This is to certify that the
dissertation entitled

VARIANCE COMPONENT MODELS IN MAPPING IMPRINTED GENES: STATISTICAL THEORY AND APPLICATIONS

presented by

Gengxin Li

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Statistics

Major Professor's Signature

6/29/2010

Date

VARIANCE COMPONENT MODELS IN MAPPING IMPRINTED GENES: STATISTICAL THEORY AND APPLICATIONS

By

Gengxin Li

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Statistics

2010

ABSTRACT

VARIANCE COMPONENT MODELS IN MAPPING IMPRINTED GENES: STATISTICAL THEORY AND APPLICATIONS

By

Gengxin Li

Genomic imprinting has been thought to play an important role in seed development in flowering plants. Seed in a flowering plant normally contains diploid embryo and triploid endosperm. Empirical studies have shown that some economically important endosperm traits are genetically controlled by imprinted genes. However, the exact number and location of imprinted genes are largely unknown due to the lack of efficient statistical mapping methods. When an iQTL segregates in experimental line crosses, combining different line crosses with similar genetic background can improve the accuracy of iQTLs inference. To make full use of the natural information of sex-specific allelic sharing among sibpairs in line crosses, general statistical variance components frameworks are proposed to map imprinted quantitative trait loci (iQTL) for the diploid tissue and the triploid tissue, individually. Considering the special characteristics of the diploid embryo genome and triploid endosperm genome, new variance components partition methods with respect to the diploid and triploid tissues are developed. An extension to multiple QTL analysis is proposed for both diploid and triploid tissues.

A number of studies have demonstrated that multivariate traits analysis can provide more significant power and higher resolution for major gene detection in linkage analysis (Evans 2002). Furthermore, when a QTL has the pleiotropic effect on several

traits, some important biologically interesting hypotheses can be performed successfully under the multivariate traits approach. It is well known that several highly correlated traits appear commonly in endosperm. So the variance components based univariate trait iQTL model is extended to bivariate traits iQTL model for mapping the parent-of-origin effect. It may expedite the process of identifying and eventually cloning genes controlling important endosperm traits.

Except for the wide application of variance components model in flowering plants, variance components analysis has been a standard means in human genetics. In brief, the genetic effect is detected by the significance of the likelihood ratio test. However, true parameters of main interest may be on the boundary of the parameter space under the null hypothesis, thus the regularity condition for declaring asymptotic chi-square distribution of the LRT statistics is not satisfied. The threshold calculation based on current methods often yields conservative hypothesis tests as discussed in a number of studies, especially in multivariate traits cases. To solve this problem, a general approximation form of the LRT under the null hypothesis of no linkage is proposed, and the chi-square mixture proportions are shown to depend on the estimated Fisher information matrix in both univariate and multivariate trait analysis.

COPYRIGHT

ACKNOWLEDGMENT

My academic life at Michigan State University is rewarding under the guidance of my advisor Dr.Yuehua Cui. I would like to express my sincere appreciation to Dr. Yuehua Cui for his valuable advices and encouragements during my Ph.D study at the department of statistics and probability. He introduces a great research area, statistical genetics, to me and makes my research study be full of pleasure. I am also grateful to be awarded two Quantitative Biology fellowships under his strong support. His outstanding performance on both research and teaching areas serves as a perfect example for pursuing my future career goal.

I sincerely thank my co-advisor Dr. Dechun Wang for making me participate in the interdisciplinary research at the department of crop and soil sciences. Dr. Dechun Wang is a professional geneticist in soybean. He provides strong helps for my application in QB interdisciplinary Ph. D program, and supervises me for the biology training in his lab. My academic training in interdisciplinary researches is further broaden under Dr. Dechun Wang's guidance.

I also express my gratitude to Dr. Connie Page, Dr. Sarat Dass, and Dr. Robert Tempelman for being my Ph.D committee members. I appreciate your suggestions for my dissertation. Special thanks to my husband, Shengqi Yao and my parents, Junru Li and Jiarong Fu.

TABLE OF CONTENTS

List of Tables viii

List of Figures x

0.1 Introduction 1
0.1.1 Gene and quantitative trait loci (QTL) 1
0.1.2 Genomic imprinting 3
0.1.3 Imprinting QTL method 4
0.1.4 Objectives and organization of the dissertation 7

1 A statistical variance components framework for mapping imprinted quantitative trait loci in experimental crosses **9**
1.1 Introduction 9
1.2 Statistical Methods 13
1.2.1 Genetic Design 13
1.2.2 The mixed-effect variance components model 16
1.2.3 Parent-specific allele sharing and covariances between two inbreeding full-sibs 19
1.2.4 Likelihood function and parameter estimation 25
1.2.4.1 The ML estimation 26
1.2.4.2 The REML Estimation 30
1.2.5 QTL IBD sharing and genomewide linkage scan 32
1.2.6 Hypothesis testing 34
1.2.7 Multiple QTL model 37
1.3 Results 41
1.3.1 Simulation design 41
1.3.2 Simulation results 44
1.3.2.1 Single QTL analysis 44
1.3.2.2 Multiple QTL analysis 52
1.4 Discussion 56

2 A general statistical framework for dissecting parent-of-origin effects underlying endosperm traits in flowering plants **61**
2.1 INTRODUCTION 61
2.2 STATISTICAL METHOD 64
2.2.1 The genetic design 64
2.2.2 The model 67
2.2.3 Parent-specific allele sharing and the covariance between two inbreeding sibs 68

2.2.4 QTL IBD sharing and genome-wide linkage scan 74
2.2.5 Hypothesis testing 75
2.2.6 Multiple iQTL model 78
2.3 SIMULATION 80
2.3.1 Single iQTL simulation 80
2.3.2 Multiple iQTL simulation 85
2.4 A CASE STUDY 87
2.5 DISCUSSION 96

3 Bivariate quantitative trait linkage analysis in mapping imprinted quantitative trait loci underlying endosperm traits in flowering plant 104
3.1 Introduction 104
3.2 Statistical method 108
3.2.1 The model 108
3.2.2 Parent-specific allele sharing & genomewide linkage scan . . . 110
3.2.3 Likelihood function and parameter estimation 111
3.2.3.1 The ML estimation 112
3.2.3.2 The REML Estimation 115
3.2.4 Hypothesis testing 119
3.3 Simulation 123
3.3.1 Simulation design 123
3.3.2 Results 124
3.4 Real Data Analysis 128
3.5 Discussion 131

4 Assessing statistical significance in genetic linkage analysis with the variance components model 134
4.1 Introduction 134
4.2 Motivating models 136
4.2.1 Model I 136
4.2.2 Model II 138
4.2.3 Model III 139
4.3 Main results 141
4.4 Simulation 162
4.5 Conclusion 162

5 Concluding remarks 167

LIST OF TABLES

1.1 The IBD sharing coefficients for full-sib pairs in a reciprocal backcross design considering allelic parental origin 15

1.2 The power, MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL with no imprinting effect under different sampling designs. 43

1.3 The power, MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL showing different imprinting effects under the 20×20 sampling design. The square roots of the mean squared errors of the parameters are given in parentheses. 48

1.4 The power, MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for data simulated with Mendelian and imprinting models based on the 20×20 sampling design. 50

1.5 The MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for data simulated with two QTLs under the 20×20 design. The square roots of the mean squared errors are given in parentheses. 53

2.1 The allelic-specific IBD sharing coefficients for full-sib pairs in a reciprocal backcross design. 66

2.2 The power, REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL with no imprinting effect under different sampling designs. The square roots of the mean squared errors of the parameter estimates are given in parentheses. . 81

2.3 The power, REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL showing different imprinting effects under the 20×20 sampling design. The square roots of the mean squared errors of the parameters are given in parentheses. 84

2.4 The estimated parameters for the three maternal effects and the variance components for two endosperm traits: mean ploidy (Mploidy) and percent of the endoreduplicated nuclei (Endo). 91

3.1 The power, REMLs of the QTL position estimated based on 100 simulation replicates for a QTL with no imprinting effect under 4 × 100 design. The square roots of the mean squared errors of the parameter estimates are given in parentheses. 125

3.2 The power, REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL showing different imprinting effects under the 4×100 sampling design. The square roots of the mean squared errors of the parameters are given in parentheses. 127

3.3 The estimated parameters for the variance components for joint bivariate endosperm traits: mean ploidy (Mploidy) and percent of the endoreduplicated nuclei (Endo). 130

4.1 The whole possible subsets of $\psi_{\nu^\vee|\mathbf{Y}}$ without covariance terms. . . . 147

4.2 The whole possible subsets of $\psi_{\nu^\vee|\mathbf{Y}}$ with covariance terms. 149

4.3 Comparisons of the performance of different approximation methods based on 1000 simulation replicates under different models. 163

LIST OF FIGURES

1.1 Possible alleles shared IBD for individuals i and j in inbreeding backcross families. Lines indicate alleles shared IBD. 21

1.2 The LR profile plot. The left and right figures correspond to the LR profiles generated using the ML and REML method, respectively. The arrow indicates the true QTL position. 46

1.3 The LR profile plot for singe QTL and multiple QTL analysis. The true QTL positions are simulated at $28cM$ and $68cM$ (see the arrow sign). The dotted curve and the solid curve represent the LR profiles by single QTL and multiple QTL analysis, respectively. The left and right figures correspond to the LR profiles generated using the ML and REML method, respectively. 55

2.1 Possible alleles shared IBD for individuals i and j in inbreeding backcross families. The solid lines indicate IBD sharing for alleles inherited from the same parent. The dotted lines indicate IBD cross-sharing for alleles inherited from different parents. 70

2.2 The LR profile plot for the single iQTL and multiple iQTL analyses. The true iQTL positions are simulated at $28cM$ and $72cM$ (see the arrow signs). The dotted and the solid curves represent the LR profiles by the single iQTL and multiple iQTL analyses, respectively. 86

2.3 The profile of the log-likelihood ratios (LR) for testing the existence of QTLs underlying the two endosperm traits across the 10 maize linkage groups ($G_1, \cdots, G_{10}$). The genome-wide LR profiles for the percentage of endoreduplication (Endo) and mean ploidy (Mploidy) traits are indicated by solid and dotted curves, respectively. The threshold values for claiming the existence of QTLs are given as the horizonal solid and dotted line for the genome-wide threshold, dashed and dash-dotted line for the chromosome-wide threshold, for the two traits Endo and Mploidy, respectively. The genomic positions corresponding to the peak of the curves that pass the corresponding thresholds are the MLEs of the QTL location. The positions of markers on the linkage groups (Coelho et al. 2007) are indicated at ticks. 88

2.4 The profile of the log-likelihood ratios (LR) for testing the existence of QTLs underlying the trait Mean Ploidy Level across the 10 maize linkage groups ($G_1, \cdots, G_{10}$). 94

2.5 The profile of LR values for testing the existence of QTLs underlying the trait Percentage of Endoreduplication across the 10 maize linkage groups ($G_1, \cdots, G_{10}$). See Figure 2.4 for more explanations of the figure. 95

3.1 The profile of the log-likelihood ratios (LR) for testing the existence of QTLs underlying the two endosperm traits across the 10 maize linkage groups ($G_1, \cdots, G_{10}$). 129

4.1 The quantile plot of the empirical p-values for Model I. For the legend: Self & Liang refers to SF; Proposed refers to the current method. See Table 4.3 for more explanation of the legend. 164

4.2 The quantile plot of the empirical p-values for Model II. For the legend: Self & Liang refers to SF; Proposed refers to the current method. See Table 4.3 for more explanation of the legend. 165

4.3 The quantile plot of the empirical p-values for Model III. For the legend: Self & Liang refers to SF; Proposed refers to the current method. See Table 4.3 for more explanation of the legend. 166

0.1 Introduction

0.1.1 Gene and quantitative trait loci (QTL)

Gregor Mendel first studied certain genetic traits to discover the inheritance of biological variations in peas. A gene is responsible for inheriting these special traits from parents. With the exploration of the DNA structure, a gene is normally defined as a stretch of DNA that acts on the protein or an RNA chain to issue instructions for a special function. For example, DMPK gene can produce a unique protein, myotonic dystrophy protein kinase, to guarantee the normal function of muscle, heart, and brain cells. There are around 30,000 protein-coding genes in human that work together to control most functions in human body. An allele is a copy of a gene that measures the variation of the DNA sequence. Usually, a gene **A** is made up of two alleles A and a. Three genetic compositions (AA, Aa, and aa) made up of two alleles (A and a) are defined as genotypes. In fact, humans share mostly the same genes with distinct combinations of alleles that make him or her genetically unique. For instance, the hair color is controlled by same genes in human, but the specific hair color, such as: red, black, blonde, and so on, is determined by different alleles combined in the same genes. Besides, genes may affect many important quantitative traits, for instance, body weight, body height, blood pressure, and so on.

Inheritance of characteristics of quantitative traits is attributed to single gene or multiple genes interacting with environmental factors. Thus, quantitative trait loci (QTL) is detectable regions of the genome that are closely linked to genes associated

with variations of quantitative traits. The association between quantitative trait loci and closely linked genes in the same chromosome is termed the genetic linkage. The recombination fraction measures the degree of this association and is utilized to create a genetic linkage map. In brief, the recombination is a process through which a chromosomal crossover happens between two QTL or genes during the meiosis. The mean number of crossovers is called map distance, such as: one centimorgan (cM) is equivalent to a recombination fraction of 1%. Because of unobservability of the QTL genotype, the closely linked neutral molecular markers is used to predict the genotype of QTL. A genetic marker is a DNA sequence that is the unit component of one chromosome. Associated with a certain locus, genetic markers are easily identifiable and highly polymorphic. Their exact locations on a chromosome can be estimated. In fact, a statistical model is built to connect the QTL genotypes and marker genotypes through phenotypes to identify and sequence genes.

So far, scientists have identified more than 10,000 mouse genes. Because mice and humans share around 95 percent identical sequence and possess same organs, more than 500 mouse models with respect to human diseases including cancer and diabetes have been developed. Many successfully developed gene techniques in mice have allowed scientists to investigate the human disease on animal models. More and more people recognize that the age of genetic medicine begins.

0.1.

0.1.2 Genomic imprinting

It is well-known that two alleles of a gene inherited from both parents affect variations of the DNA sequence jointly. If only one parental derived allele is associated with the variation of phenotype, and the other allele is unexpressed, this special epigenetic phenomena (uniparental gene expression) is termed genomic imprinting (Wolf et al. 2008). Under genomic imprinting, the expression of the same allele A from different heterozygote genotypes Aa and aA depends on the origin of inheritance of this allele. Then the maternally derived allele A (from Aa) functions differently from that of paternally derived allele A (inherited from aA). There are two types of imprinted genes, that is, one gene is maternally imprinted when the paternal copy is expressed with silent maternal copy, and a gene is paternally imprinted gene if the maternal copy is expressed with silent paternal copy. Genomic imprinting is first used to describe the elimination of paternal chromosome for spermatogenesis in sciarid flies. With the investigation of genomic imprinting, scientists have found that the DNA methylation and histone modifications are the vital mechanism to result in imprinting (Feil and Berger 2007). During this mechanism, imprinted genes are expressed differently in egg and sperm, and the different gene expression is caused by the inheritance of these epigenetic phenomena. In the healthy genome, even a mutation happens on one allele of a gene, the other allele can still be transcribed to pay off the loss from the mutation. But, if the epigenetic event takes place on the same gene, only the mutated allele is expressed, then people get a disease because of the imprinting effect. Thus, the epigenetic changes are serious to the disease without changing the genomic

sequences physically.

