with 3 known markers.

p_{000Q}	p_{000q}	p_{001Q}	p_{001q}	p_{010Q}	p_{010q}	p_{011Q}	p_{011q}
0	0.16	0.13	0.04	0	0	0.1	0
p_{100Q}	p_{100q}	p_{101Q}	p_{101q}	p_{110Q}	p_{110q}	p_{111Q}	p_{111q}

0.13	0.2	0	0	0.08	0.08	0	0.08

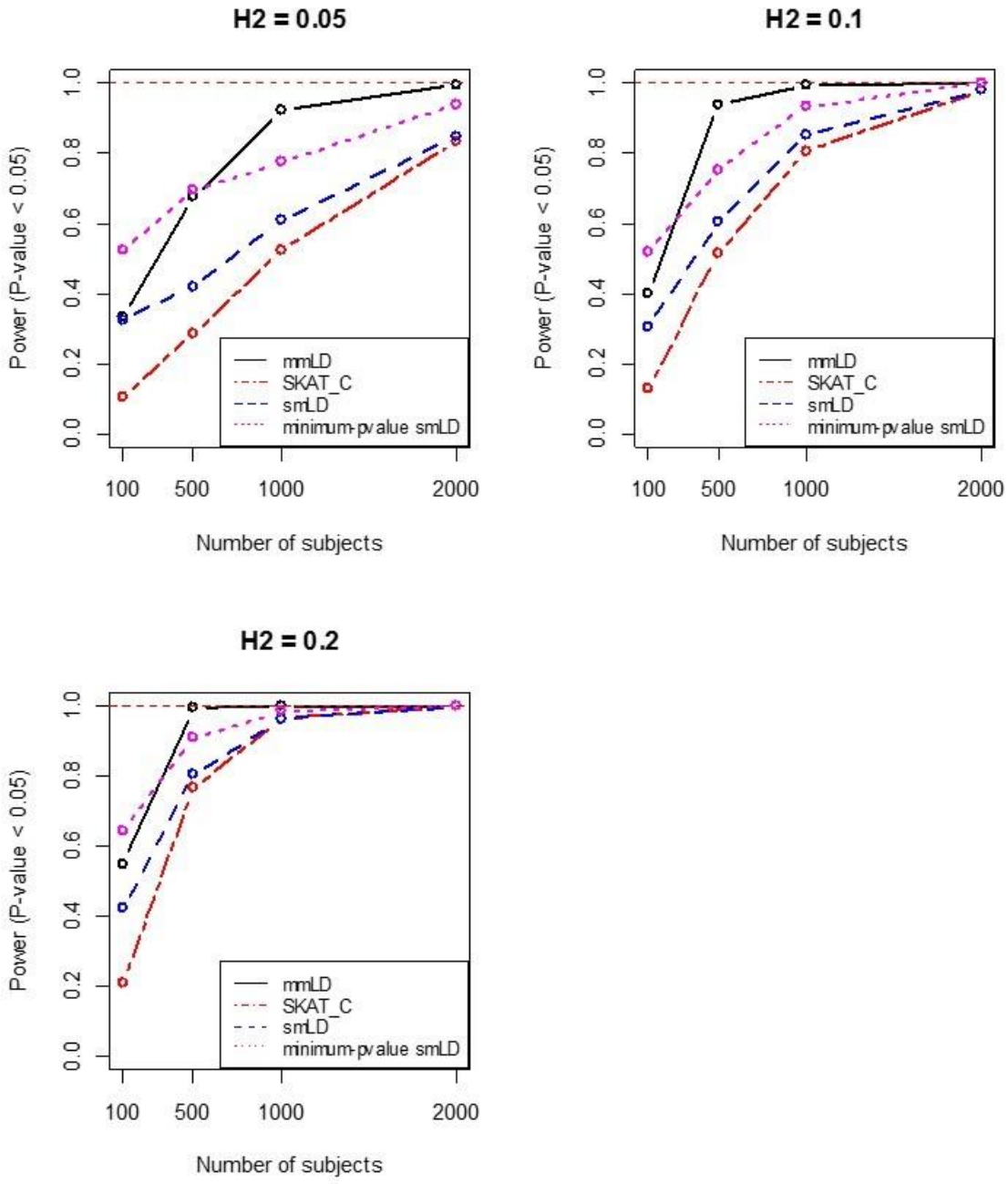

Figure 21 Power comparison of mmLD, smLD, minimum p-value smLD and SKAT_C for Scenario (1); 3 known markers

Table 31 Power comparison of mmLD, smLD, minimum p-value smLD and SKAT_C for Scenario (1); 3 known markers

Number of subjects	H^2= 0.05			
	mmLD	Adjusted smLD	Minimum p-value smLD	SKAT_C
100	0.335	0.325	0.525	0.105
500	0.675	0.420	0.695	0.285
1000	0.925	0.610	0.775	0.525
2000	0.995	0.850	0.940	0.835
	H^2= 0.1			
100	0.400	0.305	0.520	0.130
500	0.940	0.605	0.755	0.515
1000	0.995	0.855	0.935	0.805
2000	1	0.980	1	0.980
	H^2= 0.2			
100	0.550	0.425	0.645	0.210
500	0.995	0.805	0.910	0.765
1000	1	0.965	0.985	0.965
2000	1	1	1	1

Table 32 Means and standard errors of parameters for Scenario (1); 3 known markers

True value		N=100			N=500		
		H^2=0.05	H^2=0.1	H^2=0.2	H^2=0.05	H^2=0.1	H^2=0.2
p_{000Q}	0	0.052 (0.0007)	0.037 (0.0005)	0.025 (0.0003)	0.038 (0.0002)	0.032 (0.0001)	0.022 (0.0001)

p_{000q}	0.16	0.109 (0.0008)	0.123 (0.0006)	0.137 (0.0004)	0.121 (0.0002)	0.127 (0.0001)	0.138 (0.0001)
p_{001Q}	0.13	0.112 (0.0008)	0.112 (0.0007)	0.128 (0.0004)	0.114 (0.0002)	0.117 (0.0002)	0.124 (0.0001)
p_{001q}	0.04	0.056 (0.0008)	0.054 (0.0006)	0.043 (0.0004)	0.055 (0.0002)	0.054 (0.0002)	0.047 (0.0001)
p_{010Q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{010q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{011Q}	0.1	0.081 (0.0005)	0.074 (0.0004)	0.087 (0.0003)	0.077 (0.0002)	0.083 (0.0001)	0.091 (0.0001)
p_{011q}	0	0.025 (0.0005)	0.024 (0.0004)	0.013 (0.0002)	0.024 (0.0002)	0.017 (0.0001)	0.01 (0.0001)
p_{100Q}	0.13	0.153 (0.0013)	0.139 (0.0009)	0.143 (0.0006)	0.148 (0.0004)	0.151 (0.0002)	0.146 (0.0002)
p_{100q}	0.2	0.172 (0.0013)	0.194 (0.001)	0.182 (0.0006)	0.184 (0.0004)	0.18 (0.0002)	0.184 (0.0002)
p_{101Q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{101q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{110Q}	0.08	0.067 (0.0007)	0.091 (0.0006)	0.085 (0.0004)	0.08 (0.0002)	0.08 (0.0001)	0.083 (0.0001)
p_{110q}	0.08	0.092 (0.0008)	0.069 (0.0006)	0.075 (0.0004)	0.078 (0.0002)	0.079 (0.0001)	0.076 (0.0001)
p_{111Q}	0	0.023 (0.0004)	0.026 (0.0004)	0.017 (0.0002)	0.021 (0.0001)	0.018 (0.0001)	0.011 (0.0001)
p_{111q}	0.08	0.057 (0.0005)	0.057 (0.0004)	0.064 (0.0003)	0.06 (0.0001)	0.062 (0.0001)	0.068 (0.0001)
μ_0	10	18.2 (0.135)	15.3 (0.092)	12.4 (0.033)	15.5 (0.041)	12.9 (0.018)	11.5 (0.009)
μ_1	5	4.2 (0.111)	4.7 (0.063)	4.1 (0.029)	4.4 (0.032)	4.7 (0.014)	4.5 (0.008)
μ_2	0	-8.2 (0.159)	-5.2 (0.085)	-2.4 (0.036)	-4.9 (0.04)	-3.5 (0.017)	-1.5 (0.01)
True value		N=1000			N=2000		
		H^2=0.05	H^2=0.1	H^2=0.2	H^2=0.05	H^2=0.1	H^2=0.2
p_{000Q}	0	0.038 (0.0001)	0.031 (0.0001)	0.021 (0.0001)	0.039 (0.0001)	0.028 (0.0001)	0.017 (0.0001)

p_{000q}	0.16	0.122 (0.0001)	0.128 (0.0001)	0.139 (0.0001)	0.121 (0.0001)	0.132 (0.0001)	0.142 (0.0001)
p_{001Q}	0.13	0.114 (0.0001)	0.122 (0.0001)	0.127 (0.0001)	0.115 (0.0001)	0.119 (0.0001)	0.127 (0.0001)
p_{001q}	0.04	0.056 (0.0001)	0.049 (0.0001)	0.043 (0.0001)	0.055 (0.0001)	0.051 (0.0001)	0.043 (0.0001)
p_{010Q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{010q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{011Q}	0.1	0.079 (0.0001)	0.084 (0.0001)	0.091 (0.0001)	0.08 (0.0001)	0.085 (0.0001)	0.093 (0)
p_{011q}	0	0.021 (0.0001)	0.015 (0.0001)	0.009 (0)	0.02 (0.0001)	0.014 (0.0001)	0.008 (0)
p_{100Q}	0.13	0.154 (0.0002)	0.151 (0.0002)	0.146 (0.0001)	0.153 (0.0001)	0.147 (0.0001)	0.142 (0.0001)
p_{100q}	0.2	0.177 (0.0002)	0.179 (0.0002)	0.183 (0.0001)	0.177 (0.0001)	0.184 (0.0001)	0.188 (0.0001)
p_{101Q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{101q}	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
p_{110Q}	0.08	0.081 (0.0001)	0.084 (0.0001)	0.086 (0.0001)	0.084 (0.0001)	0.083 (0.0001)	0.085 (0.0001)
p_{110q}	0.08	0.078 (0.0001)	0.077 (0.0001)	0.075 (0.0001)	0.076 (0.0001)	0.077 (0.0001)	0.075 (0.0001)
p_{111Q}	0	0.02 (0.0001)	0.017 (0.0001)	0.009 (0)	0.02 (0)	0.013 (0)	0.009 (0)
p_{111q}	0.08	0.06 (0.0001)	0.063 (0.0001)	0.071 (0.0001)	0.06 (0.0001)	0.067 (0)	0.071 (0)
μ_0	10	14.6 (0.022)	12.2 (0.012)	11.1 (0.007)	13.9 (0.016)	12.3 (0.009)	10.9 (0.005)
μ_1	5	4.6 (0.016)	4.5 (0.009)	4.5 (0.007)	4.5 (0.012)	4.6 (0.008)	4.6 (0.004)
μ_2	0	-4.8 (0.023)	-3 (0.011)	-1.5 (0.007)	-4.4 (0.017)	-2.4 (0.008)	-1.2 (0.004)

Scenario 1-3. Four known markers

In this simulation, four markers are considered here and the parameters in *Table 33*. Similar to the result of three markers, the mmLD shows higher power performance compared to the SKAT_C, smLD and minimum p-value smLD in *Figure 22* and *Table 34*. Through these results, it is clear that the mmLD shows stable performance compared to other existing methods for the non-genotyped QTL. *Table 35* shows the estimates for the true values.

Table 33 Parameters of haplotype frequencies for Scenario (1) with 4 known markers.

p_{0000Q}	p_{0000q}	p_{0001Q}	p_{0001q}	p_{0010Q}	p_{0010q}	p_{0011Q}	p_{0011q}
0.025	0.12	0.03	0.015	0	0	0	0
p_{0100Q}	p_{0100q}	p_{0101Q}	p_{0101q}	p_{0110Q}	p_{0110q}	p_{0111Q}	p_{0111q}
0.05	0.1	0.09	0	0.02	0.04	0	0.01
p_{1000Q}	p_{1000q}	p_{1001Q}	p_{1001q}	p_{1010Q}	p_{1010q}	p_{1011Q}	p_{1011q}
0.11	0	0	0.02	0	0.1	0.07	0
p_{1100Q}	p_{1100q}	p_{1101Q}	p_{1101q}	p_{1110Q}	p_{1110q}	p_{1111Q}	p_{1111q}
0	0	0	0.02	0.07	0.04	0	0.07