In the past few years, scientists have made a lot of efforts in the understanding of genomic imprinting. Specifically, the significant phenotypic variations caused by imprinted genes have been confirmed in areas of the fetal growth and behavior. It has been increasingly recognized that imprinted genes may influence cancer, obesity, diabetes and many other disease in human and mammal, and many imprinted genes are identified to regulate embryonic development in plants. For example, Prader-Willi Syndrome, a genetic disease, makes patients to be extremely fat. It is caused by the deletion of 7 genes on the paternal chromosome 15 where the maternal copy is silent. Besides, other severe genetic diseases caused by the imprinting effect are Embryonal rhabdomyosarcoma for kidney cancer, Osteosarcoma for bone cancer, and Angelman syndrome for delayed development, and so on. In maize endosperm, imprinted genes are thought to control the endoreduplication (Dilkes et al. 2002) procedure through which larger fruits or seeds are obtained (Grime and Mowforth 1982). The disrupted gene (IGF2) encoding paternally transmitted insulin-like growth factor II results in growth deficiency in mice (DeChiara et al. 1991). Currently more than 600 imprinted genes have been predicted in mouse genome (Luedi et al. 2005). But the accurate locations and the genetic effect of most imprinted genes remain largely unknown.

0.1.3 Imprinting QTL method

From a quantitative genetic theory point of view, imprinting results in genetic gain and evolutionarily favorable. Considering a gene **A** with two alleles A and a, the allele

frequency of A is p, and for a is q. Because of genomic imprinting, heterozygotes Aa and aA are expressed differently, then distinct genotypic values can be defined by the additional imprinting effect i When $i = 0$, the model is reduced to the traditional

Genotype	Frequency	value
AA	p^2	a
Aa	pq	$d+i$
aA	pq	$d-i$
aa	q^2	-a

Mendelian model. Simple algebra shows that the genetic variance with and without imprinting is given as

$$\sigma^2_{g_i} = 2pq\alpha_i^2 + (2pqd)^2 + 2pqi^2, \qquad \text{Imprinting}$$

$$\sigma^2_g = 2pq\alpha^2 + (2pqd)^2, \qquad \text{No imprinting}$$

where α_i and α are the average effects with respect to imprinting and no-imprinting, respectively. The additional variance term $2pqi^2$ due to imprinting is always non-negative. Thus, imprinting leads to increased genetic variance and is evolutionarily favorable. This explains why after so many years' natural selection, genomic imprinting is still preserved.

The imprinted inheritance violates the Mendelian theory and brings challenges in statistical modelling. The statistical framework in mapping imprinted genes or QTL was initiated with a fixed effect model in which the genetic effect is treated as a fixed term. Many studies under this framework were developed to test imprinted QTL

with controlled crosses of outbred parents (Knott et al. 1998; de Koning et al. 2000 2002). But, the allelic heterozygosity of two outbred parents may induce confounding effects for genomic imprinting. The genetic difference based on these methods may not be explained by the real imprinting effect (Lin et al. 2003). When backcross and F_2 populations with inbred lines were analyzed, the regression-based maximum likelihood approaches in mapping the imprinted QTL were proposed (Cui 2006; Cui et al. 2006, 2007). It has been shown that methods focusing on genetic variances are more powerful to infer QTL effects than the allele substitution method assuming a fixed effect (Xie et al. 1998). When an iQTL segregates in multiple line crosses, the detection of iQTL may be improved by combining different line crosses with similar genetic background. However, no studies based on the variance components method have been proposed to identify iQTLs with multiple line crosses.

The variance components method is based on the identical-by-decent (IBD) principle in which sib pairs have more similar phenotypic trait values when they share more proportion of alleles IBD. Variance components model in mapping the parent-of-origin effect in human was first proposed by Hanson et al. (2001). In this approach, the additive genetic variance is decomposed into two terms, a component due to the expression of the maternal allele and a component due to that of the paternal allele. However, the direct application of this variance components method to a fully or partially inbreeding population is infeasible. The structure of inbreeding populations is more complicated than that of non-inbreeding populations. Constructing a variance components method based on inbred populations is still a challenging problem.

Endosperm in flowering plants is developed from the process of double fertilization, and ended up with a triploid tissue. A number of studies have shown that many endosperm traits are affected by genomic imprinting. Statistical methods based on the fixed effect model were proposed to map Mendelian QTL controlling endosperm traits (Wu et al. 2002; Xu et al. 2003; Cui et al. 2005, 2006). However, no studies are investigated for mapping imprinted QTL in endosperm inbreeding population due to the difficulty in modeling the inheritance patterns in a triploid organism with imprinting. In a collaboration with scientists, a data set has been generated for the purpose of identifying imprinted genes controlling for endosperm development. This example motivates us to develop efficient methods while considering the unique genetic structure of a triploid tissue.

0.1.4 Objectives and organization of the dissertation

In this dissertation, I will focus on developing efficient variance components models for the purpose of identifying imprinted genes in experimental crosses. Major goals of this dissertation are summarized as follows:

- Propose a general statistical variance components framework by utilizing the natural information of sex-specific allelic sharing among sib pairs in line crosses, to map imprinted quantitative trait loci (iQTL) underlying traits in a diploid mapping population.

- Extend the method to map iQTLs underlying endosperm traits.

- Extend the single trait model to multi-trait analysis for mapping iQTL underlying bivariate or highly correlated endosperm traits. New biologically interesting hypotheses, such as, testing the pleiotropic effect of (i)QTL or testing pleiotropic effect against close linkage will be designed.

- Conduct a theoretical investigation of the likelihood ratio test (LRT) under the proposed mapping framework.

The dissertation is organized as follows. Chapter 1 will illustrate the variance components based statistical mapping framework for diploid inbreeding populations. The variance components based iQTL mapping approach for the triploid endosperm will be discussed in Chapter 2. The predominance of the bivariate trait analysis will be studied in chapter 3. The asymptotic properties of the likelihood ratio test under the variance components model will be investigated in chapter 4, followed by the final concluding remarks in chapter 5.

Chapter 1

A statistical variance components framework for mapping imprinted quantitative trait loci in experimental crosses

1.1 Introduction

The genetic architecture of complex phenotypes in agriculture, evolution and biomedicine are generally complex involving a network of multiple genetic and environmental factors that interact with one another in complicated ways (Lynch and Walsh 1998). The development of molecular markers makes it possible to identify genetic loci (i.e., quantitative trait loci or QTLs) underlie various traits of interest. Genetic designs

with controlled crosses are generally pursued to generate mapping populations aimed to identify QTLs underlying the variation of phenotypes. Statistical method for QTL mapping with experimental crosses dates back to the seminal work of Lander and Botstein (1989). Various extensions have been developed since then (e.g., Zeng 1994; Kao et al. 1999).

For a diploid organism, the expression products of most functional regions from each one of a chromosome pair are equal. A broken of this equivalence, that is, nonequivalent genetic contribution of each parental genome to offspring phenotype, can result in *genomic imprinting*, a phenomenon also called parent-of-origin effect (Pfeifer 2000). Since its discovery, imprinting-like phenomena have been commonly observed in mammals and seed plants (reviewed by Burt and Trivers 2006). However, statistical methods for identifying imprinted genes have not been extensively studied and well developed.

The imprinted inheritance violates the Mendelian theory and brings challenges in statistical modelling. Currently there are two frameworks in mapping imprinted genes. One is based on the random effect model with pedigree-based natural population such as humans. Hanson et al. (2001) first proposed a variance components framework by partitioning the additive variance component as two parts, a component due to maternal gene and a component due to paternal gene. The variance component method is developed based on the identical-by-decent (IBD) idea in which the expression of the gene for a pair of individuals is expected to be similar if they share alleles IBD. Liu et al. (2007) recently applied the model to map iQTL underlying ca-

nine hip dysplasia in a structured canine population. However, the current IBD-based variance components method for mapping imprinted genes assumes non-inbreeding population. Their applications are immediately limited with fully or partially inbreeding population such as the controlled inbreeding design in plants and animals. With inbred mapping population in humans, Abney et al. (2000) proposed a method to estimate variance components of quantitative traits. However, the extension of the method to map imprinted gene is not straightforward. No variance components method has been proposed to map imprinted genes with inbred population in the literature.

Another general framework for mapping imprinted genes is based on the fixed-effect model in which the effects of genetic factors are considered as fixed. A number of studies were proposed under this framework for mapping imprinted QTL (iQTL) with controlled crosses of outbred parents (Knott et al. 1998; Koning et al. 2000; Koning et al. 2002). One potential limitation of these methods is that allelic heterozygosity at a locus between two outbred parents could cause confounding effects for genomic imprinting. The genetic differences detected by such a fixed-effect model could be caused by allelic heterozygosity of the parents rather than the imprinted effect of iQTL (Lin et al. 2003). A natural alternative for the mapping population is the inbred lines. Fixed-effect models based on backcross and F_2 population were recently proposed under the maximum likelihood framework (Cui 2007; Cui et al. 2006 2007; Li et al. 2008). When inbred lines are used, Xie et al. (1998) pointed out that it is more meaningful to inference QTL effect by its variance rather than by the

allele substitution effect. The QTL variance is generally calculated conditional on the cross, and it, as a variable, is different from one cross to another (Xie et al. 1998). In a single line cross the estimated QTL variance can not be simply extended to a statistical inference space beyond that (Xie et al. 1998). Multiple parental lines are needed for QTL variance inference. A solution to this is to combine data from multiple line crosses (Xie et al. 1998). An IBD-based variance component method was proposed by Xie et al. (1998) with multiple line crosses. Extension of the IBD-based variance component method with multiple line crosses to iQTL mapping has not been studied.

Motivated by the limitations of current methods aforementioned and by the pressing need for efficient iQTL mapping procedure, in this article, we propose a statistical variance components framework for iQTL mapping by combining data from multiple inbred line crosses. The proposed model is robust in iQTL variance inference by extending the iQTL inference space from single line cross to multiple line crosses. A parent-specific IBD sharing partition method is proposed by considering the inbreeding structure in line crosses. As discussed in Cui (2007), the phenotype of an offspring is not only controlled by its own genetic profiles, but also by maternal genotype. The effect of maternal genotype on the phenotype of her offspring, termed maternal effect, is one potential source of confounding effect in the inference of genomic imprinting. The existence of such parental effect may lead to incorrect interpretations of imprinting when they are not properly accounted for in the analysis. Parameters that model the maternal effect are also included and adjusted when testing imprinting.

With the developed model, we propose an interval-based method for genomewide scan and testing of iQTL. Both maximum likelihood (ML) and restricted maximum likelihood (REML) methods are proposed and compared for parameter estimation and power analysis. An extension to multiple QTL is also proposed in which the multiple QTL model provides improved resolution for QTL inference. Extensive simulations are conducted to compare the performance of the proposed model under different sampling designs with different combinations of family and offspring size. Comparisons of the ML and REML methods, single QTL and multiple QTL methods are discussed. The proposed method provides a general framework in iQTL mapping with multiple line crosses and has significant implications in real application.

1.2 Statistical Methods

1.2.1 Genetic Design

The dissection of imprinting effects in line crosses depends on appropriate mating designs where the allele parental origin can be traced and distinguished. Most commonly used inbred line crosses are the backcross, F_2 and recombinant inbred line (RIL). Reciprocal backcross design has been proposed in iQTL mapping (Cui 2007; Cui et al. 2007). Considering parental origin of an allele, we use the subscripts m and f to refer an allele inherited from the maternal and paternal parents, respectively. The merit of a backcross design is that two reciprocal heterozygotes in offsprings, $A_m a_f$ and $a_m A_f$, can be distinguished and their mean effects can be estimated and

tested to assess imprinting (Cui 2007; Cui et al. 2007). While all individuals in an F_2 segregation population share the same parental information, theoretically it is impossible to distinguish the phenotypic distribution of $A_m a_f$ and $a_m A_f$ without extra information. Considering sex-specific recombination rates, Cui et al. (2006) recently developed an imprinting model by incorporating this information into an interval mapping framework. No study has been reported to use RILs for iQTL mapping.

The methods proposed in Cui (2007) and Cui et al. (2007) are fixed-effects QTL models where the effects of an iQTL are considered as fixed. While only four backcross families are considered, when extending to multiple backcross families, the inference of iQTL variance calculation is less efficient. The variance components method, initially proposed in human linkage analysis (Amos 1994) offers a powerful alternative in assessing genomic imprinting (Hanson et al. 2001). In this paper, we will extend the variance components method to inbred line populations by combining different backcross lines to map iQTL.

Table 1.1: The IBD sharing coefficients for full-sib pairs in a reciprocal backcross design considering allelic parental origin

Backcross	Offspring genotype	Parent-specific IBD sharing						Total IBD	
		π_{mm}		π_{ff}		$\pi_{m/f}$		π	
		Q_mQ_f	Q_mq_f	Q_mQ_f	Q_mq_f	Q_mQ_f	Q_mq_f	Q_mQ_f	Q_mq_f
$\underline{QQ \times Qq}$	Q_mQ_f	0.5	0.5	0.5	0	1	0.5	2	1
	Q_mq_f	0.5	0.5	0	0.5	0.5	0	1	1
		Q_mQ_f	q_mQ_f	Q_mQ_f	q_mQ_f	Q_mQ_f	q_mQ_f	Q_mQ_f	q_mQ_f
$\underline{Qq \times QQ}$	Q_mQ_f	0.5	0	0.5	0.5	1	0.5	2	1
	q_mQ_f	0	0.5	0.5	0.5	0.5	0	1	1
		q_mQ_f	q_mq_f	q_mQ_f	q_mq_f	q_mQ_f	q_mq_f	q_mQ_f	q_mq_f
$\underline{qq \times Qq}$	q_mQ_f	0.5	0.5	0.5	0	0	0.5	1	1
	q_mq_f	0.5	0.5	0	0.5	0.5	1	1	2
		Q_mq_f	q_mq_f	Q_mq_f	q_mq_f	Q_mq_f	q_mq_f	Q_mq_f	q_mq_f
$\underline{Qq \times qq}$	Q_mq_f	0.5	0	0.5	0.5	0	0.5	1	1
	q_mq_f	0	0.5	0.5	0.5	0.5	1	1	2

A typical backcross design often starts with the cross between one of the parental lines and their F_1 progeny to create a segregation population. Then large number of offsprings are collected for QTL mapping. When imprinting effect is considered, reciprocal backcrosses are needed. A basic design framework is illustrated in Table 2.1 in Cui (2007). The two reciprocal backcrosses are treated as the base mapping units. Multiple backcross families are sampled based on these crosses. For simplicity, we sample equal number of families for each backcross category. For example, a sample of 8 families would require two of each of the four backcrosses. Noted that the variance components method assesses the degree of allele sharing among siblings. When it is applied to inbred line crosses, each backcross population is considered as one family and different families are considered as independent. For fixed total sample size, one issue is to assess whether we should sample large number of families each with small offspring size or small number of families each with large offspring size. For example, to sample 400 individuals, shall we sample 4 backcross families each with 100 offsprings or 100 families each with 4 progenies or other sampling strategies? The choice of optimal designs is intensively evaluated through simulations.