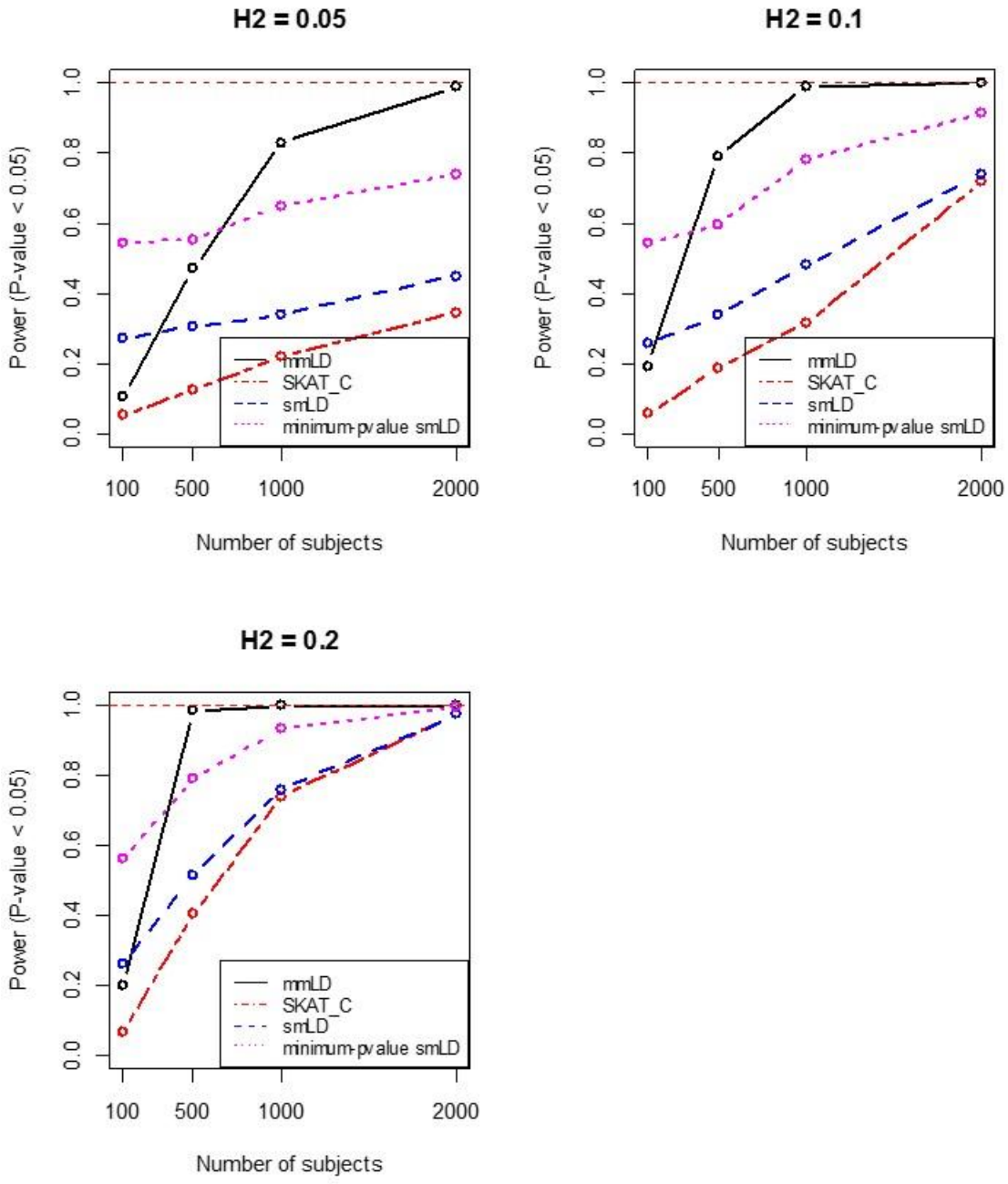

Figure 22 Power comparison of mmLD, smLD, minimum p-value smLD and SKAT_C for Scenario (1); 4 known markers

Table 34 Power comparison of mmLD, smLD, minimum p-value smLD and SKAT_C for Scenario (1); 4 known markers

Number of subjects	H^2= 0.05			
	mmLD	Adjusted smLD	Minimum p-value smLD	SKAT_C
100	0.105	0.270	0.545	0.050
500	0.470	0.305	0.555	0.125
1000	0.830	0.340	0.650	0.220
2000	0.990	0.450	0.740	0.345
	H^2= 0.1			
100	0.190	0.255	0.545	0.055
500	0.790	0.340	0.595	0.185
1000	0.990	0.480	0.780	0.315
2000	1	0.740	0.915	0.720
	H^2= 0.2			
100	0.200	0.255	0.545	0.065
500	0.985	0.340	0.595	0.405
1000	1	0.480	0.780	0.740
2000	1	0.740	0.915	0.975

Table 35 Means and standard errors of parameters for Scenario (1); 4 known markers

True value	N=100			N=500		
	H^2=0.05	H^2=0.1	H^2=0.2	H^2=0.05	H^2=0.1	H^2=0.2

p_{0000Q}	0.025	0.059 (0.0003)	0.048 (0.00025)	0.045 (0.00022)	0.053 (0.00014)	0.047 (0.00011)	0.041 (0.00009)
p_{0000q}	0.12	0.088 (0.0003)	0.094 (0.00028)	0.105 (0.00025)	0.091 (0.00014)	0.098 (0.00011)	0.104 (0.0001)
p_{0001Q}	0.03	0.024 (0.00015)	0.023 (0.00014)	0.025 (0.00014)	0.027 (0.00008)	0.027 (0.00007)	0.027 (0.00006)
p_{0001q}	0.015	0.023 (0.00015)	0.023 (0.00015)	0.018 (0.00012)	0.02 (0.00008)	0.019 (0.00006)	0.018 (0.00006)
p_{0010Q}	0	0 (0.00002)	0 (0.00001)	0 (0.00001)	0 (NA)	0 (NA)	0 (NA)
p_{0010q}	0	0 (0.00003)	0.001 (0.00002)	0 (0.00002)	0 (NA)	0 (NA)	0 (NA)
p_{0011Q}	0	0 (0.00001)	0 (0.00001)	0 (0.00001)	0 (NA)	0 (NA)	0 (NA)
p_{0011q}	0	0 (0.00001)	0 (0.00002)	0 (0.00001)	0 (NA)	0 (NA)	0 (NA)
p_{0100Q}	0.05	0.066 (0.00032)	0.069 (0.00029)	0.064 (0.00027)	0.066 (0.00013)	0.062 (0.00012)	0.06 (0.0001)
p_{0100q}	0.1	0.088 (0.00033)	0.081 (0.00028)	0.084 (0.00028)	0.083 (0.00013)	0.088 (0.00012)	0.091 (0.00011)
p_{0101Q}	0.09	0.054 (0.00023)	0.062 (0.00023)	0.067 (0.00021)	0.065 (0.00011)	0.069 (0.00008)	0.076 (0.00007)
p_{0101q}	0	0.036 (0.00022)	0.028 (0.00021)	0.024 (0.00017)	0.024 (0.0001)	0.019 (0.00008)	0.013 (0.00007)
p_{0110Q}	0.02	0.022 (0.00015)	0.022 (0.00015)	0.023 (0.00014)	0.026 (0.00007)	0.024 (0.00006)	0.021 (0.00005)
p_{0110q}	0.04	0.034 (0.00013)	0.036 (0.00015)	0.035 (0.00015)	0.035 (0.00008)	0.037 (0.00007)	0.039 (0.00006)
p_{0111Q}	0	0.003 (0.00005)	0.003 (0.00004)	0.004 (0.00005)	0.003 (0.00002)	0.004 (0.00002)	0.003 (0.00002)
p_{0111q}	0.01	0.006 (0.00007)	0.009 (0.00007)	0.007 (0.00006)	0.007 (0.00003)	0.006 (0.00003)	0.007 (0.00003)
p_{1000Q}	0.11	0.074 (0.00022)	0.075 (0.00024)	0.083 (0.0002)	0.079 (0.00011)	0.084 (0.00009)	0.096 (0.00008)
p_{1000q}	0	0.038 (0.00021)	0.035 (0.00022)	0.023 (0.00016)	0.031 (0.0001)	0.026 (0.00009)	0.014 (0.00007)
p_{1001Q}	0	0.009 (0.00008)	0.007 (0.00008)	0.007 (0.00008)	0.008 (0.00004)	0.006 (0.00004)	0.005 (0.00003)
p_{1001q}	0.02	0 (NA)	0 (0.00001)	0.001 (0.00001)	0 (NA)	0 (NA)	0 (NA)
p_{1010Q}	0	0.034 (0.00022)	0.032 (0.0002)	0.025 (0.00017)	0.032 (0.00011)	0.022 (0.00007)	0.018 (0.00007)
p_{1010q}	0.1	0.067 (0.00024)	0.068 (0.00023)	0.075 (0.00022)	0.069 (0.00012)	0.078 (0.00009)	0.083 (0.00009)
p_{1011Q}	0.07	0.041 (0.00018)	0.048 (0.00019)	0.049 (0.00017)	0.048 (0.00009)	0.054 (0.00008)	0.059 (0.00006)
p_{1011q}	0	0.029 (0.00018)	0.021 (0.00016)	0.019 (0.00014)	0.021 (0.00009)	0.016 (0.00007)	0.01 (0.00005)