1.2.2 The mixed-effect variance components model

Suppose there is a putative QTL with two segregating alleles Q and q, located in an interval responsible for the variation of a quantitative trait. The phenotype, y_{ik}, for individual i measured in backcross family $k(= 1, \cdots, K)$ can be written as a linear

function of QTL, polygene and environmental effects,

$$y_{ik} = \mu + a_{ik} + G_{ik} + e_{ik}, \quad k = 1, \cdots, K; \; i = 1, \cdots, n_k \tag{1.2.1}$$

where n_k is the number of offspring in the kth backcross family; μ denotes the overall mean; a_{ik} is the random additive effect of the major monogenic QTL assuming normal distribution with mean zero; G_{ik} is the polygenic effect that reflects the effects of unlinked genes and is assumed to be normally distributed with mean zero; and $e_{ik} \sim N(0, \sigma_e^2)$ is the random environmental error uncorrelated to other effects. The phenotypic variance-covariance for the kth family can be expressed as,

$$\mathbf{\Sigma_k} = \mathbf{\Pi}_k \sigma_a^2 + \mathbf{\Phi}_g \sigma_g^2 + \mathbf{I}\sigma_e^2 \tag{1.2.2}$$

where σ_a^2 and σ_g^2 are the additive and polygene variances; $\mathbf{\Pi}_k$ is a matrix containing the proportion of marker alleles shared IBD for individuals in the kth backcross family; $\mathbf{\Phi}_g$ is a matrix of the expected proportion of alleles shared IBD, and $\mathbf{I}$ is the identity matrix. The calculation of the IBD sharing matrix with inbred lines can be found in Xie et al. (1998) which is based on the Malécot's coefficient of coancestry (Malécot 1948).

Noted that a backcross offspring with genotype $Q_m q_f$ may be obtained by the $QQ \times Qq$ or the $Qq \times qq$ cross. When there is a significant maternal effect, the mean expression for genotype $Q_m q_f$ may be different depending on whether its maternal parents carrying QQ or Qq genotype. As described in Cui (2007), maternal effect

is one source of potential confounding factor for genomic imprinting. It should be appropriately modeled and adjusted when testing imprinting. Here, we model the cytoplasmic maternal effects as fixed effects, and the overall mean μ is replaced by μ_k which models the maternal effect of the kth distinct backcross family.

To accommodate parent-of-origin effects, the QTL additive effect (a) can be partitioned as two terms: (1) a component that reflects the influence of the QTL carried on the maternally derived chromosome (a_m); and (2) a component that reflects the influence of the QTL carried on the paternally derived chromosome (a_f). The model that accommodates the parent-specific effects can be expressed as,

$$y_{ik} = \mu_k + a_{ikm} + a_{ikf} + G_{ik} + e_{ik}, \quad k = 1, \cdots, K; \; i = 1, \cdots, n_k$$

For data vector $\mathbf{y}$ in family k, the above model can be re-expressed as,

$$\mathbf{y}_k = X_k\beta + \mathbf{a}_{km} + \mathbf{a}_{kf} + \mathbf{G}_k + \mathbf{e}_k, \quad k = 1, \cdots, K \tag{1.2.3}$$

where X_k is an indicator matrix corresponding to the kth backcross family and β contains parameters associated with the three maternal effects; $\mathbf{a}_{km} \sim N(\mathbf{0}, \mathbf{\Pi}_{m|k}\sigma_m^2)$, $\mathbf{a}_{kf} \sim N(\mathbf{0}, \mathbf{\Pi}_{f|k}\sigma_f^2)$, $\mathbf{G}_k \sim N(\mathbf{0}, \mathbf{\Phi}_g\sigma_g^2)$, $\mathbf{e}_k \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where $\mathbf{\Pi}_{m|k}$ and $\mathbf{\Pi}_{f|k}$ are matrices containing the proportion of marker alleles shared IBD that are derived from the mother and father, respectively; $\mathbf{\Phi}_g$ is a matrix of the expected proportion of alleles shared IBD, and $\mathbf{I}$ is the identity matrix; σ_m^2 and σ_f^2 are the variance of alleles inherited from the maternal and paternal parents, respectively.

With non-inbreeding mapping population, Hanson et al. (2001) expressed the phenotypic variance-covariance for the kth family as,

$$\mathbf{\Sigma_k} = \mathbf{\Pi}_{m|k}\sigma_m^2 + \mathbf{\Pi}_{f|k}\sigma_f^2 + \mathbf{\Phi}_g\sigma_g^2 + \mathbf{I}\sigma_e^2 \tag{1.2.4}$$

However, for an inbred mapping population, this IBD-based variance partition method can not be directly applied. New method considering the inbreeding structure is needed.

1.2.3 Parent-specific allele sharing and covariances between two inbreeding full-sibs

Before we get the phenotypic variance-covariance of a pair of individuals i and j, let us first consider the parent-specific allele sharing status. Within each BC family, there are two alleles segregating at each locus. Because of inbreeding, the IBD values between two backcross individuals are different from those calculated from outbred full-sibs. Consider two sibs i and j in the kth backcross family. Without considering allelic parental origin, Xie et al. (1998) proposed to calculate the IBD value at a QTL as,

$$\pi_{ij} = 2\theta_{ij} = \begin{cases} 2 \text{ for } QQ - QQ \\ \\ 1 \text{ for } QQ - Qq \text{ or } Qq - Qq \end{cases} \tag{1.2.5}$$

with θ_{ij} being the Malécot's coefficient of coancestry (Malécot 1948). Thus, for an inbred population, π_{ij} is not the actual IBD value between individuals i and j, rather

interpreted as twice the coefficient of coancestry (Xie et al. 1998; Harris 1964). For individuals with itself,

$$\pi_{ii} = 1 + F_i = \begin{cases} 2 \text{ for } QQ - QQ \\ \\ 1 \text{ for } Qq - Qq \end{cases} \tag{1.2.6}$$

where F_i is the inbreeding coefficient for individual i at the QTL. The elements in $\mathbf{\Phi}_g$ matrix are just the expected values of π_{ij} and π_{ii} which are ϕ_{ij}=5/4 and $\phi_{ii} = 3/2$ (Xie et al. 1998).

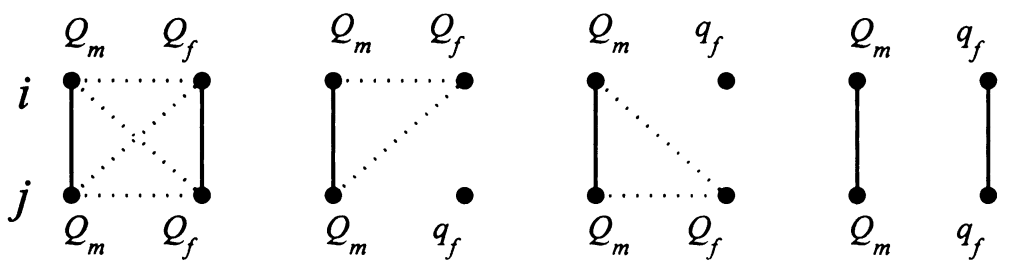

Figure 1.1: Possible alleles shared IBD for individuals i and j in inbreeding backcross families. Lines indicate alleles shared IBD.

When allelic parental origin is considered, the IBD sharing matrix can also be calculated based on the coefficient of coancestry. By definition, the coefficient of coancestry is defined as the probability that two randomly drawn alleles from individuals i and j are identical by descent. Fig. 1.1 displays possible alleles shared IBD for sibs drawn in backcross families. Consider two backcross individuals i (with two alleles A_{im} and A_{if}) and j (with two alleles A_{jm} and A_{jf}). Define θ_{ij} as the coefficient of coancestry between individuals i and j. By definition, θ_{ij} can be calculated as,

$$\begin{aligned}\theta_{ij} &= \frac{1}{4}\{\Pr(A_{im}=A_{jm})+\Pr(A_{im}=A_{jf})+\Pr(A_{if}=A_{jm})+\Pr(A_{if}=A_{jf})\} \\ &= \frac{1}{4}(\theta_{imjm}+\theta_{imjf}+\theta_{ifjm}+\theta_{ifjf})\end{aligned}$$

where $\theta_{i.j.}$ can be interpreted as the allelic kinship coefficient, i.e., the probability that a randomly chosen allele from individual i is IBD to a randomly chosen allele from individual j. Note that the two terms θ_{imjf} and θ_{ifjm} are not distinguishable. However, their sum is unique and therefore the two terms can be combined as one single term, denoted as $\theta_{im/jf}(=\theta_{imjf}+\theta_{ifjm})$. After the manipulation, the coefficient of coancestry for individuals i and j can be expressed as $\theta_{ij}=\frac{1}{4}(\theta_{imjm}+\theta_{im/jf}+\theta_{ifjf})$ which is composed of three components.

Following Xie et al. (1998), the alleles shared IBD between individuals i and j can be expressed as,

$$\begin{aligned}\pi_{ij} = 2\theta_{ij} &= \frac{1}{2}(\theta_{i_m j_m} + \theta_{i_m/j_f} + \theta_{i_f j_f}) \\ &= \pi_{i_m j_m} + \pi_{i_m/j_f} + \pi_{i_f j_f} \qquad (1.2.7)\end{aligned}$$

where $\pi_{i_m j_m} = \frac{1}{2}\theta_{i_m j_m}$ and $\pi_{i_f j_f} = \frac{1}{2}\theta_{i_f j_f}$ are the alleles shared IBD derived from the mother and father, respectively; $\pi_{i_m/j_f} = \frac{1}{2}\theta_{i_m/j_f}$ is the alleles shared IBD due to alleles cross sharing, a special case for inbreeding sibs. Without inbreeding, π_{i_m/j_f} takes value of zero.

For completely inbreeding population, the inbreeding coefficient F_i is 1 if alleles inherited from both parents are the same since these alleles can be traced back to the same grandparent. For example, for an individual with genotype $Q_m Q_f$, $\Pr(Q_m = Q_f) = 1$ since both alleles Q_m and Q_f are inherited from the same grandparent. Therefore, for individuals with itself, $\pi_{ii} = 1 + F_i$ would be the same as $\pi_{ij} (i \neq j)$ when i and j carry the same genotypes. The expected proportion of alleles shared IBD ϕ_{ij} can also be calculated.

Thus, the proportion of alleles shared IBD can be partitioned as three components for inbreeding sibs, rather than two components considering parent-of-origin effects proposed by Hanson et al. (2001). To further illustrate the idea, we use one backcross family to demonstrate the derivation. A full list of possible IBD sharing values for the two reciprocal backcrosses are given in Table 2.1. Considering a backcross family initiated with the $Qq \times QQ$ cross. Randomly selecting two individuals i and j with

genotype Q_mQ_f and Q_mQ_f, the Malécot's coefficient of coancestry can be calculated as,

$$\begin{aligned}\pi_{ij} = 2\theta_{ij} &= \frac{1}{2}\{\Pr(Q_{im} = Q_{jm}) + \Pr(Q_{im} = Q_{jf}) + \Pr(Q_{if} = Q_{jm}) \\ &\quad +\Pr(Q_{if} = Q_{jf})\} \\ &= \frac{1}{2}[1 + 1 + 1 + 1] = 2\end{aligned}$$

Thus, $\pi_{i_m j_m} = \pi_{i_f j_f} = 0.5$ and $\pi_{i_m/j_f} = 1$. For sib pairs i (with genotype Q_mQ_f) and j (with genotype Q_mq_f), $\pi_{i_m j_m} = 0.5$, $\pi_{i_f j_f} = 0$ and $\pi_{i_m/j_f} = 0.5$, and $\pi_{ij} = 1$ which is the same as given in (1.2.5) without considering parent-of-origin partition.

Considering the allelic sharing status in a complete inbreeding population, the relationship between the maternal and paternal alleles is no longer independent if the two alleles are in identical form. There exists a covariance term (denoted as σ^2_{mf}) due to alleles cross sharing for two inbreeding full-sibs when calculating the phenotypic variance. Corresponding to the partition of the IBD-sharing considering allelic parental origin, the major QTL additive variance component can be partitioned into three components, i.e., σ^2_f, σ^2_m and σ^2_{mf}, in which σ^2_{mf} can be interpreted as the covariance due to alleles cross sharing in inbreeding families. Thus, the trait covariance between two individuals i and j can be expressed as,

$$Cov(y_i, y_j) = \pi_{i_m j_m}\sigma^2_m + \pi_{i_f j_f}\sigma^2_f + \pi_{i_m/j_f}\sigma^2_{mf} + \phi_{ij}\sigma^2_g + I_{ij}\sigma^2_e$$

where I_{ij} is an indicator variable taking value 1 if $i = j$ and 0 if $i \neq j$. The variance-covariance matrix for a phenotypic vector in the kth backcross family can then be expressed as,

$$\mathbf{\Sigma_k} = \mathbf{\Pi}_{m|k}\sigma^2_m + \mathbf{\Pi}_{m/f|k}\sigma^2_{mf} + \mathbf{\Pi}_{f|k}\sigma^2_f + \mathbf{\Phi}_g\sigma^2_g + \mathbf{I}\sigma^2_e \qquad (1.2.8)$$

where the elements of $\mathbf{\Pi}_{m|k}$, $\mathbf{\Pi}_{f|k}$ and $\mathbf{\Pi}_{m/f|k}$ can be found in Table 2.1.

For non-inbreeding sib pairs with random mating, $\pi_{i_m/j_f} = 0$ and hence $\text{Cov}(a_m, a_f) = 0$. Model (2.2.4) reduces to $\mathbf{\Sigma_k} = \mathbf{\Pi}_{m|k}\sigma^2_m + \mathbf{\Pi}_{f|k}\sigma^2_f + \mathbf{\Phi}_g\sigma^2_g + \mathbf{I}\sigma^2_e$, the same as the variance components partition model considering parent-of-origin effects given in Hanson et al. (2001).

1.2.4 Likelihood function and parameter estimation

Assuming multivariate normality, the density function of observing a particular vector of data $\mathbf{y}$ for family k is given by,

$$f(\mathbf{y}_k; \mu_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{n_k/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}_k - \mu_k)^{\mathrm{T}}\mathbf{\Sigma_k}^{-1}(\mathbf{y}_k - \mu_k)\right]$$

where $\mathbf{y}_k = (y_{1k}, \ldots y_{n_k k})^T$ is a n_k x 1 vector of phenotypes for the kth backcross family and n_k is the kth backcross family size. The overall log likelihood function for K independent backcross families is give by,

$$\ell = \sum_{k=1}^{K} \log[f(\mathbf{y}_k; \mu_k, \mathbf{\Sigma}_k)] \qquad (1.2.9)$$

Note that the maternal effect μ_k is the same for families with the same maternal genotype. Thus, only three maternal effects need to be estimated. Two commonly used methods can be applied to estimate parameters in a mixed effects model, the ML method and the REML method. Both methods have been applied in genetic linkage analysis in a variance components model framework (Amos 1994; Almasy and Blangero 1998). In general, ML estimators tend to be downwardly biased given that it does not account for the loss in degrees of freedom resulted from estimation of the fixed effects (Corbeil and Searle 1976). The REML is based on a linear transformation of the data such that the fixed effects are eliminated from the model, hence it provides less biased estimators. Even though standard softwares such as SAS have standard procedures to estimate parameters for a mixed effects model, the estimation for the proposed model can not be directly fitted into a standard software. The estimation procedures for the two methods are detailed here.

1.2.4.1 The ML estimation

The phenotype vector in the kth backcross family follows a multivariate normal distribution, i.e., $\mathbf{y}_k \sim MVN(X_k\beta, \mathbf{\Sigma}_k)$. Parameters that need to be estimated are $\Omega = (\beta, \sigma_m^2, \sigma_f^2, \sigma_{mf}^2, \sigma_g^2, \sigma_e^2)$ with $\beta = (\mu_1, \mu_2, \mu_3)$.