p_{1100Q}	0	0 (NA)	0 (0.00001)	0 (0.00001)	0 (NA)	0 (NA)	0 (NA)
p_{1100q}	0	0 (NA)	0 (NA)	0.001 (0.00002)	0 (NA)	0 (NA)	0 (NA)
p_{1101Q}	0	0.007 (0.00006)	0.007 (0.00007)	0.005 (0.00005)	0.007 (0.00004)	0.005 (0.00003)	0.005 (0.00003)
p_{1101q}	0.02	0.012 (0.00009)	0.013 (0.00008)	0.015 (0.00008)	0.013 (0.00004)	0.014 (0.00003)	0.015 (0.00004)
p_{1110Q}	0.07	0.06 (0.00027)	0.061 (0.00024)	0.061 (0.00023)	0.06 (0.00013)	0.066 (0.0001)	0.066 (0.00009)
p_{1110q}	0.04	0.048 (0.00023)	0.051 (0.00025)	0.048 (0.00023)	0.05 (0.00013)	0.044 (0.0001)	0.044 (0.0001)
p_{1111Q}	0	0.027 (0.00018)	0.024 (0.00016)	0.017 (0.00015)	0.022 (0.00009)	0.016 (0.00007)	0.011 (0.00005)
p_{1111q}	0.07	0.041 (0.00019)	0.043 (0.0002)	0.055 (0.00017)	0.048 (0.0001)	0.054 (0.00008)	0.059 (0.00007)
μ_0	10	20.1 (0.047)	15.5 (0.038)	12.6 (0.02)	16.1 (0.025)	13.7 (0.013)	11.6 (0.008)
μ_1	5	4.6 (0.039)	4.4 (0.03)	4.4 (0.019)	4.1 (0.021)	4.6 (0.01)	4.6 (0.007)
μ_2	0	-8.4 (0.05)	-4.6 (0.036)	-2.2 (0.023)	-5.8 (0.032)	-3.3 (0.013)	-1.7 (0.009)
True value		N=1000			N=2000		
		H^2=0.05	H^2=0.1	H^2=0.2	H^2=0.05	H^2=0.1	H^2=0.2
p_{0000Q}	0.025	0.052 (0.00009)	0.046 (0.00009)	0.039 (0.00007)	0.051 (0.00006)	0.045 (0.00006)	0.037 (0.00005)
p_{0000q}	0.12	0.094 (0.0001)	0.099 (0.00009)	0.107 (0.00007)	0.094 (0.00007)	0.101 (0.00006)	0.108 (0.00005)
p_{0001Q}	0.03	0.026 (0.00005)	0.027 (0.00005)	0.028 (0.00004)	0.025 (0.00005)	0.027 (0.00004)	0.029 (0.00004)
p_{0001q}	0.015	0.019 (0.00005)	0.018 (0.00005)	0.016 (0.00004)	0.019 (0.00005)	0.018 (0.00004)	0.016 (0.00003)
p_{0010Q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{0010q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{0011Q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{0011q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{0100Q}	0.05	0.064 (0.00009)	0.063 (0.00008)	0.058 (0.00007)	0.064 (0.00007)	0.06 (0.00006)	0.059 (0.00005)
p_{0100q}	0.1	0.085 (0.0001)	0.088 (0.00008)	0.092 (0.00008)	0.086 (0.00008)	0.09 (0.00007)	0.091 (0.00005)
p_{0101Q}	0.09	0.067 (0.00007)	0.072 (0.00007)	0.078 (0.00006)	0.069 (0.00006)	0.074 (0.00005)	0.081 (0.00005)

p_{0101q}	0	0.022 (0.00007)	0.018 (0.00006)	0.013 (0.00005)	0.021 (0.00005)	0.016 (0.00005)	0.01 (0.00004)
p_{0110Q}	0.02	0.026 (0.00006)	0.024 (0.00004)	0.023 (0.00005)	0.025 (0.00004)	0.024 (0.00004)	0.022 (0.00003)
p_{0110q}	0.04	0.034 (0.00006)	0.036 (0.00005)	0.037 (0.00005)	0.035 (0.00004)	0.037 (0.00004)	0.038 (0.00003)
p_{0111Q}	0	0.004 (0.00002)	0.003 (0.00002)	0.002 (0.00001)	0.003 (0.00002)	0.003 (0.00001)	0.002 (0.00001)
p_{0111q}	0.01	0.007 (0.00002)	0.007 (0.00002)	0.008 (0.00002)	0.007 (0.00002)	0.007 (0.00002)	0.008 (0.00001)
p_{1000Q}	0.11	0.083 (0.00007)	0.088 (0.00007)	0.097 (0.00006)	0.084 (0.00006)	0.091 (0.00006)	0.098 (0.00005)
p_{1000q}	0	0.027 (0.00007)	0.022 (0.00006)	0.014 (0.00005)	0.026 (0.00006)	0.02 (0.00005)	0.012 (0.00004)
p_{1001Q}	0	0.007 (0.00003)	0.005 (0.00003)	0.004 (0.00002)	0.005 (0.00002)	0.005 (0.00002)	0.004 (0.00002)
p_{1001q}	0.02	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{1010Q}	0	0.027 (0.00008)	0.022 (0.00006)	0.015 (0.00004)	0.026 (0.00006)	0.02 (0.00004)	0.014 (0.00003)
p_{1010q}	0.1	0.073 (0.00008)	0.078 (0.00007)	0.084 (0.00006)	0.074 (0.00006)	0.079 (0.00005)	0.086 (0.00004)
p_{1011Q}	0.07	0.053 (0.00007)	0.056 (0.00005)	0.061 (0.00005)	0.054 (0.00005)	0.057 (0.00004)	0.062 (0.00004)
p_{1011q}	0	0.016 (0.00006)	0.015 (0.00006)	0.009 (0.00004)	0.017 (0.00005)	0.012 (0.00004)	0.008 (0.00003)
p_{1100Q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{1100q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{1101Q}	0	0.006 (0.00003)	0.005 (0.00002)	0.003 (0.00002)	0.006 (0.00002)	0.004 (0.00002)	0.003 (0.00001)
p_{1101q}	0.02	0.014 (0.00003)	0.016 (0.00003)	0.017 (0.00002)	0.014 (0.00002)	0.015 (0.00002)	0.017 (0.00002)
p_{1110Q}	0.07	0.063 (0.00008)	0.064 (0.00008)	0.066 (0.00007)	0.063 (0.00006)	0.066 (0.00006)	0.068 (0.00004)
p_{1110q}	0.04	0.046 (0.00009)	0.045 (0.00008)	0.043 (0.00006)	0.047 (0.00006)	0.044 (0.00006)	0.042 (0.00005)
p_{1111Q}	0	0.019 (0.00006)	0.016 (0.00005)	0.011 (0.00004)	0.018 (0.00005)	0.014 (0.00004)	0.01 (0.00003)
p_{1111q}	0.07	0.052 (0.00006)	0.054 (0.00006)	0.059 (0.00005)	0.051 (0.00005)	0.057 (0.00004)	0.06 (0.00004)
μ_0	10	15.2 (0.015)	13 (0.01)	11.5 (0.006)	14.9 (0.014)	12.7 (0.008)	11.2 (0.004)
μ_1	5	4.1 (0.014)	4.7 (0.008)	4.7 (0.005)	4.7 (0.01)	4.7 (0.006)	4.8 (0.003)
μ_2	0	-5.1 (0.018)	-3.1 (0.01)	-1.7 (0.005)	-5 (0.012)	-2.8 (0.007)	-1.5 (0.003)

Scenario 2: QTL is genotyped as a marker

In this scenario, one marker is assumed to be a QTL. The simulated setting is given in *Table 36*, in which three markers are included and the third marker is set to be the QTL.