Define $\sigma^2 = \sigma_m^2 + \sigma_f^2 + \sigma_{mf}^2 + \sigma_g^2 + \sigma_e^2$, $h_m^2 = \frac{\sigma_m^2}{\sigma^2}$, $h_f^2 = \frac{\sigma_f^2}{\sigma^2}$, $h_{mf}^2 = \frac{\sigma_{mf}^2}{\sigma^2}$, $h_g^2 = \frac{\sigma_g^2}{\sigma^2}$, and $h_e^2 = 1 - h_m^2 - h_f^2 - h_{mf}^2 - h_g^2$. σ^2 is the total phenotypic variance and hence h_m^2 and h_f^2 can be considered as the heritability of maternal and paternal

alleles, $h_m^2 + h_f^2 + h_{mf}^2$ is the total genetic heritability due to the major QTL, h_g^2 is the polygene heritability and $h^2 = h_m^2 + h_f^2 + h_{mf}^2 + h_g^2$ is the overall heritability. The phenotypic variance-covariance between any two individuals i and j in the kth backcross family can then be re-expressed as:

$$Var \begin{pmatrix} y_{ik} \\ y_{jk} \end{pmatrix} = \sigma^2 H_{ij|k}$$

where

$$H_{ij|k} = \begin{pmatrix} \delta_i & \delta_{ij} \\ \delta_{ij} & \delta_j \end{pmatrix}$$

with $\delta_i = \pi_{i_m i_m} h_m^2 + \pi_{i_f i_f} h_f^2 + \pi_{i_m / i_f} h_{mf}^2 + \phi_{ii} h_g^2 + h_e^2$; δ_j is defined similarly; and $\delta_{ij} = \pi_{i_m j_m} h_m^2 + \pi_{i_f j_f} h_f^2 + \pi_{i_m / j_f} h_{mf}^2 + \phi_{ij} h_g^2$

If there are n_k sibs in each backcross family, $H_k = \{H_{ij|k}\}_{n_k \times n_k}$ is simply a $n_k \times n_k$ matrix. Instead of estimating $\Omega = (\beta, \sigma_m^2, \sigma_f^2, \sigma_{mf}^2, \sigma_g^2, \sigma_e^2)$, we can estimate Ω=(β, σ^2, h_m^2, h_f^2, h_{mf}^2, h_g^2) and solve above equations to get the original variance estimates. Now the log-likelihood can be expressed as,

$$\begin{aligned} \ell(\Omega) &= \sum_{k=1}^{K} \log[f(\mathbf{y}_k|\Omega)] \\ &\propto -\sum_{k=1}^{K} \{\frac{n_k}{2}\log\sigma^2 - \frac{1}{2}\log|H_k| - \frac{1}{2\sigma^2}(\mathbf{y}_k - X_k\beta)' H_k^{-1} (\mathbf{y}_k - X_k\beta)\} \end{aligned} \tag{1.2.10}$$

Maximizing likelihood (3.2.3) is equivalent to maximize (1.2.10). Here, we take an iterated estimation procedure to estimate the parameters contained in Ω. For given

values of $h_m^2, h_f^2, h_{mf}^2, h_g^2$, we can get the maximum likelihood estimates (MLE) of parameters (β, σ^2) by setting the partial derivative of the log-likelihood function (1.2.10) to zero, i.e.,

$$\hat{\beta} = \sum_{k=1}^{K} (X_k^T H_k^{-1} X_k)^{-1} (X_k^T H_k^{-1} \mathbf{y}_k)$$

$$\hat{\sigma}^2 = \frac{1}{\sum_{k=1}^{K} n_k} \sum_{k=1}^{K} (\mathbf{y}_k - X_k\hat{\beta})^T H_k^{-1} (\mathbf{y}_k - X_k\hat{\beta})$$

It can be seen that $\hat{\beta}$ and $\hat{\sigma}^2$ are functions of h_m^2, h_f^2, h_{mf}^2 and h_g^2. Plug the updated parameter values for β and σ^2 into likelihood equation (1.2.10), the log-likelihood function can be simplified as,

$$\ell(\Omega) = \sum_{k=1}^{K} \log[f(\mathbf{y}_k|\Omega)] \propto -\sum_{k=1}^{K} \frac{n_k}{2} \log\hat{\sigma}^2 - \frac{1}{2} \sum_{k=1}^{K} \log|H_k| \qquad (1.2.11)$$

The simplex algorithm can be applied to maximize the function (1.2.11) with respect to parameters h_m^2, h_f^2, h_{mf}^2 and h_g^2 subject to the the constraints that $0 \leq h_m^2, h_f^2, h_{mf}^2, h_g^2 \leq 1$ and $0 \leq h^2 \leq 1$.

To guarantee a positive definite covariance matrix when searching for these heritability values over the constraint parameter space, a reparameterization technique is adopted (Xu and Atchley 1995). Taking $\delta_{ij} = h^2(\pi_{i_m j_m}\gamma_m^2 + \pi_{i_f j_f}\gamma_f^2 + \pi_{i_m/j_f}\gamma_{mf}^2 + \phi_{ij}\gamma_g^2)$ where $\gamma_m^2 = \frac{h_m^2}{h^2}, \gamma_f^2 = \frac{h_f^2}{h^2}, \gamma_{mf}^2 = \frac{h_{mf}^2}{h^2}$, $\gamma_g^2 = \frac{h_g^2}{h^2}$, and $h^2 = h_m^2 + h_f^2 + h_{mf}^2 + h_g^2$. We now have four new unknowns with the constraints: $0 \leq h^2 \leq 1$,

$\gamma_m^2 + \gamma_f^2 + \gamma_{mf}^2 + \gamma_g^2 = 1$ and $\gamma_m^2, \gamma_f^2, \gamma_{mf}^2, \gamma_g^2 \geq 0$.

The new constraints can be easily satisfied by a reparameterization technique. Let u, v_m, v_f, v_{mf} and v_g be any real numbers. Estimating h^2, γ_m^2, γ_f^2, γ_{mf}^2 and γ_g^2 can be done by maximizing the likelihood function (1.2.11) via searching through the real domain space with respect to u, v_m, v_f, v_{mf} and v_g with the reparameterization

$$h^2 = \frac{e^u}{1+e^u},$$

$$\gamma_m^2 = \frac{e^{v_m}}{e^{v_m} + e^{v_f} + e^{v_{mf}} + e^{v_g}},$$

$$\gamma_f^2 = \frac{e^{v_f}}{e^{v_m} + e^{v_f} + e^{v_{mf}} + e^{v_g}},$$

$$\gamma_{mf}^2 = \frac{e^{v_{mf}}}{e^{v_m} + e^{v_f} + e^{v_{mf}} + e^{v_g}},$$

and

$$\gamma_g^2 = \frac{e^{v_g}}{e^{v_m} + e^{v_f} + e^{v_{mf}} + e^{v_g}}$$

MLEs of h^2, γ_m^2, γ_f^2, γ_{mf}^2 and γ_g^2 can be obtained through the estimated values for u, v_m, v_f, v_{mf} and v_g according to the invariance property of MLEs. These estimated MLEs are used to update h^2, h_m^2, h_f^2, h_{mf}^2 and h_g^2, and hence σ^2 and β. The iteration steps continue until convergence.

1.2.4.2 The REML Estimation

The REML method was first proposed by Patterson and Thompson (1971). This method has been broadly applied to estimate variance components in a mixed-effect model framework. Taking $\Omega = (\beta, \Theta)$ where $\Theta=(\sigma_m^2, \sigma_f^2, \sigma_{mf}^2, \sigma_g^2, \sigma_e^2)$. The REML method starts with maximizing the following likelihood function,

$$\ell^*(\Theta) = \sum_{k=1}^{K} \log[f(\mathbf{y}_k|\Theta)] = -\frac{1}{2}\sum_{k=1}^{K}\left\{\log|\mathbf{\Sigma}_k| + \log(|X_k'\mathbf{\Sigma}_k^{-1}X_k|) + \mathbf{y}_k'P_k\mathbf{y}_k\right\} \tag{1.2.12}$$

where $P_k = \mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_k^{-1}X_k(X_k'\mathbf{\Sigma}_k^{-1}X_k)^{-1}X_k'\mathbf{\Sigma}_k^{-1}$. We can combine all family data together as one $N \times 1$ vector denoted as $\mathbf{y}$ where $N = \sum_{k=1}^{K} n_k$. All the X_k and the variance-covariance matrix $\mathbf{\Sigma}_k$ corresponding to each family can be combined. The log-likelihood function for the combined data is expressed as,

$$\ell^*(\Theta) = \log[f(\mathbf{y}|\Theta)] = -\frac{1}{2}\left\{\log|\mathbf{\Sigma}| + \log(|X'\mathbf{\Sigma}^{-1}X|) + \mathbf{y}'P\mathbf{y}\right\} \tag{1.2.13}$$

where $\mathbf{\Sigma}$ is a block diagonal matrix with the kth diagonal block $\mathbf{\Sigma}_k$ corresponding to the kth family and off-diagonal blocks being zeros; P is also a block diagonal matrix with block elements given by P_k. The dimension of $\mathbf{\Sigma}$ is $N \times N$. With this combination, we develop the following REML estimation procedure.

We apply the Fisher scoring algorithm to estimate the unknowns, which has the form,

$$\Theta^{(t+1)} = \Theta^{(t)} + \mathcal{I}(\Theta^{(t)})^{-1}\frac{\partial \ell^*(\Theta)}{\partial \Theta}|\Theta^{(t)}$$

where $\mathcal{I}(\Theta^{(t)})$ is the Fisher information matrix evaluated at $\Theta^{(t)}$ which can be expressed as,

$$\mathcal{I}\begin{pmatrix} \sigma^2_m \\ \sigma^2_f \\ \sigma^2_{mf} \\ \sigma^2_g \\ \sigma^2_e \end{pmatrix} = \frac{1}{2}$$

$$\begin{pmatrix} tr(P\Pi_m P\Pi_m) & tr(P\Pi_m P\Pi_f) & tr(P\Pi_m P\Pi_{m/f}), \\ tr(P\Pi_f P\Pi_m) & tr(P\Pi_f P\Pi_f) & tr(P\Pi_f P\Pi_{m/f}), \\ tr(P\Pi_{m/f} P\Pi_m) & tr(P\Pi_{m/f} P\Pi_f) & tr(P\Pi_{m/f} P\Pi_{m/f}), \\ tr(P\Phi_g P\Pi_m) & tr(P\Phi_g P\Pi_f) & tr(P\Phi_g P\Pi_{m/f}), \\ tr(P\Pi_m P) & tr(P\Pi_f P) & tr(P\Pi_{m/f} P), \end{pmatrix}$$

$$\begin{pmatrix} tr(P\Pi_m P\Phi_g) & tr(P\Pi_m P) \\ tr(P\Pi_f P\Phi_g) & tr(P\Pi_f P) \\ tr(P\Pi_{m/f} P\Phi_g) & tr(P\Pi_{m/f} P) \\ tr(P\Phi_g P\Phi_g) & tr(P\Phi_g P) \\ tr(P\Phi_g P) & tr(PP) \end{pmatrix}$$

The first-derivative of the log-likelihood function ℓ^* with respective to each vari-

ance components is given by,

$$\frac{\partial \ell^*}{\partial \sigma_m^2} = -\frac{1}{2}(tr(P\Pi_m) - \mathbf{y}^T P\Pi_m P\mathbf{y}),$$

$$\frac{\partial \ell^*}{\partial \sigma_f^2} = -\frac{1}{2}(tr(P\Pi_f) - \mathbf{y}^T P\Pi_f P\mathbf{y}),$$

$$\frac{\partial \ell^*}{\partial \sigma_{mf}^2} = -\frac{1}{2}(tr(P\Pi_{m/f}) - \mathbf{y}^T P\Pi_{m/f} P\mathbf{y}),$$

$$\frac{\partial \ell^*}{\partial \sigma_g^2} = -\frac{1}{2}(tr(P\Phi_g) - \mathbf{y}^T P\Phi_g P\mathbf{y}),$$

$$\frac{\partial \ell^*}{\partial \sigma_e^2} = -\frac{1}{2}(tr(PI_N) - \mathbf{y}^T PP\mathbf{y})$$

The REML estimator of β is the generalized least squares estimator, i.e.,

$$\hat{\beta} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} Y$$

1.2.5 QTL IBD sharing and genomewide linkage scan

The above IBD computation procedure assumes that a putative QTL is located right on a marker. When a QTL is located within an interval, a more efficient approach would be to do an interval scan and to test the imprinting property of QTLs at positions across the entire linkage group. Under the proposed framework, essentially we need to estimate the proportion of putative QTL alleles shared IBD at every genome position. Here we propose a method to calculate QTL alleles shared IBD inside an

interval conditional on the flanking markers. The so called expected conditional IBD values can be derived at each test position as a function of recombination fraction between the two flanking markers, and the one between one flanking marker and the QTL. We use one backcross initiated with the cross $QQ \times Qq$ as an example to illustrate the idea. For a putative QTL with two alleles Q and q, four QTL genotype pairs $QQ - QQ$, $QQ - Qq$, $Qq - QQ$ and $Qq - Qq$ can be formed. If the QTL genotype is observed, the corresponding QTL alleles shared IBD can be calculated (see Table 2.1). In general, the QTL genotype is unobservable, but its conditional distribution can be calculated from the two flanking markers. For individuals i and j with flanking marker genotypes g_i and g_j, let $\pi_{v|G_iG_j}$ be the IBD values calculated at the QTL position between individual i carrying QTL genotype G_i (=1 or 2 corresponding to QQ or Qq, respectively) and individual j carrying genotype G_j (similarly 1 or 2), where $v = i_m j_m, i_f j_f$ or i_m/j_f. For example, $\pi_{i_m j_m|G_iG_j}$ is the proportion of IBD sharing between individual i carrying QTL genotype G_i and individual j carrying genotype G_j for alleles derived from the mother.

Let $\varphi_{G_i|g_i}$ and $\varphi_{G_j|g_j}$ be the conditional distribution of QTL genotype G_i and G_j for individuals i and j given on the flanking markers g_i and g_j, respectively. This conditional probabilities can be easily calculated and can be found at standard QTL mapping literature (see Wu et al. 2007). The probability to observe $\pi_{v|G_iG_j}$ is just $\varphi_{G_i|g_i}\varphi_{G_j|g_j}$. Thus, the expected IBD values between individual i and j at the tested QTL position conditioning on the flanking markers g_i and g_j can be calculated as, $\hat{\pi}_v = \mathrm{E}(\pi_{v|G_iG_j}) = \sum_{G_i=1}^{2} \sum_{G_j=1}^{2} \pi_{v|G_iG_j}\varphi_{G_i|g_i}\varphi_{G_j|g_j}$. For the above

example, the IBD values derived from the maternal and paternal parents can be calculated as $\hat{\pi}_{imjm}=E(\pi_{imjm}|G_iG_j)=0.5\varphi_{1|g_i}\varphi_{1|g_j}+ 0.5\varphi_{1|g_i}\varphi_{2|g_j} + 0.5\varphi_{2|g_i}\varphi_{1|g_j} + 0.5\varphi_{2|g_i}\varphi_{2|g_j}$ and $\hat{\pi}_{ifjf}=E(\pi_{ifjf}|G_iG_j)=0.5\varphi_{1|g_i}\varphi_{1|g_j}+ 0.5\varphi_{2|g_i}\varphi_{2|g_j}$. Similarly, we can calculate the conditional expectation of IBD sharing for other backcross families.

Since $\varphi_{G_i|g_i}$ and $\varphi_{G_j|g_j}$ are functions of recombinations, the conditional QTL IBD values vary at different testing positions. Once the estimated IBD matrix is calculated at every 1 or 2cM on an interval bracketed by two markers throughout the entire genome, a grid search can be done at all testing positions. The amount of support for a QTL at a particular map position can be displayed graphically through the use of likelihood ratio profiles, which plot the likelihood ratio test statistic as a function of testing positions of putative QTLs (see details in hypothesis testing section). The peaks of the profile plot that passes certain significant threshold corresponds to the positions of significant QTLs.