It is expected that the smLD would work the best in this case since QTL is indeed genotyped. Therefore, considering each marker should have better power than considering the linkage disequilibrium of several markers. *Table 37* and *Figure 23* display the power comparisons of four methods. Although the power of the mmLD is slightly less than that of the smLD, it still shows comparable power to the smLD. This suggests that the mmLD can be expected to provide the robust result of power even when QTL is genotyped.

Table 36 Parameters of haplotype frequencies for Scenario (2); 3 known markers

p_{000Q}	p_{000q}	p_{001Q}	p_{001q}	p_{010Q}	p_{010q}	p_{011Q}	p_{011q}
0.09	0	0	0.12	0.08	0	0	0.12
p_{100Q}	p_{100q}	p_{101Q}	p_{101q}	p_{110Q}	p_{110q}	p_{111Q}	p_{111q}
0.2	0	0	0	0.15	0	0	0.24

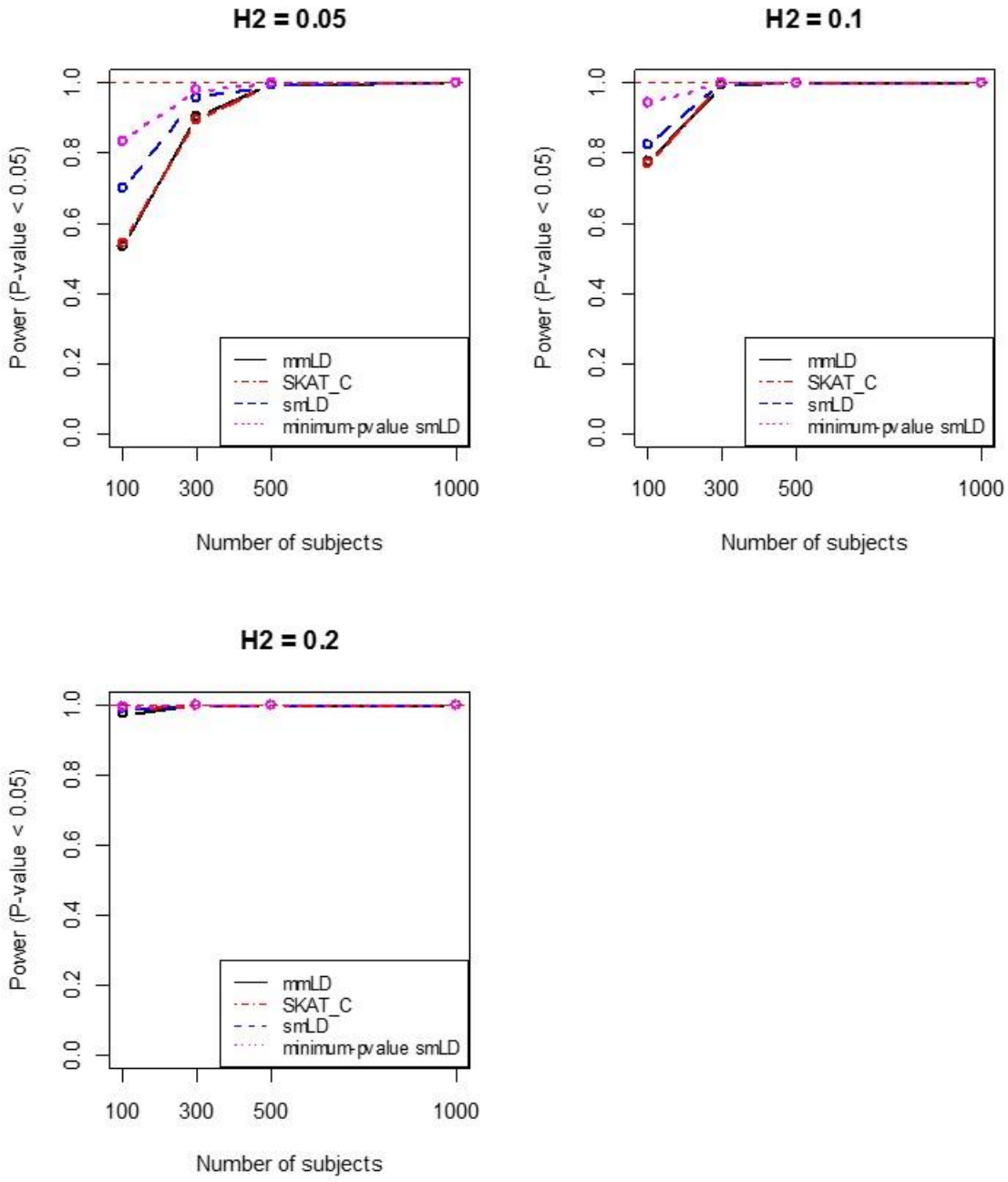

Figure 23 Power comparison of mmLD, smLD, minimum p-value smLD and SKAT_C for Scenario (2); 3 known markers

Table 37 Power comparison of mmLD, smLD, minimum p-value smLD and SKAT_C for Scenario (2); 3 known markers

Number of subjects	H^2= 0.05			
	mmLD	Adjusted smLD	Minimum p-value smLD	SKAT_C
100	0.535	0.700	0.835	0.545
300	0.905	0.960	0.980	0.895
500	0.995	0.995	1	1
1000	1	1	1	1
	H^2= 0.1			
100	0.775	0.825	0.945	0.770
300	0.995	1	1	1
500	1	1	1	1
1000	1	1	1	1
	H^2= 0.2			
100	0.975	0.990	0.995	0.990
300	1	1	1	1
500	1	1	1	1
1000	1	1	1	1

Table 38 Means and standard errors of parameters for Scenario (2); 3 known markers

True value		N=100			N=300		
		H^2=0.05	H^2=0.1	H^2=0.2	H^2=0.05	H^2=0.1	H^2=0.2
p_{000Q}	0.04	0.024 (0.0002)	0.027 (0.0001)	0.029 (0.0001)	0.029 (0.0001)	0.031 (0.0001)	0.031 (0.0001)
p_{000q}	0	0.012 (0.0001)	0.012 (0.0001)	0.01 (0.0001)	0.011 (0.0001)	0.01 (0.0001)	0.007 (0)