1.2.6 Hypothesis testing

With the estimated parameters using either the ML or REML method, we are interested in testing the existence of QTLs across the genome and assess their imprinting mechanism. The first hypothesis is to test the existence of major QTLs, termed overall QTL test, which can be formulated as,

$$\begin{cases} H_0 : \sigma_m^2 = \sigma_f^2 = \sigma_{mf}^2 = 0 \\ \\ H_1 : \text{at least one parameter is not zero.} \end{cases} \tag{1.2.14}$$

Likelihood ratio (LR) test is applied which is computed between the full (there is a QTL) and the reduced model (there is no QTL) corresponding to H_1 and H_0, respectively. Let $\widetilde{\Omega}$ and $\widehat{\Omega}$ be the estimates of the unknown parameters under H_0 and H_1, respectively. The log-likelihood ratio can be calculated as,

$$\mathrm{LR}_1 = -2[\log L(\widetilde{\Omega}|\mathbf{y}) - \log L(\widehat{\Omega}|\mathbf{y})]$$

When testing the hypothesis, the polygene and the residual variances are nuisance parameters which are constrained to be nonnegative. The three tested genetic variance components under the null are lied on the boundaries of their alternative parameter spaces. Following Self and Liang (1987), when the null is true, LR_1 asymptotically follows a mixture of χ^2 distribution on $0, \cdots, 3$ degrees of freedom (df) with the mixture proportion for the χ_k^2 components being given in Theorem 2.2.1 in Chapter 2. The theoretical distribution can be used to assess significance in linkage scan. However, since there are many point tests across the genome, the point-wise significance value may not guarantee an appropriate genomewide error rate. Another approach to assess significance is to use nonparametric permutation tests in which the critical threshold value can be empirically calculated on the basis of repeatedly shuffling the relationships between marker genotypes and phenotypes (Churchill and Doerge 1994).

In simulation studies, we also simulate the null distribution and compare it with the theoretical distribution.

For those detected QTLs, the next step is to assess their imprinting property. An identified QTL can be imprinted, completely imprinted, partially imprinted or not imprinted at all. These can be tested through the following sequential tests. The first imprinting test is to assess whether a QTL shows imprinting effect, which can be done by formulating the following hypotheses,

$$\begin{cases} H_0 : \sigma_m^2 = \sigma_f^2 = \sigma^2 \\ H_1 : \sigma_m^2 \neq \sigma_f^2 \end{cases} \tag{1.2.15}$$

Rejection of H_0 provides evidence of genomic imprinting and the QTL is called iQTL. Again likelihood ratio test can be applied in which the log-likelihood ratio test statistics asymptotically follows a χ^2 with one df (Hanson et al. 2001). We denote the log-likelihood ratio test statistic as $\mathrm{LR}_{\mathrm{imp}}$. If the null is rejected, one would be interested to test if the detected iQTL is completely maternally or paternally imprinted. The corresponding hypotheses can be formulated as,

$$\begin{cases} H_0 : \sigma_m^2 = 0, \\ H_1 : \sigma_m^2 \neq 0. \end{cases}$$

for testing completely maternal imprinting and

$$\begin{cases} H_0 : \sigma_f^2 = 0, \\ H_1 : \sigma_f^2 \neq 0. \end{cases}$$

for testing completely paternal imprinting. The likelihood ratio test statistics for the above two tests asymptotically follow a 50:50 mixture of χ_0^2 and χ_1^2 distribution (Self and Liang 1987). Rejection of complete imprinting indicates partial imprinting.

1.2.7 Multiple QTL model

In reality, more than one QTL may contribute to the phenotypic variation located in one chromosome region or across the whole genome. The polygenic effect in model (1.2.3) absorbs the effects of multiple QTLs located on other chromosomes. However, when there are multiple QTLs located on the same linkage group as the tested QTL, if their effects are not properly adjusted, the estimation could be biased due to interference caused by theses QTLs outside of the testing interval (Zeng 1994; Martinez and Cuirnow 1992; Janson 1994; Zeng 1993). A multiple QTL model that can test the putative QTL effect while adjusting the effects of interference QTLs deserves more attention.

Zeng (1993) previously showed that IBD variables share the same property as the indicator variables in which the shared proportion of alleles IBD for a QTL conditional on the IBD of one flanking marker is independent of that of a QTL on the other side of that flanking marker. Thus, conditional on one flanking marker, the interference of QTLs located on the other side of the marker can be eliminated. By conditional

on the IBD of the flanking markers, the IBD sharing of a QTL is uncorrelated with that outside this interval. Xu and Atchley (1995) showed that one marker is enough to block the interference caused by other QTLs located on the same linkage group. The authors derived the next-to-flanking markers structure to block additional QTL effects from both sides of testing region in one chromosome. We derive a multiple QTL model adopting a similar idea as Xu and Atchley (1995). Assume there are total S QTLs located on a linkage group. Considering parent-specific allelic effects, the multiple QTL model can be expressed in general as,

$$y_{ik} = \mu_k + \sum_{s=1}^{S} a_{ikms} + \sum_{s=1}^{S} a_{ikfs} + G_{ik} + e_{ik}, \quad k = 1, \cdots, K; \; i = 1, \cdots, n_k \quad (1.2.16)$$

In an interval-based linkage scan, only one putative QTL is considered at each testing position conditioning on the effects of all other QTLs. Assuming there are total L and R QTLs located on the left and right side of the putative QTL on a linkage group, model (2.2.6) can be modified as,

$$y_{ik} = \mu_k + \sum_{l=1}^{L} a_{ikl} + (a_{ikm} + a_{ikf}) + \sum_{r=1}^{R} a_{ikr} + G_{ik} + e_{ik}, \quad k = 1, \cdots, K; \; i = 1, \cdots, n_k \quad (1.2.17)$$

where a_{ikl} and a_{ikr} are the lth and rth QTL random effects on the left and right side of the putative QTL, respectively. When testing the putative QTL effect, we are only interested in blocking the total effects of QTLs outside of the tested interval. Therefore, in the modified model, the effects of QTLs outside of the tested interval are not partitioned. This however does not affect the inference of the tested QTL.

As shown by Zeng (1993) and Jansen (1994; 1993), one marker is enough to block the correlation between a locus on its left and a locus on its right. Therefore, only two additional markers flanking the current interval are needed to block interference caused by outside QTLs (Xu and Atchley 1995). Let $\mathcal{M}_l$ and $\mathcal{M}_r$ denote two flanking markers for the tested interval, and $\mathcal{L}$ and $\mathcal{R}$ denote the two markers next to $\mathcal{M}_l$ and $\mathcal{M}_{l+1}$ with the marker order $\mathcal{L}$-$\mathcal{M}_l$-$\mathcal{M}_{l+1}$-$\mathcal{R}$. With the modified model given in (1.2.17), the covariance of phenotypes between individuals i and j in the kth backcross family can be expressed as,

$$\begin{aligned} Cov(y_{ik}, y_{jk}) &= \sum_{l=1}^{L} Cov(a_{ikl}, a_{jkl}) + Cov(a_{ikm}, a_{jkm}) + Cov(a_{ikf}, a_{jkf}) + \\ &\quad + Cov(a_{ikm}, a_{jkf}) + \sum_{r=1}^{R} Cov(a_{ikr}, a_{jkr}) + \phi_{ij}\sigma_g^2 + I_{ij}\sigma_e^2 \\ &= \sum_{l=1}^{L} \pi_{l|k}\sigma_l^2 + \pi_{i_m j_m}\sigma_m^2 + \pi_{i_m/j_f}\sigma_{mf}^2 + \pi_{i_f j_f}\sigma_f^2 \\ &\quad + \sum_{r=1}^{R} \pi_{r|k}\sigma_r^2 + \phi_{ij}\sigma_g^2 + I_{ij}\sigma_e^2 \end{aligned} \tag{1.2.18}$$

where $\pi_{l|k}$ and $\pi_{r|k}$ are the IBD values for QTLs located on the left and right side of the putative QTL in the kth backcross family, and can be calculated following (1.2.5) and (1.2.6) if their genotype information is known. Unfortunately, the number and exact locations of QTLs outside of the testing interval are unknown. Hence $\pi_{l|k}$ and $\pi_{r|k}$ are not observable. Xu and Atchley (1995) showed that when $\pi_{l|k}$ and $\pi_{r|k}$ are unknown, they can be estimated by some composite terms $K(\theta_{l\mathcal{L}}, \pi_{\mathcal{L}|k})$ and $K(\theta_{l\mathcal{R}}, \pi_{\mathcal{R}|k})$, where $K(\theta_{l\mathcal{L}}, \pi_{\mathcal{L}|k})$ is a function of the recombination fraction

between the lth QTL and the left marker $\mathcal{L}$ as well as a function of $\pi_{\mathcal{L}|k}$, the IBD value for a pair of individuals at the left marker $\mathcal{L}$. $K(\theta_{l\mathcal{R}}, \pi_{\mathcal{R}|k})$ can be similarly defined. Following Xu and Atchley (1995), $K(\theta_{l\mathcal{L}}, \pi_{\mathcal{L}|k})$ can be expressed as a function of $\pi_{\mathcal{L}|k}$ multiplied by a function of recombination frequency between the lth QTL and the marker $\mathcal{L}$, $f(\theta_{l\mathcal{L}})$, i.e., $K(\theta_{l\mathcal{L}}, \pi_{\mathcal{L}|k}) = \pi_{\mathcal{L}|k} f(\theta_{l\mathcal{L}})$. Similarly, $K(\theta_{l\mathcal{R}}, \pi_{\mathcal{R}|k}) = \pi_{\mathcal{R}|k} f(\theta_{r\mathcal{R}})$. When doing an interval scan, the covariance function given in (4.2.6) between individuals i and j can be re-expressed as,

$$
\begin{aligned}
& Cov(y_{ik}, y_{jk} | \pi_{\mathcal{L}|k}, \hat{\pi}_{i_m j_m}, \hat{\pi}_{i_m/j_f}, \hat{\pi}_{i_f j_f}, \pi_{\mathcal{R}|k}) \\
=& \sum_{l=1}^{L} K(\theta_{l\mathcal{L}}, \pi_{\mathcal{L}|k}) \sigma_l^2 + \hat{\pi}_{i_m j_m} \sigma_m^2 + \hat{\pi}_{i_m/j_f|k} \sigma_{mf}^2 + \hat{\pi}_{i_f j_f} \sigma_f^2 \\
& + \sum_{r=1}^{R} K(\theta_{l\mathcal{R}}, \pi_{\mathcal{R}|k}) \sigma_r^2 + \phi_{ij} \sigma_g^2 + I_{ij} \sigma_e^2 \\
=& \pi_{\mathcal{L}|k} \sum_{l=1}^{L} f(\theta_{l\mathcal{L}}) \sigma_l^2 + \hat{\pi}_{i_m j_m} \sigma_m^2 + \hat{\pi}_{i_m/j_f} \sigma_{mf}^2 + \hat{\pi}_{i_f j_f} \sigma_f^2 \\
& + \pi_{\mathcal{R}|k} \sum_{r=1}^{R} f(\theta_{r\mathcal{R}}) \sigma_r^2 + \phi_{ij} \sigma_g^2 + I_{ij} \sigma_e^2 \\
=& \pi_{\mathcal{L}|k} \sigma_L^2 + \hat{\pi}_{i_m j_m} \sigma_m^2 + \hat{\pi}_{i_m/j_f|k} \sigma_{mf}^2 + \hat{\pi}_{i_f i_f} \sigma_f^2 + \pi_{\mathcal{R}|k} \sigma_R^2 + \phi_{ij} \sigma_g^2 + I_{ij} \sigma_e^2
\end{aligned}
$$

Instead of estimating individual variance components σ_l^2 and σ_r^2, now we estimate the composite term $\sum_{l=1}^{L} f(\theta_{l\mathcal{L}}) \sigma_l^2 = \sigma_L^2$ and $\sum_{r=1}^{R} f(\theta_{r\mathcal{R}}) \sigma_r^2 = \sigma_R^2$. By conditioning the IBD sharing information for the left and right markers $\mathcal{L}$ and $\mathcal{R}$, the effects of those interference QTLs are blocked. σ_L^2 and σ_R^2 absorb the random effects of all QTLs that are outside of the testing interval but are on the same linkage group as

the putative QTL. Estimation of the variance components terms follows the same procedure as the single QTL analysis with slight modification to consider multiple variance components.

1.3 Results

1.3.1 Simulation design

To investigate the performance of the proposed models and estimation methods, we conduct intensive computer simulations. We start with the single QTL simulation followed by the multiple QTL analysis. Six evenly spaced markers ($\mathcal{M}_1 - \mathcal{M}_6$) are simulated. The total length for the simulated linkage group is 100cM. We assume that all the backcross families share the same linkage map constructed using Haldane map function. For simplicity, we assume the sample size for all backcross families is the same (i.e., $n_k = n$). The position of the simulated QTL is assumed to be located at 48cM away from the first marker ($\mathcal{M}_1$). The effect of the putative QTL is simulated by assuming different imprinting mechanisms, i,e., no imprinting, completely imprinting and partial imprinting. Once QTL genotypes are simulated, phenotypes can be simulated by randomly drawing multivariate normal distribution with the covariance structure given in (2.2.4) with different parameter combinations.

To evaluate the effect of family and offspring size combination on testing power and parameter estimation, we simulate data assuming different sample size combinations. We fix the total sample size as 400 and vary the family and offspring size with

different combinations, i.e., 4×100, 8×50, 20×20 and 100×4. The first number for each combination indicates the family size. For example, in the combination 20×20, 20 families each containing 20 offspring are simulated. For each sib-pair, the IBD value at a putative position at every 2cM along the linkage group is calculated as described in the previous section. For each simulation scenario, 100 simulation replications are recorded and the ML and the REML methods are used to estimate the unknown parameters.

Table 1.2: The power, MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL with no imprinting effect under different sampling designs.

$n_F \times n_k$	Estimation method	Position 48cM	$\mu 1$ 10	$\mu 2$ 8	$\mu 3$ 6	σ_m^2 1.5	σ_f^2 1.5	σ_{mf}^2 0.5	σ_g^2 0.5	σ_e^2 2	Power[1]	Power[2]	Type I error
4×100	ML	48.78	9.85	8.08	5.80	0.91	0.73	0.78	0.17	2.12	0.91	0.99	0.42
		(17.87)	(1.869)	(1.246)	(1.768)	(1.296)	(1.122)	(1.242)	(0.572)	(0.229)			
	REML	47.2	9.83	8.13	6.23	1.51	1.42	1.52	1.06	2.02	0.92	0.96	0.17
		(11.95)	(2.45)	(1.39)	(2.42)	(1.65)	(1.33)	(2.88)	(2.35)	(0.57)			
8×50	ML	50.04	9.68	7.97	5.87	1.09	1.28	0.98	0.37	2.06	0.81	1.00	0.26
		(14.73)	(1.373)	(0.736)	(1.157)	(0.943)	(0.990)	(1.302)	(0.701)	(0.233)			
	REML	46.9	9.84	8.12	5.91	1.61	1.64	1.02	0.70	2.03	0.98	1.00	0.08
		(7.93)	(1.16)	(0.83)	(1.10)	(0.99)	(1.26)	(1.44)	(1.10)	(0.37)			
20×20	ML	48.96	9.99	7.95	6.05	1.34	1.25	0.60	0.32	2.01	0.98	0.98	0.12
		(10.70)	(0.706)	(0.495)	(0.633)	(0.768)	(0.771)	(0.810)	(0.478)	(0.182)			
	REML	49.74	10.04	7.94	5.89	1.55	1.48	0.69	0.58	1.98	0.95	0.98	0.06
		(13.96)	(0.597)	(0.458)	(0.675)	(0.964)	(0.786)	(0.838)	(0.608)	(0.257)			
100×4	ML	51.34	10.02	7.95	6.02	1.50	1.45	0.58	0.41	2.01	0.55	0.86	0.11
		(18.00)	(0.321)	(0.277)	(0.345)	(0.875)	(0.878)	(0.553)	(0.444)	(0.242)			
	REML	48.4	9.95	8.01	6.01	1.51	1.57	0.62	0.52	1.99	0.67	0.81	0.07
		(20.84)	(0.305)	(0.220)	(0.319)	(0.902)	(0.892)	(0.654)	(0.485)	(0.222)			

Power[1] is calculated using the empirical distribution through simulation. Power[2] is calculated using the theoretical distribution assuming mixture chi-square distribution. Type I error refers to the imprinting type I error.

1.3.2 Simulation results

1.3.2.1 Single QTL analysis

The single QTL model assumes one QTL is located at the third interval in the simulated linkage group, 48cM away from the first marker. Results using both ML and REML estimation methods are summarized in Table 2.2. n_F denotes the number of families and n_k denotes the number of offspring for each family. Without loss of generality, we assume equal offspring size for all families in each simulation scenario. The simulated parameter values are listed under each parameter. The root mean square errors (RMSEs) are recorded for each parameter estimate to assess the estimation precision. Overall, the fixed effects (three means) and most variance components can be better estimated with large number of families. For example, the RMSE of parameter μ_1 is reduced from 1.869 (2.45) to 0.321 (0.305) when the number of families increases from 4 to 100 with the ML (REML) estimation method. The only exception is the two variance components terms (σ^2_m and σ^2_f) which are better estimated with the 20×20 combination design. Through the combination of different line crosses, the parameter inference space is expanded, and as a result, better estimations are achieved as expected. However, the QTL position is better estimated with the 8×50 and 20×20 designs than the other two among the four simulation scenarios. The 100×4 design gives the worst QTL position estimation with the largest RMSEs for both estimation methods. Therefore, a balance of family and offspring size is needed. A moderate family size with moderate offspring size would be necessary in order to achieve reasonable parameter estimation for both QTL effects and position.

Table 2.2 also lists the results of power analysis under different scenarios with two different estimation methods. Power^1 denotes the empirical power calculated from the simulated null distribution corresponding to hypothesis test (2.2.6). We simulate the null distribution by simulating data assuming no QTL effect (i.e., σ^2_m=σ^2_f=σ^2_{mf}=0). The LR test statistics is calculated for each simulation run and the 95% cutoff is reported as the threshold value. Power^2 refers to the theoretical power which is calculated assuming the mixture chi-square distribution. Results show that the threshold calculated from the theoretical distribution is smaller than the one calculated from the simulation. Thus the testing power based on the theoretical cutoff is greater than the empirical power. The testing powers under different sampling designs are very comparable except for the 100×4 design in which the power is dramatically reduced compared to other designs. No remarkable difference in power for both estimation methods is observed.

Fig. 1.2 shows the log-likelihood ratio test statistic calculated under the four sampling designs across the simulated linkage group by using both ML and REML estimation methods. The plotted LR curve is from averaged LR values out of 100 replications. It is clear that large offspring size always gives large test statistics. As the family size increases from 4 to 100 and so decreased offspring size, we observe a huge LR value decrease. Clearly, the 100×4 design is less powerful than the others. The last column listed in Table 2.2 shows the type I error for testing genomic imprinting, i.e., H_0: $\sigma^2_m = \sigma^2_f$. The simulated data assume no imprinting ($\sigma^2_m = \sigma^2_f$=1.5). The imprinting test is only conducted at the position where the overall QTL test

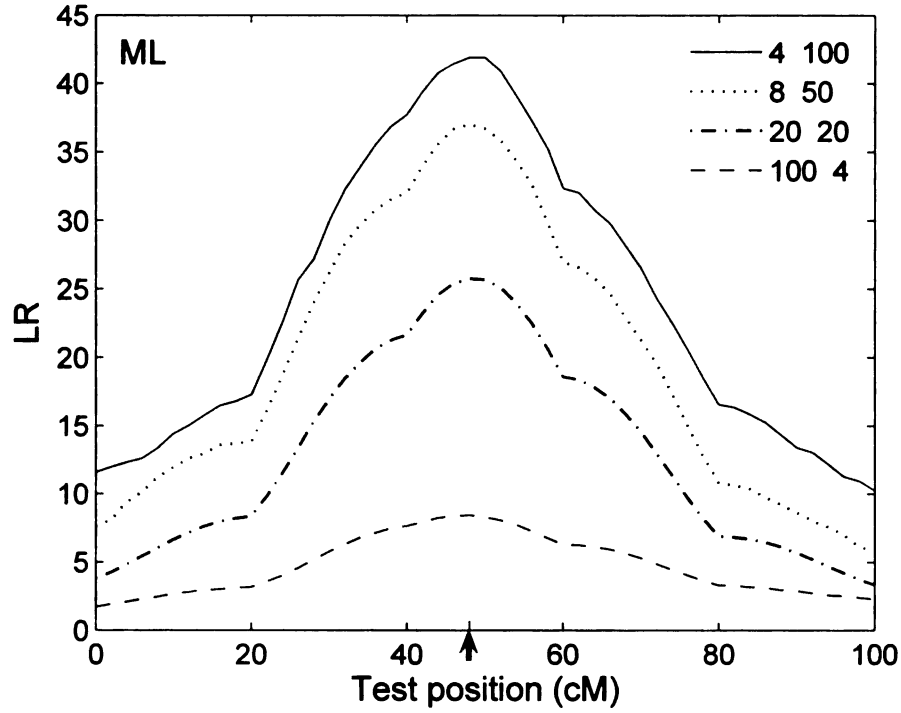

Figure 1.2: The LR profile plot. The left and right figures correspond to the LR profiles generated using the ML and REML method, respectively. The arrow indicates the true QTL position.

shows significance. The imprinting test statistic LR_{imp} is compared with a chi-square distribution with 1 df. Overall, the REML estimation method results in smaller type I error rate than the ML method does. As the number of families increase, the type I error decreases. The 4×100 design yields the largest type I error.

In comparison of the ML and REML methods, the REML method gives smaller estimation biases but larger RMSEs than the ML method does. This reflects the large variability of the REML estimation. In terms of computation speed, the ML method is faster than the REML method. Even though the QTL position estimation

is better estimated by using the REML method when family size is small, as family size increases, the REML method performs worse than the ML method (Table 2.2). In checking the LR profile plot in Fig. 1.2 and the power analysis in Table 2.2, we do not observe significant gain in power by using the REML method. The two methods do no dominate each other and are very comparable in power analysis. With large sample size and limited computing resources, one might want to try the ML method first. However, the REML method is suggested when testing imprinting since it has small type I error.

In a short summary of the results listed in Table 2.2, the 8×50 and 20×20 designs give better QTL position estimation and testing power. In terms of the type I error for imprinting test, the 20×20 and 100×4 designs provide reasonable type I error. Thus, a practical guidance is to choose the 20×20 design, and one should always avoid designs with extremely large or extremely small family size.

Table 1.3: The power, MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL showing different imprinting effects under the 20×20 sampling design. The square roots of the mean squared errors of the parameters are given in parentheses.

Estimation	Position	$\mu 1$	$\mu 2$	$\mu 3$	σ_m^2	σ_f^2	σ_{mf}^2	σ_g^2	σ_e^2			
method	48cM	10	8	6			0.5	0.5	2	Power[1]	Power[2]	ipower
					3	0						
ML	48.98	10.06	8.02	5.94	2.94	0.10	0.65	0.25	2.04	0.98	0.99	0.91
	(7.36)	(0.716)	(0.388)	(0.722)	(1.197)	(0.217)	(0.694)	(0.426)	(0.207)			
REML	48.54	9.96	7.95	5.90	2.95	0.09	0.64	0.57	1.97	0.97	0.99	0.86
	(9.62)	(0.705)	(0.428)	(0.747)	(1.501)	(0.180)	(0.598)	(0.609)	(0.226)			
					0	3						
ML	46.62	10.05	8.02	6.08	0.09	2.76	0.69	0.30	2.05	0.97	1.00	0.89
	(8.11)	(0.680)	(0.442)	(0.584)	(0.201)	(1.052)	(0.722)	(0.430)	(0.193)			
REML	48.7	10.04	7.98	5.91	0.09	2.95	0.61	0.57	1.98	0.96	0.98	0.88
	(7.40)	(0.560)	(0.506)	(0.615)	(0.195)	(1.279)	(0.663)	(0.544)	(0.225)			
					1	2						
ML	47.8	10.09	8.00	6.11	0.84	2.04	0.66	0.32	2.02	0.97	1.00	0.24
	(7.83)	(0.634)	(0.468)	(0.666)	(0.615)	(0.994)	(0.811)	(0.475)	(0.203)			
REML	49.66	10.02	7.96	5.90	1.08	1.73	0.68	0.61	2.00	0.97	0.99	0.09
	(11.06)	(0.606)	(0.548)	(0.668)	(0.679)	(0.958)	(0.707)	(0.616)	(0.216)			

Power[1] and power[2] correspond to the overall QTL effect test (2.2.6) calculated using the empirical and theoretical cutoff, respectively; ipower refers to the imprinting test power corresponding to test (2.2.8). See Table 2.2 for explanations of other parameters.

To evaluate the proposed model under different imprinting mechanisms, we simulated data assuming different degree of imprinting. Since the results in Table 2.2 indicate that a 20×20 design provides relatively reasonable parameter estimation, good power and small type I error rate for imprinting test, the evaluation of imprinting analysis is thus focused on this design. The results for 100 simulation replication are summarized in Table 2.3. Three imprinting models are assumed: complete maternal imprinting ($\sigma_m^2 = 0$ and σ_f^2=3), complete paternal imprinting ($\sigma_m^2 = 3$ and σ_f^2=0), and partial maternal imprinting ($\sigma_m^2 = 1$ and σ_f^2=2). Both ML and REML estimators are reported. Overall, the two estimation methods produce very comparable results with less biased estimations by the REML method as we expected. All the parameters can be properly estimated with reasonable precision.

Large imprinting power is observed when the variance difference between the two parent-specific variance components is large. When the difference between the two parent-specific variance components is reduced, the power to detect imprinting is largely reduced. For example, when data are simulated assuming complete paternal imprinting, the power is 0.91(0.86) by using the ML(REML) estimation method. With partially imprinted data, the imprinting power reduces to 0.24(0.09) by using the ML(REML) method, even though it can be increased by increasing the offspring sample size (data not shown).

Table 1.4: The power, MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for data simulated with Mendelian and imprinting models based on the 20×20 sampling design.

Simulation model	Analysis model	Position 48cM	$\mu 1$	$\mu 2$	$\mu 3$	h_a^2			h_g^2	σ_e^2		
						h_m^2	h_f^2	h_{mf}^2	0.083	2	Power[1]	Power[2]
M	M	48.08	10.123	7.890	6.023		0.544		0.093	1.982	1.00	1.00
		(2.981)	(0.887)	(0.730)	(1.046)		*(0.121)*		(0.137)	(0.250)		
	I	48.260	9.922	8.081	6.074	0.263	0.272	0.239	0.049	1.983	1.00	1.00
		(3.015)	(1.010)	(0.720)	(0.949)	*(0.097)*	*(0.110)*	*(0.124)*	(0.080)	(0.249)		
I_1	M	48.940	10.079	7.913	6.023		0.323		0.111	2.022	0.86	0.96
		(11.660)	(0.588)	(0.511)	(0.712)		*(0.287)*		(0.134)	(0.205)		
	I	47.86	10.077	7.900	6.018	0.008	0.473	0.089	0.091	2.015	0.95	0.98
		(7.524)	(0.579)	(0.575)	(0.665)	*(0.022)*	*(0.151)*	*(0.099)*	(0.053)	(0.186)		
I_2	M	48.640	10.070	7.917	6.028		0.331		0.103	2.023	0.91	0.98
		(8.389)	(0.613)	(0.517)	(0.706)		*(0.278)*		(0.129)	(0.195)		
	I	49.200	10.059	7.926	6.008	0.079	0.396	0.089	0.089	2.023	0.93	0.98
		(9.818)	(0.617)	(0.518)	(0.673)	*(0.075)*	*(0.156)*	*(0.097)*	(0.050)	(0.184)		

M and I. refers to Mendelian and imprinting model, respectively. Simulated parameters for model M: ($h_a^2 = 0.583$); I_1: (h_m^2, h_f^2, h_{mf}^2)=(0, 0.5, 0.083); and I_2: (h_m^2, h_f^2, h_{mf}^2)=(0.083, 0.417, 0.083). Power[1] and Power [2] correspond to the power calculated using the empirical cutoff and the theoretical threshold, respectively. The numbers given in the parenthesis with normal and italic fonts correspond to the RMSEs and standard errors of the parameter estimates, respectively. See Table 2.2 for other explanations.

In reality, whether a QTL is imprinted or not is an unknown prior. When a QTL has Mendelian effect and is not imprinted, is there any power loss by analyzing with the proposed imprinting model? Or when a QTL is actually imprinted, is there any power loss by analyzing with regular variance components approach? To answer these two questions, we simulated data under different scenarios and analyzed with both Mendelian and imprinting models. The first and second column in Table 1.4 refer to the simulation and analysis models, respectively. M refers to the Mendelian model without variance components partition and I refers to the imprinting model with allelic-specific partition of the variance components. For comparison purpose, heritabilities are recorded instead of original variance components estimates. The polygene and residual variances are fixed as 0.5 ($h_g^2 = 0.083$) and 2, respectively for all the simulation scenarios. We first simulated data with one additive genetic effect without partitioning variance into allelic specific components. This is equivalent to simulate data assuming the Mendelian model. A single additive variance component of 3.5 is assumed which corresponds to a heritability of $h_a^2 = 0.583$. The second scenario is to simulate data with three allelic-specific variance components. Simulation models I_1 and I_2 correspond to a complete maternal imprinting model (i.e., $h_m^2 = 0$ and $h_f^2 = 0.5$) and a partial maternal imprinting model (i.e., $h_m^2 = 0.083$ and $h_f^2 = 0.417$), respectively. The variance component σ_{mf}^2 is assumed to be 0.5 ($h_{mf}^2 = 0.083$) for I_1 and I_2. In all the simulations, we use the 20×20 design to make the comparison. Similar results are expected under the other sampling designs. Since the true variance components values for the imprinting model is unknown when data

are simulated assuming Mendelian effect and vice versa, only standard deviations for these parameter estimates are recorded (listed as italic font in the parentheses).

The simulation results are summarized in Table 1.4. When the simulated model is Mendelian, QTL position is better estimated with the Mendelian model than with the imprinting model. No remarkable difference in power is observed for both models. The estimated parent-specific variances due to maternal and paternal alleles are almost identical and no imprinting is detected. When data are simulated assuming imprinting (model I_1 and I_2), large power is observed when analyzed with the imprinting model. For example, the power is 86% when analyze the I_1 imprinting data by the Mendelian model. The power is increased to 95% when data are analyzed by the imprinting model. When imprinting data are analyzed with the Mendelian model, the major QTL variance is under-estimated and the polygene variance is slightly over-estimated. No remarkable differences are observed for the estimation of the three fixed mean effects and the residual variance under all simulation cases. In any case, the imprinting model performs better or no worse than the Mendelian model. Thus, it is generally safe to apply the imprinting analysis for data shown any inheritance pattern.