p_{001Q}	0.25	0.07 (0.0005)	0.061 (0.0003)	0.053 (0.0002)	0.082 (0.0002)	0.063 (0.0002)	0.039 (0.0002)
p_{001q}	0	0.182 (0.0005)	0.188 (0.0003)	0.203 (0.0002)	0.166 (0.0003)	0.185 (0.0002)	0.21 (0.0002)
p_{010Q}	0	0.001 (0.00004)	0.001 (0.00002)	0.002 (0.00002)	0 (0.00001)	0 (0.00001)	0 (0.00001)
p_{010q}	0	0.001 (0.00004)	0.001 (0.00002)	0.001 (0.00001)	0 (0.00001)	0 (0.00001)	0 (0.000003)
p_{011Q}	0.12	0.04 (0.0003)	0.037 (0.0002)	0.028 (0.0001)	0.038 (0.0001)	0.03 (0.0001)	0.021 (0.0001)
p_{011q}	0	0.08 (0.0004)	0.082 (0.0002)	0.091 (0.0002)	0.081 (0.0002)	0.088 (0.0001)	0.099 (0.0001)
p_{100Q}	0.23	0.187 (0.0005)	0.188 (0.0003)	0.204 (0.0002)	0.178 (0.0002)	0.19 (0.0002)	0.203 (0.0001)
p_{100q}	0	0.044 (0.0004)	0.04 (0.0003)	0.027 (0.0002)	0.054 (0.0002)	0.042 (0.0002)	0.026 (0.0001)
p_{101Q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{101q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{110Q}	0.06	0.043 (0.0002)	0.042 (0.0002)	0.045 (0.0001)	0.045 (0.0001)	0.046 (0.0001)	0.052 (0.0001)
p_{110q}	0	0.019 (0.0002)	0.017 (0.0001)	0.012 (0.0001)	0.015 (0.0001)	0.014 (0.0001)	0.008 (0)
p_{111Q}	0.3	0.086 (0.0005)	0.082 (0.0004)	0.063 (0.0002)	0.098 (0.0003)	0.073 (0.0002)	0.045 (0.0002)
p_{111q}	0	0.211 (0.0005)	0.22 (0.0004)	0.233 (0.0003)	0.204 (0.0003)	0.227 (0.0002)	0.257 (0.0002)
μ_0	10	17.9 (0.068)	14.2 (0.033)	12.2 (0.016)	15.6 (0.032)	13.3 (0.017)	11.6 (0.008)
μ_1	5	4.7 (0.056)	4.3 (0.027)	4.5 (0.011)	3.6 (0.026)	4.4 (0.013)	4.8 (0.008)
μ_2	0	-9 (0.058)	-4.6 (0.026)	-3.2 (0.014)	-6.7 (0.031)	-4.4 (0.014)	-2.1 (0.008)
True value		N=500			N=1000		
		H^2=0.05	H^2=0.1	H^2=0.2	H^2=0.05	H^2=0.1	H^2=0.2
p_{000Q}	0.04	0.029 (0.0001)	0.03 (0.00005)	0.033 (0.00004)	0.031 (0.00004)	0.033 (00.00004)	0.035 (0.00003)
p_{000q}	0	0.011 (0.0001)	0.009 (0.00004)	0.006 (0.00003)	0.009 (00.00004)	0.006 (0.00003)	0.005 (0.00002)

p_{001Q}	0.25	0.077 (0.0002)	0.059 (0.0002)	0.035 (0.0001)	0.07 (0.0002)	0.053 (0.0001)	0.027 (0.0001)
p_{001q}	0	0.175 (0.0002)	0.191 (0.0002)	0.216 (0.0001)	0.179 (0.0002)	0.197 (0.0001)	0.223 (0.0001)
p_{010Q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{010q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{011Q}	0.12	0.036 (0.0001)	0.029 (0.0001)	0.017 (0.0001)	0.033 (0.0001)	0.026 (0.0001)	0.013 (0.0001)
p_{011q}	0	0.084 (0.0001)	0.091 (0.0001)	0.103 (0.0001)	0.087 (0.0001)	0.095 (0.0001)	0.106 (0.0001)
p_{100Q}	0.23	0.18 (0.0002)	0.192 (0.0001)	0.208 (0.0001)	0.184 (0.0001)	0.199 (0.0001)	0.21 (0.0001)
p_{100q}	0	0.05 (0.0002)	0.039 (0.0001)	0.023 (0.0001)	0.046 (0.0001)	0.032 (0.0001)	0.019 (0.0001)
p_{101Q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{101q}	0	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
p_{110Q}	0.06	0.044 (0.0001)	0.05 (0.0001)	0.052 (0.00005)	0.046 (0.0001)	0.051 (0.00005)	0.054 (0.00004)
p_{110q}	0	0.015 (0.0001)	0.01 (0.00005)	0.008 (0.00004)	0.014 (00.00005)	0.01 (0.00004)	0.006 (0.00003)
p_{111Q}	0.3	0.088 (0.0002)	0.066 (0.0002)	0.042 (0.0001)	0.085 (0.0002)	0.063 (0.0002)	0.031 (0.0001)
p_{111q}	0	0.211 (0.0002)	0.234 (0.0002)	0.259 (0.0002)	0.215 (0.0002)	0.236 (0.0002)	0.27 (0.0001)
μ_0	10	15.2 (0.024)	12.8 (0.013)	11.4 (0.007)	14.4 (0.016)	12.1 (0.008)	11 (0.005)
μ_1	5	4.1 (0.018)	4.5 (0.011)	4.8 (0.006)	4.2 (0.013)	4.4 (0.007)	4.9 (0.004)
μ_2	0	-6.6 (0.021)	-3.8 (0.012)	-1.9 (0.007)	-5.8 (0.016)	-3.3 (0.01)	-1.4 (0.005)

3.2.4. Power change with the number of known markers

Logically, it is expected that power would be increased as more markers are included in the LD mapping framework, although computationally it would be much harder for more markers. However, it is also expected that the marginal gain of each additional markers would decrease if markers are correlated. So a question of interest is how many markers we should include in the LD mapping framework.

In order to investigate this, we tracked the power change by the number of known markers involved. In this simulation, we considered seven known markers and one QTL. The phenotype is assumed to follow a mixture Gaussian distribution with μ_0 = 20, μ_1 = 40, μ_2 = 60 and the heritability value (H^2) was set up 0.05. The sample size was 2,000 and each simulation was conducted 200 times. Details of simulated haplotype frequencies are given in *Table 39*. *Figure 24* shows the power change of the mmLD, SKAT_C, smLD, and minimum p-value smLD. When three known markers were applied, the power of mmLD reaches 1 and stays there. Also, the powers of the other methods reach almost 1 when four or five markers are considered. The overall power of mmLD is consistently higher than those of SKAT_C, smLD or minimum p-value smLD. This suggests that in practice, we probably need to consider only 4 or 5- marker LD mapping.

Table 39 Haplotype frequencies of seven known markers and one QTL

SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	**QTL**	Simulated haplotype frequencies
0	0	0	0	0	0	0	1	0.009
0	0	0	0	1	1	1	1	0.03
0	0	0	1	1	0	1	1	0.01
0	0	0	1	1	1	1	1	0.05
0	0	1	0	1	0	0	0	0.02
0	1	0	1	1	0	1	0	0.1
0	1	0	1	1	1	1	0	0.004
0	1	0	0	1	1	1	1	0.01
0	1	0	1	1	1	1	1	0.08
0	1	1	0	0	1	1	1	0.04
0	1	1	0	1	0	0	1	0.02
1	0	0	0	0	1	0	0	0.1
1	0	0	1	0	1	0	1	0.04
1	0	1	0	0	0	0	1	0.2
1	0	1	0	1	1	0	1	0.08
1	0	1	1	0	1	1	1	0.05
1	1	0	0	0	1	1	0	0.007
1	1	1	1	0	0	0	0	0.05
1	1	1	1	1	1	1	0	0.07
1	1	1	0	1	1	0	1	0.03