1.3.2.2 Multiple QTL analysis

Table 1.5: The MLEs and REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for data simulated with two QTLs under the 20×20 design. The square roots of the mean squared errors are given in parentheses.

	Q_1				Q_2			
Estimation method	Position 28cM	σ_m^2	σ_f^2	σ_{mf}^2 0.25	Position 68cM	σ_m^2	σ_f^2	σ_{mf}^2 0.25
		0.75	0.75			0.75	0.75	
ML	31.06	0.960	0.819	0.381	64.560	0.806	0.891	0.414
	(12.825)	(0.760)	(0.693)	(0.533)	(12.126)	(0.552)	(0.679)	(0.646)
REML	32.32	0.949	0.924	0.440	64.60	0.943	1.002	0.461
	(13.951)	(0.739)	(0.686)	(0.678)	(12.169)	(0.713)	(0.711)	(0.601)
		1.5	0			0.75	0.75	
ML	29.52	1.568	0.210	0.601	63.86	1.210	0.532	0.533
	(12.153)	(0.872)	(0.382)	(0.836)	(11.913)	(0.932)	(0.572)	(0.847)
REML	31.22	1.668	0.282	0.433	64.16	1.355	0.665	0.529
	(12.557)	(0.881)	(0.457)	(0.598)	(13.030)	(1.128)	(0.554)	(0.650)

Q_1 and Q_2 refer to two QTLs located at 28cM and 68cM.

To see the relative merit of multiple QTL analysis against single QTL analysis when multiple QTLs are located on the same linkage group, two QTLs are simulated with QTL 1 (denoted as Q_1) located at the second interval, $28cM$ away from the first marker ($\mathcal{M}_1$) and QTL 2 (denoted as Q_2) located at the fourth interval, 68cM away from the first marker. Two simulation scenarios are considered. The first scenario considers two non-imprinted QTLs with equal genetic effects. The second scenario assume Q_1 is imprinted and Q_2 is not imprinted. Simulated parameters for the two QTLs are listed in Table 1.5. Data are simulated assuming the 20×20 design. Parameters are estimated by the ML and REML approaches with 100 replicates.

Fig. 1.3 shows the LR profile plots for the single and multiple QTL analysis. The single QTL model indicates three major peaks. The highest peak for the single QTL analysis is located at the wrong QTL interval where no QTL is assumed. The so called "ghost image" of QTL can be removed and the positions of the two QTLs can be precisely mapped on the chromosome by the multiple QTL model. Two clear peaks indicating the correct QTL positions (arrow signs) are observed by the multiple QTL analysis. However, we observe a remarkable reduction in LR values by multiple QTL analysis compared to those by the single QTL analysis. Since the threshold for multiple QTL analysis is unknown, we can not make the conclusion that multiple QTL analysis is less powerful than the single QTL analysis. It is possible that we may gain accuracy in QTL position estimation at the cost of power loss. Similar phenomenon and issues were also observed and discussed in the literatures (Zeng 1994; Xu and Alchley 1995).

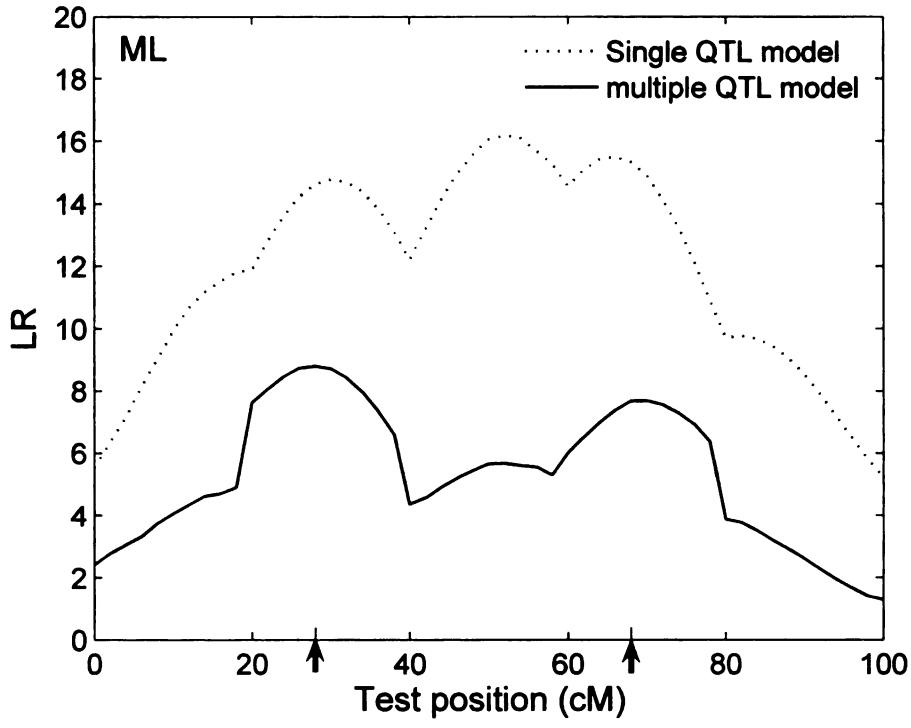

Figure 1.3: The LR profile plot for singe QTL and multiple QTL analysis. The true QTL positions are simulated at $28cM$ and $68cM$ (see the arrow sign). The dotted curve and the solid curve represent the LR profiles by single QTL and multiple QTL analysis, respectively. The left and right figures correspond to the LR profiles generated using the ML and REML method, respectively.

The results of the multiple QTL analysis are summarized in Table 1.5. The fixed mean effects, the polygene and residual variance components can be reasonably estimated with small RMSEs, similar results shown in Table 2.2 for the 20×20 design and hence are not reported here. Only the genetic factors for the two simulated QTLs are reported. It can be seen that both ML and REML methods provide reasonable parameter estimates and are very comparable. Under the first simulation scenario in which both QTLs are not imprinted, the genetic effects are all slightly over-estimated

by both methods. This might be due to the interference of the two QTLs in the same linkage group. The multiple QTL model may not completely block the effects of QTLs outside of the tested interval. For the second simulation scenario, an interesting pattern is observed. When one QTL is imprinted (Q_1), the maternal and paternal variance components for the second one (Q_2) tend to be estimated with bias in the direction as the first imprinted QTL, i.e., σ_m^2 tends to be over-estimated and σ_f^2 tends to be under-estimated. As we gain accuracy in QTL position estimation, we lose precision for the parameter estimation. These effects are expected as described in Zeng (1994) and Xu and Atchley (1995). More investigations are needed in multiple QTL analysis in order to maintain a good balance of QTL position and parameter inference.

1.4 Discussion

Statistical methods assuming fixed effect models for iQTL mapping in controlled outbred and inbred lines have been proposed (e.g., Koning et al. 2000; Cui 2007; Cui et al. 2006 2007). Considering the limitation of fixed-effect models, a random model that estimates the QTL variance by extending single line cross to multiple line crosses should be more powerful in QTL variance inference (Xie et al. 1998). The IBD-based variance components method assuming random genetic effect for iQTL mapping has been developed in human linkage analysis (Hanson et al. 2001). However, no study has been proposed to map iQTL using variance components method with inbred or partially inbred line cross. In this article, we have first time presented an IBD-based

variance components framework to search for the existence and distribution of iQTL throughout the entire genome in multiple experimental line crosses. The idea of the method is demonstrated through a backcross design. It can also be extended to multiple F_2 line crosses using the sex-specific recombination information as proposed by Cui et al. (2006).

The key point of the proposed iQTL variance components analysis is to partition the additive genetic variance into parent-specific components. We have proposed a new parent-specific allelic sharing method which characterizes the relatedness of parent-specific alleles between pairs of individuals in a backcross pedigree. The calculation of parent-specific allelic sharing is based on the information of the coefficient of coancestry. More complicated calculation of the coefficient of coancestry can be found at Harris (1964). The quantification of the coefficient of the coancestry proposed by Harris (1964) can also be utilized to calculate the parent-specific IBD sharing in an inbred human population, and thus for iQTL mapping in inbred human populations.

There have been extensive studies in literature about various methods in the estimation of variance components in a mixed-effect model framework. The ML and REML are two commonly applied methods in variance components estimation with less biased estimation by the REML method. Simulations show that the ML method yields high precision in parameter estimation but with relatively large bias than the REML method. Power analysis indicates that the ML method is a little more powerful than the REML method but with large type I error when testing imprinting. In terms of computing speed, the ML method is faster than the REML method. Thus, no

single method dominates the other. In terms of overall QTL test, we suggest to use the ML method for the genomewide linkage scan and use the REML method for the imprinting test.

The effect of sampling design is investigated by extensive simulations. Results indicate that one can always achieve large power with large offspring size when the total sample size is fixed. The LR value differences under different sampling designs are shown in Fig. 1.2. However, the combination of small families each with large offsprings gives poor parameter estimation and large type I error for imprinting test (Table 2.2). As the number of families increase, we observe less biased parameter estimates for both fixed and random effects, but with poor QTL position estimation and small power. This information implies that it is necessary to enlarge the number of families to improve precision of parameter estimation. Meanwhile, a balance of family and offspring size is needed to maintain good QTL detection power and position estimation. Our simulations indicate that for a fixed total sample size (n=400), both 8×50 and 20×20 designs yield comparable results and both designs outperform the other two designs (Table 2.2). Moreover, the 20×20 design produces relatively small type I error in imprinting test. With the 20×20 design, results in Table 1.4 indicate that the imprinting model is better or as good as the regular Mendelian analysis without considering imprinting. In real data analysis, it should be safe to apply the proposed imprinting model for data with any imprinting pattern.

In this study, we have extended the single marker-based analysis to an interval-based mapping for genomewide scan and testing of iQTL effects. Considering the

interference of QTLs located on the same linkage group, we have extended the single QTL model to multiple QTL analysis following the derivation of Xu and Atchley (1995). Simulation results indicate the relative merit of the multiple QTL analysis with improved QTL position inference, but with possible power loss (Fig. 1.3). This, however, has been a common issue in multiple QTL modelling (Zeng 1994; Xu and Atchley 1995). More investigations are needed in deriving efficient and robust multiple QTL mapping models to improve precision without suffering too much from power loss.

The theoretical distribution for the likelihood ratio test has been a challenging problem in QTL mapping. Dupuis and Siegmund (1999) first proposed theoretical properties for LR test statistics in a genomewide linkage scan for QTLs in an interval mapping frameworh with a fixed-effect model. Currently, most linkage analysis using the variance components method assume that the LR test statistic follows a mixture of chi-square distribution (Allison et al. 1999). The mixture distribution is derived following Self and Liang (1987). With multiple testings and multiple nuisance parameters in a genomewide scan, the assumptions to get the mixture chi-square distribution may not be satisfied. Moreover, the multivariate normal assumption for the phenotypic data required to get the mixture distribution may not even valid. No theoretical work has been done to investigate this in a IBD-based variance components linkage mapping. Our simulations indicate that the theoretical threshold calculated from the mixture chi-square distribution is smaller than the simulated cutoff. Thus, the power calculated with the theoretical threshold is slightly inflated. A modified

mixture chi-quare distribution may be more appropriate. More theoretical investigations are needed in this regard.

Chapter 2

A general statistical framework for dissecting parent-of-origin effects underlying endosperm traits in flowering plants

2.1 INTRODUCTION

The life cycle of an angiosperm starts with the process of double fertilization, where the fertilization of the haploid egg with one sperm cell forms the embryo, and the fusion of the two polar nuclei with another sperm cell develops into endosperm (Chaudhury et al. 2001). Thus, endosperm is a tissue unique to angiosperm. The embryo and endosperm are genetically identical, except that the endosperm is triploid com-

posed of one set of paternal and two identical sets of maternal chromosomes. In cereals, the endosperm of a grain is the major storage organ providing nutrition for early-stage seed development, and more than that, serves as the major source of food for human beings. The identification of important genes that underlie the variation of quantitative traits of various interests in endosperm, is thus paramountly important.

Genomic imprinting refers to the situation where the expression of the same genes is different depending on their parental origin (Pfifer 2000). It has been increasingly recognized that many endosperm traits are controlled by genomic imprinting. For example, endoreduplication is a commonly observed phenomenon which shows a maternally controlled parent-of-origin effect in maize endosperm (Dilkes et al. 2002). Cells undergo endoreduplication are typically larger than other cells, which consequently results in larger fruits or seeds beneficial to human beings (Grime and Mowforth 1982). Other reports of genomic imprinting with paternal imprinting in maize endosperm include, for instance, the r gene in the regulation of anthocyanin (Kermicle 1970), the seed storage protein regulatory gene *dsrl* (Chaudhuri and Messing 1994), the *MEA* gene affecting seed development (Kinoshita et al. 1999) and some *α-tubulin* genes (Lund et al 1995). These studies underscore the value of developing statistical methods that empower geneticists to identify the distribution and effects of imprinted genes controlling endosperm traits.

Statistical methods for mapping imprinted genes or imprinted quantitative trait loci (iQTL) have been extensively studied. Focusing on different genetic designs and different segregation populations, methods were developed in mapping iQTL

underlying quantitative traits in controlled experimental crosses (e.g. Cui et al. 2006, 2007; Wolf et al. 2008), in outbred population (e.g., de Koning et al. 2002) and in human population (e.g., Hanson et al. 2001; Shete et al. 2003). Broadly speaking, these methods can be categorized into two frameworks: one based on the fixed effect model where the iQTL effect is considered as fixed (e.g., Cui et al. 2006, 2007; de Koning et al. 2002), and the other considering iQTL effect as random and estimating the genetic variances contributed by an iQTL (e.g. Hanson et al. 2001; Shete et al. 2003; Li and Cui 2009a). The method proposed by Li and Cui (2009a) extended the variance components model to experimental crosses and showed relative merits in mapping iQTLs with inbred lines. However, all these approaches for iQTL mapping were developed based on diploid populations, whereby chromosomes are paired. Their applications are immediately limited when the ploidy level of the study population is more than two for instance, the triploid endosperm.

In this study, we propose to extend our previous work in iQTL mapping with variance components approach in experimental crosses (Li and Cui 2009a), and consider the unique genetic make-up of the triploid endosperm genome to map iQTLs underlying triploid endosperm traits. Cytoplasmic maternal effects are also considered and adjusted when testing for genomic imprinting. Motivated by a real experiment, we propose a reciprocal backcross design initiated with two inbred lines. Likelihood ratio test (LRT) is applied to test the significance of the variance components and its asymptotic distribution is evaluated under irregular conditions.

The article is organized as follows. Section 2 will illustrate the basic genetic design

and the statistical mapping framework. We propose a new approach for calculating the parental specific allelic sharing among inbreeding triploid sibs. Statistical hypothesis testings are proposed to assess iQTL effects. The limiting distribution of the LRT under the proposed mapping framework is studied. Multiple QTL model is also proposed to separate closely linked QTLs. Section 3 and 4 will be devoted to simulations and real application followed by a general discussion in section 5.

2.2 STATISTICAL METHOD

2.2.1 The genetic design

Using experimental crosses for QTL mapping has been the traditional means in targeting genetic regions harboring potential genes responsible for quantitative trait variations. Toward the goal of mapping iQTL underlying endosperm traits in line crosses, we propose a reciprocal backcross design. A similar design was proposed by Li and Cui (2009a) for diploid mapping populations. In brief, two inbred parents with genotypes AA and aa are crossed to produce an F_1 population (Aa). F_1 individuals are then backcrossed with one of the parents to generate backcross populations. We can use both parents as the maternal strain to cross with an F_1 individual to generate two backcross segregation populations. Or we can use F_1 individuals as the maternal strains to cross with both parents to produce another two sets of segregation populations. The so called reciprocal backcross design generates four different segregation populations with each one being considered as one family. Large number of backcross

families can be obtained by simply replicating each one of the above crosses.