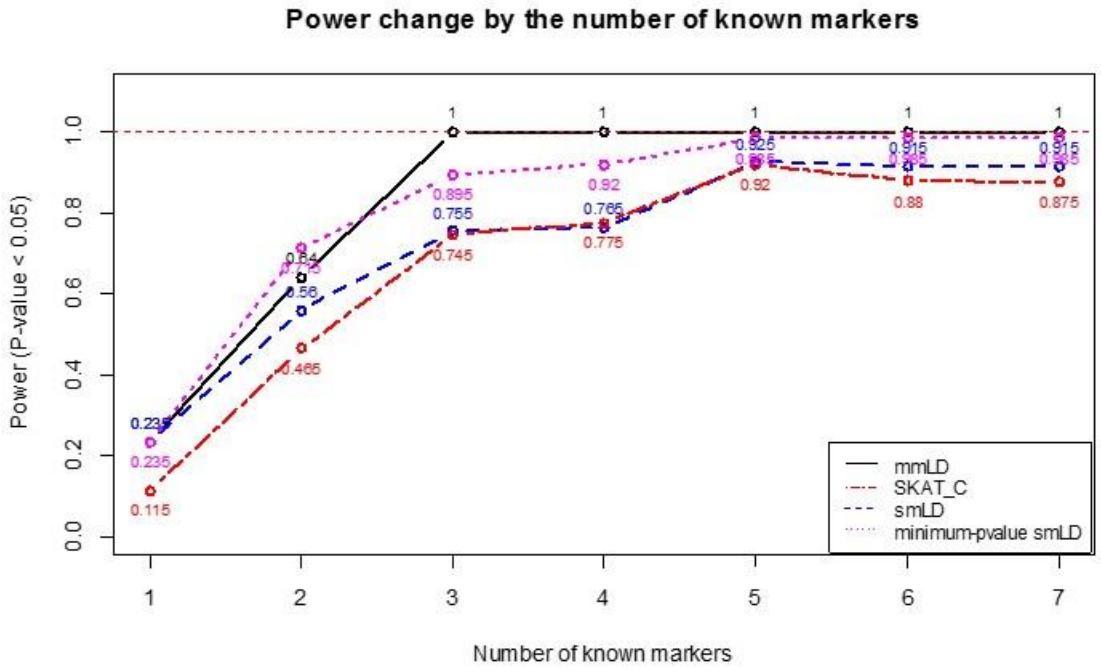

Figure 24 Power change by the number of known markers

3.3. Real data application

In order to further evaluate the applicability of the mmLD mapping with real genomic structure, it was applied to a real dataset, the "GAW17". The data is provided by Texas Biomedical Research Institute at Genetic Analysis Workshop [31]. The GAW17 is a "mini-exome" scan, using real sequence data for several hundred genes from the 1000 Genomes Project [31, 32].

Since the original phenotypic data provided by the GAW17 is simulated based on a few rare variants with minor allele frequencies (MAF < 5%), it is not directly applicable to our mmLD model. Thus, although the real genomic sequences were used here, we re-generated the phenotypic data with a specified QTL, and some rare variants (MAF < 5%) were removed in this analysis. The QTL is set to be the SNP "C3S784", located on the 3rd chromosome and indicated by the blue solid line in *Figure 24*. We assume that the phenotype follows a mixture Gaussian normal distribution with three means and the same variance by three genotypes of the QTL, with μ_0 = 20, μ_1 = 40, and μ_2 = 60 and its heritability value (H^2) was 0.1. The mmLD was then applied to scan the whole chromosome with a sliding window searching for five markers.

Figure 25 shows the scatter plot of negative logarithm of p-values for the 3rd chromosome. Since there are 113 SNPs considered here, the significance cut-off of negative logarithmic p-value was set to be 3.35, which is calculated based on the Bonferroni correction. The upper plot is the output of the mmLD and the lower is the output of the SKAT_C.

As shown in *Figure 25*, it is clear that the several p-values corresponding to the nearby true regions of the QTL pass the significance level in the both methods. Although both detected the true region of the QTL, the degrees of significant p-values of the mmLD are less than those of the SKAT_C. However, SKAT_C shows false signal near 30^{th} loci while the mmLD detected only true regions of the QTL.

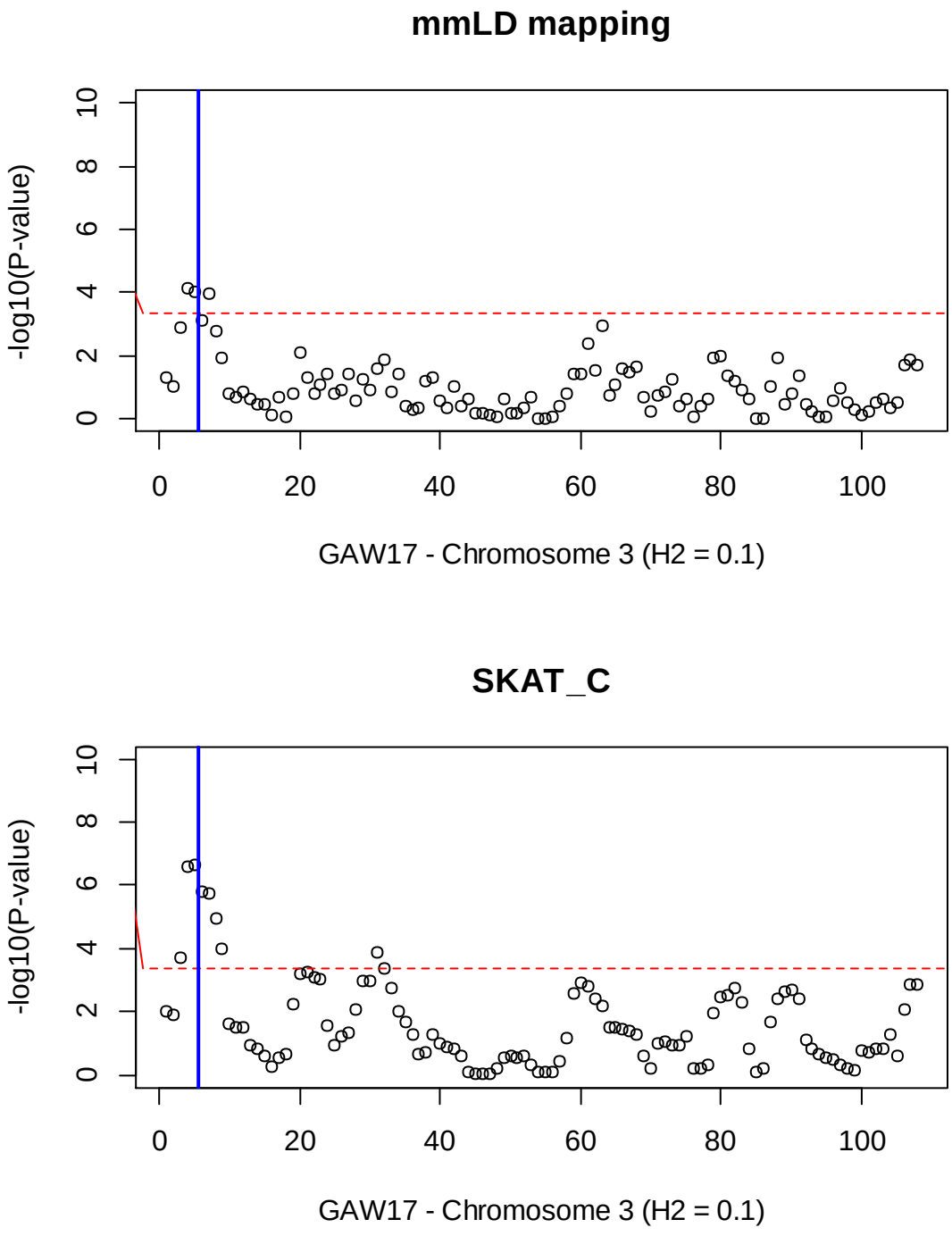

Figure 25 Scatter plot of negative logarithmic p-value for chromosome 3 of GAW17

Chapter 4 Conclusion and Discussion

The purpose of this study is to extend the two-marker LD mapping of QTL (tmLD) method proposed by Yang et al [18] to a multi-marker case. The tmLD is a novel statistical method that intends to identify QTL with adjacent two markers; however, it has some limitations to handle multiple markers simultaneously: The first is how to efficiently estimate haplotypic and genotypic frequencies, given their complicated relationship; and the second is how to find proper degrees of freedom for the likelihood ratio test.

Although there are a few open software tools for estimating haplotype frequencies from genetic data, none of them can be directly incorporated into the QTL mapping framework. We have investigated the regularity of calculation of genotypic probabilities from haplotype frequencies and suggested an algorithm for the joint genotypic probabilities.

To address the issue of the degrees of freedom in the likelihood ratio test, we found that it should be determined by the number of haplotypes with non-zero frequencies. This is because haplotypes with zero frequencies are not estimated from the data and therefore do not contribute to the number of parameters. In addition, we proposed a sequential likelihood ratio test procedure to determine the degrees of freedom from haplotypes with small frequencies. In this process, multiple testing issues may occur due to the iterative

testing scheme. We further used the Bonferroni correction method to control the family-wise type I error, which showed minimal difference from the proposed sequential LRT alone. Thus, we expect that multiple-testing issue should not be a big concern in sequential LRT procedure.

For method comparison, the mmLD showed either higher or almost equal power performance compared to SKAT_C [19, 20] and adjusted single-marker association test (smLD, minimum p-value smLD), in both scenarios when QTL is either genotyped or non-genotyped. We expect that the mmLD can be a useful tool for future GWAS.

One important assumption in the mmLD is that the QTL is in linkage disequilibrium with its adjacent markers. Hence, it is best applicable for the QTL detection of inheritable traits. For newly occurring somatic mutations that are not strongly related to genetic markers, we expect the mmLD would not perform very well.

Chapter 5 Further Study

In this dissertation, I suggested a novel model to detect the existence of the unknown QTL with multiple adjacent markers. However, there are still several issues to be studied. Below are several directions for future studies.

(1) In this study, we mainly focus on the continuous phenotypic traits. However the idea of mmLD can also be extended to other important trait types, such as binary data for case-control studies, or longitudinal traits for development. Extension to other biological traits will greatly enhance the applicability of the mmLD.

(2) The current mmLD framework assumes only one QTL in a LD block, which may not be true in real data. For example, mutations of the same gene at different location may have the same biological consequences. So if there are indeed two or more QTL in one LD block, how to efficiently and effectively detecting them is also a question of interest.

(3) It is well known that genes form a network to function together. In the current mmLD mapping, only the marginal effect of each QTL is considered. It would be very interesting to extend this framework to incorporate the gene-gene interaction, or epistasis effects into the mmLD framework.

REFERENCES

1. Medicine, U.S.N.L.o. *What is a gene?* July 13, 2015 [cited 2015 07/20/2015]; Available from: http://ghr.nlm.nih.gov/handbook/basics/gene.

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

3. Wu, R., C.-X. Ma, and G. Casella, *Statistical genetics of quantitative traits : linkage, maps, and QTL*. 2007, New York: Springer. xvi, 365 p.

4. *International HapMap Project*. Available from: http://hapmap.ncbi.nlm.nih.gov/.

5. Reich, D.E., et al., *Linkage disequilibrium in the human genome.* Nature, 2001. **411**(6834): p. 199-204.

6. Lambert, B.W., J.D. Terwilliger, and K.M. Weiss, *ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth.* Bioinformatics, 2008. **24**(16): p. 1821-2.

7. Stern, C., *The Hardy-Weinberg Law.* Science, 1943. **97**(2510): p. 137-8.

8. Castle, W.E., *The laws of heredity of Galton and mEndel, and some laws governing improvement by selection.* Proceedings of the American Academy of Arts and Sciences, 1903. **39**(1/12): p. 223-242.

9. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.

10. Wu, S., *A robust approach for genetic mapping of complex traits*

2008, University of Florida: 3381465. p. 137.

11. Excoffier, L. and M. Slatkin, *Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.* Mol Biol Evol, 1995. **12**(5): p. 921-7.

12. Stephens, M., N.J. Smith, and P. Donnelly, *A new statistical method for haplotype reconstruction from population data.* Am J Hum Genet, 2001. **68**(4): p. 978-89.

13. Wray, N.R., *Estimating Trait Heritability.* Nature Education 2008. **1**(1).

14. Balding, D.J., *A tutorial on statistical methods for population association studies.* Nat Rev Genet, 2006. **7**(10): p. 781-91.

15. Wu, R., C.X. Ma, and G. Casella, *Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations.* Genetics, 2002. **160**(2): p. 779-92.

16. Wang, Z. and R. Wu, *A statistical model for high-resolution mapping of quantitative trait loci determining HIV dynamics.* Stat Med, 2004. **23**(19): p. 3033-51.

17. Wu, R. and Z.B. Zeng, *Joint linkage and linkage disequilibrium mapping in natural populations.* Genetics, 2001. **157**(2): p. 899-909.

18. Yang, J., et al., *Genome-wide two-marker linkage disequilibrium mapping of quantitative trait loci.* BMC Genet, 2014. **15**: p. 20.

19. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet, 2011. **89**(1): p. 82-93.

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

21. Liu, J., et al., *Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method.* Stat Interface, 2013. **6**(1): p. 99-115.

22. Zhang, C.H., *Nearly Unbiased Variable Selection under Minimax Concave Penalty.* Annals of Statistics, 2010. **38**(2): p. 894-942.

23. Wu, S., J. Yang, and R. Wu, *Semiparametric functional mapping of quantitative trait loci governing long-term HIV dynamics.* Bioinformatics, 2007. **23**(13): p. i569-76.

24. Lawrence, E., *Henderson's dictionary of biology*. 14th ed. 2008, Harlow, England ; New York: Pearson Benjamin Cummings Prentice Hall. xii, 759 p.

25. Lou, X.Y., et al., *A haplotype-based algorithm for multilocus linkage disequilibrium mapping of quantitative trait loci with epistasis.* Genetics, 2003. **163**(4): p. 1533-48.

26. Sinnwell JP, S.D. *haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous*. 2013 Available from: http://cran.r-project.org/web/packages/haplo.stats/index.html.

27. Ott, J. *User's Guide to the Estimating Haplotype program*. 2013; Available from: http://www.jurgott.org/linkage/eh.htm.

28. Clayton, D. *SNPHAP*. 2011; Available from: https://www-gene.cimr.cam.ac.uk/staff/clayton/software/.

29. Kempthorne, O., *An introduction to genetic statistics*. Wiley publications in statistics. 1957, New York,: Wiley. 545 p.

30. Ionita-Laza, I., et al., *Sequence kernel association tests for the combined effect of rare and common variants.* Am J Hum Genet, 2013. **92**(6): p. 841-53.

31. Institute, T.B.R. *GAW17 Data*. 2010; GAW grant, R01 GM031575]. Available from: http://www.gaworkshop.org/gaw17/data.html.

32. Genomes. *A Deep Catalog of Human Genetic Variation*. NIH R01 MH059490]. Available from: http://www.1000genomes.org/.