To distinguish the allelic parental origin, we use subscript letter f and m to denote an allele inherited from the father and mother, respectively. A list of possible offspring genotypes considering the unique genetic make-ups in the triploid endosperm genome is detailed in the second column in Table 2.1. Clearly, the endosperm genome carries one extra maternal copy due to the unique double fertilization step in flowering plants. When a dosage effect is considered, we do expect different expression values triggered by endosperm and embryo genes.

Table 2.1: The allelic-specific IBD sharing coefficients for full-sib pairs in a reciprocal backcross design.

Backcross	Offspring genotype	Parent-specific IBD sharing						Total IBD	
		π_{mm}		π_{ff}		$\pi_{m/f}$		π	
		$QmQmQ_f$	$QmQmq_f$	$QmQmQ_f$	$QmQmq_f$	$QmQmQ_f$	$QmQmq_f$	$QmQmQ_f$	$QmQmq_f$
$QQ \times Qq$	$QmQmQ_f$	4/3	4/3	1/3	0	4/3	2/3	3	2
	$QmQmq_f$	4/3	4/3	0	1/3	2/3	0	2	5/3
		$QmQmQ_f$	$qmqmQ_f$	$QmQmQ_f$	$qmqmQ_f$	$QmQmQ_f$	$qmqmQ_f$	$QmQmQ_f$	$qmqmQ_f$
$Qq \times QQ$	$QmQmQ_f$	4/3	0	1/3	1/3	4/3	2/3	3	1
	$qmqmQ_f$	0	4/3	1/3	1/3	2/3	0	1	5/3
		$qmqmQ_f$	$qmqmq_f$	$qmqmQ_f$	$qmqmq_f$	$qmqmQ_f$	$qmqmq_f$	$qmqmQ_f$	$qmqmq_f$
$qq \times Qq$	$qmqmQ_f$	4/3	4/3	1/3	0	0	2/3	5/3	2
	$qmqmq_f$	4/3	4/3	0	1/3	2/3	4/3	2	3
		$QmQmq_f$	$qmqmq_f$	$QmQmq_f$	$qmqmq_f$	$QmQmq_f$	$qmqmq_f$	$QmQmq_f$	$qmqmq_f$
$Qq \times qq$	$QmQmq_f$	4/3	0	1/3	1/3	0	2/3	5/3	1
	$qmqmq_f$	0	4/3	1/3	1/3	2/3	4/3	1	3

2.2.2 The model

In QTL mapping, different line crosses can be combined together to increase the parameter inference space via a variance components method (Xie et al. 1998). VC method has been shown to be powerful in assessing genomic imprinting in human linkage analysis (Hanson et al. 2001). Recently, Li and Cui (2009a) extended the VC model to experimental crosses and proposed an iQTL mapping framework via combining different line crosses for iQTL detection. We extend our previous work to triploid endosperm tissue considering the unique genetic components in the endosperm genome.

Suppose total K families are collected which are composed of the four distinct backcross families. Assume n_k individuals are sampled in the kth family. The phenotypic variation of a quantitative trait in family k (denoted as y_k) can be explained by the genotype-specific cytoplasmic maternal effect (denoted as μ_k), additive QTL effect (denoted as a_k), polygene effect (denoted as g_k), and random residual effect (denoted as e_k). To incorporate the parent-of-origin effect, the additive QTL effect (a_k) can be further partitioned into two separate effects, an effect due to the expression of the maternal allele (denoted as a_{km}) and an effect due to the expression of the paternal allele (denoted as a_{kf}). The model can thus be expressed as

$$y_{ki} = \mu_k + 2a_{kmi} + a_{kfi} + g_{ki} + e_{ki}, \quad k = 1, \cdots, K; \; i = 1, \cdots, n_k \qquad (2.2.1)$$

where a_{kmi}, a_{kfi}, g_{ki} and e_{ki} are random effects with normal distribution, i.e.,

$a_{kmi} \sim N(0, \sigma_m^2)$, $a_{kfi} \sim N(0, \sigma_f^2)$, $g_{ki} \sim N(0, \sigma_g^2)$, $e_{ki} \sim N(0, \sigma_e^2)$; g_{ki} and e_{ki} are uncorrelated to a_{kmi} and a_{kfi}; the coefficient 2 for a_{kmi} adjusts for the effects of two identical maternal copies; μ_k models the maternal genotype-specific effect. With four distinct segregation populations, we have only three distinct maternal genotypes, AA, Aa and aa. Thus the parameter μ_k can be collapsed into three distinct values denoted as μ_1, μ_2 and μ_3 corresponding to maternal genotypes AA, Aa and aa, respectively. Let $\beta = (\mu_1, \mu_2, \mu_3)$, then model (3.2.1) can be rewritten in a vector form as

$$\boldsymbol{y}_k = X_k\beta + 2\boldsymbol{a}_{km} + \boldsymbol{a}_{kf} + \boldsymbol{g}_k + \boldsymbol{e}_k, \quad k = 1, \cdots, K \tag{2.2.2}$$

where X_k is an $n_k \times 3$ matrix with one column of ones and two columns of zeros.

2.2.3 Parent-specific allele sharing and the covariance between two inbreeding sibs

One of the major tasks in IBD-based iQTL mapping with variance components model is to calculating the IBD sharing probabilities and the phenotypic covariances between sibs. Such a method has been developed in human population (Hanson et al. 2001), which however, can not be applied to a complete inbreeding population in experimental crosses, because the allelic sharing relationship among sibpairs does not follow the pattern as the one derived from a natural non-inbreeding population. Instead, the IBD sharing probability can be calculated based on the Malécot's coefficient of coancestry (1948) for an inbreeding population. Li and Cui (2009a) recently explored different allelic sharing patterns among sibpairs in a reciprocal backcross design with

a diploid tissue. We extend the method to the triploid endosperm genome and derive the covariances among sibpairs in a triploid tissue.

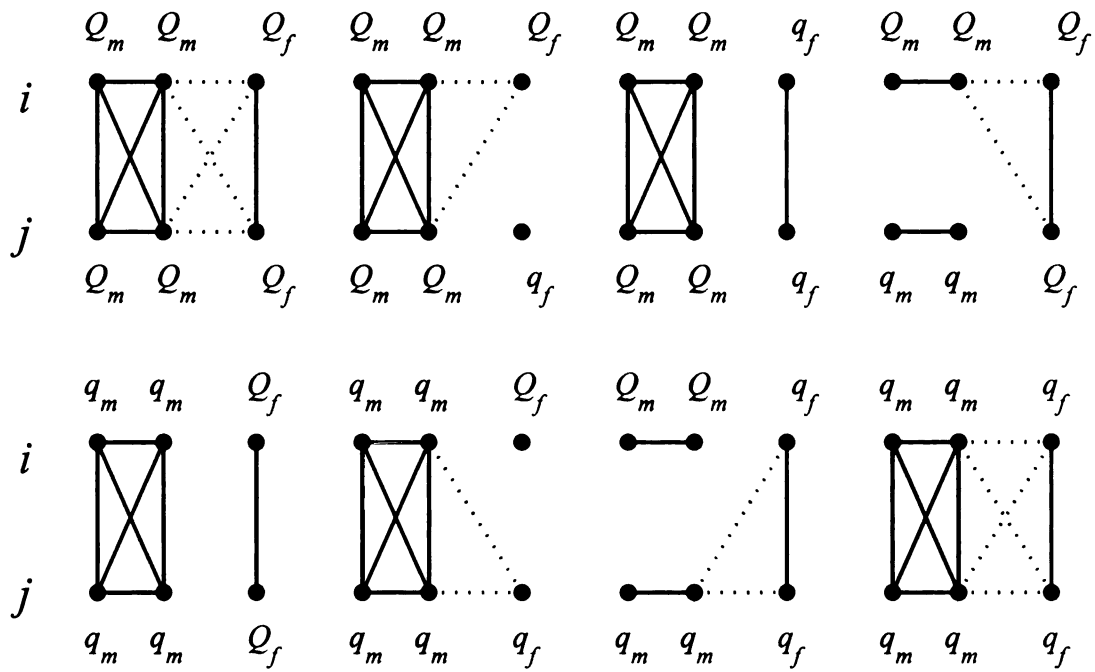

Figure 2.1: Possible alleles shared IBD for individuals i and j in inbreeding backcross families. The solid lines indicate IBD sharing for alleles inherited from the same parent. The dotted lines indicate IBD cross-sharing for alleles inherited from different parents.

Consider two individuals i and j randomly selected from one backcross family with phenotype y_i and y_j. Figure 2.1 shows all possible allelic sharing patterns between individuals i and j. The solid line indicates IBD sharing for alleles derived from the same parent and the dotted line indicates IBD cross-sharing for alleles derived from different parents. The allelic cross-sharing is unique to inbreeding populations, whereby this cross-sharing probability reduces to zero for non-inbreeding populations. Here we propose to calculate the IBD sharing between individuals i and j (denoted as π_{ij}) for a triploid genome as

$$\pi_{ij} = \begin{cases} 3\theta_{ij} & \text{if } i \neq j \\ \frac{1}{3}(5 + 3F_i) & \text{if } i = j \end{cases} \tag{2.2.3}$$

where θ_{ij} is the Malécot's coefficient of coancestry and F_i is the inbreeding coefficient (Harris 1964; Cockerham 1983; Lynch and Walsh 1998). By definition, θ_{ij} is calculated as the probability of two randomly selected alleles from individuals i and j being identical by descent. The calculation of π_{ij} is different from the usual IBD sharing calculation in non-inbreeding populations. It is rather interpreted as triple the Malécot's coefficient of coancestry (Xie et al. 1998). For easy notation, we still adopt the term "IBD sharing probability" for π_{ij} in the rest of the presentation. The calculation of the inbreeding coefficient follows the procedure given in Lynch and Walsh (1998).

To illustrate the idea, consider two backcross individuals i (with genotype $A_m A_m A_f$) and j (with genotype $B_m B_m B_f$). The coefficient of coancestry θ_{ij} between these

two individuals can be expressed as,

$$\begin{aligned}\theta_{ij} &= \frac{1}{9}\{\Pr(A_{m1} \equiv B_{m1}) + \Pr(A_{m1} \equiv B_{m2}) + \Pr(A_{m2} \equiv B_{m1}) \\ &\quad + \Pr(A_{m2} \equiv B_{m2}) + \Pr(A_{m1} \equiv B_f) + \Pr(A_{m2} \equiv B_f) + \Pr(A_f \equiv B_{m1}) \\ &\quad + \Pr(A_f \equiv B_{m2}) + \Pr(A_f \equiv B_f)\} \\ &= \frac{1}{9}(4\theta_{imjm} + 2\theta_{imj_f} + 2\theta_{i_fjm} + \theta_{i_fj_f})\end{aligned}$$

where the notation $\equiv$ refers to identical by decent; the subscript numbers 1 and 2 indicate two maternally inherited alleles; $\theta_{i.j.}$ is defined as the allelic kinship coefficient (Lynch and Walsh 1998). Noted that the two terms θ_{imj_f} and θ_{i_fjm} are indistinguishable, but their sum denoted as $\theta_{im/j_f}(= \theta_{imj_f} + \theta_{i_fjm})$ is unique. Thus, we have $\theta_{ij} = \frac{1}{9}(4\theta_{imjm} + 2\theta_{im/j_f} + \theta_{i_fj_f})$. Following equation (2.2.3), we have

$$\pi_{ij} = 3\theta_{ij} = \frac{4}{3}\theta_{imjm} + \frac{2}{3}\theta_{im/j_f} + \frac{1}{3}\theta_{i_fj_f} = \pi_{imjm} + \pi_{im/j_f} + \pi_{i_fj_f} \text{ for } i \neq j$$

It can be seen that the IBD sharing between any two individuals can be decomposed as three separate components, one due to the IBD sharing for alleles derived from the maternal parent ($\pi_{imjm} = \frac{4}{3}\theta_{imjm}$), one due to the cross-sharing for alleles derived from different parents ($\pi_{im/j_f} = \frac{2}{3}\theta_{im/j_f}$), and one due to the IBD sharing for alleles derived from the paternal parent ($\pi_{i_fj_f} = \frac{1}{3}\theta_{i_fj_f}$). An exhaustive list of all possible IBD sharing probabilities for the four backcross families is given in Table 2.1.

Dropping the family index k, the covariance between any two individuals i and j can be expressed as,

$$\begin{aligned}\mathrm{Cov}(y_i, y_j|\pi_{i_m j_m}, \pi_{i_m/j_f}, \pi_{i_f j_f}) &= \mathrm{Cov}(2a_{mi} + a_{fi} + g_i + e_i, \\ &\quad 2a_{mj} + a_{fj} + g_j + e_j) \\ &= 4\pi'_{i_m j_m}\sigma^2_m + 2\pi'_{i_m/j_f}\sigma^2_{mf} + \pi_{i_f j_f}\sigma^2_f \\ &\quad +\phi_{ij}\sigma^2_g + I_{ij}\sigma^2_e\end{aligned}$$

where $\pi'_{i_m j_m} = \frac{1}{4}(\pi_{i_m j_m})$ and $\pi'_{i_m/j_f} = \frac{1}{2}(\pi_{i_m/j_f})$ are the IBD sharing and cross-sharing probabilities by considering one single maternal allele; σ^2_{mf} measures the variation of trait distribution due to alleles cross-sharing; ϕ_{ij} is the expected alleles shared IBD; I_{ij} is an indicator variable taking value 1 if $i = j$ and 0 if $i \neq j$. For a natural population without inbreeding, there is no allele cross-sharing for an individual with itself, hence $\pi_{i_m/j_f} = 0$. For a diploid non-inbreeding population, the trait covariance can be simplified as the one given in Shete et al. (2003). In matrix form, the phenotypic variance-covariance for individuals in the kth backcross family can then be expressed as

$$\mathbf{\Sigma_k} = \mathbf{\Pi}_{m|k}\sigma^2_m + \mathbf{\Pi}_{m/f|k}\sigma^2_{mf} + \mathbf{\Pi}_{f|k}\sigma^2_f + \mathbf{\Phi}_{g|k}\sigma^2_g + \mathbf{I}\sigma^2_e \qquad (2.2.4)$$

where the elements of $\mathbf{\Pi}_{m|k}$, $\mathbf{\Pi}_{f|k}$ and $\mathbf{\Pi}_{m/f|k}$ can be found in Table 2.1.

2.2.4 QTL IBD sharing and genome-wide linkage scan

The above described IBD sharing probability is calculated at a known marker position. Unless markers are dense enough, we have to search across the genome for potential (i)QTL positions and their effects. In general, the QTL position can be viewed as a fixed parameter by searching for a putative QTL at every 1 or 2 cM on a map interval bracketed by two markers throughout the entire linkage map. Thus, we need to estimate the QTL IBD sharing at every scan position. Since the conditional probability of an endosperm QTL given upon two flanking markers is the same as the one derived from a diploid genome (Cui and Wu 2005), the same procedure termed as the expected conditional IBD sharing described in Li and Cui (2009a) can be applied to calculate the QTL IBD sharing probability at every scan position.

Assuming multivariate normality of the trait distribution for data in each family and assuming independence between families, the joint log-likelihood function when K backcross families are sampled can be formulated as

$$\ell = \sum_{k=1}^{K} \log[f(\boldsymbol{y}_k; \mu_k, \boldsymbol{\Sigma}_k)] \tag{2.2.5}$$

where f is the multivariate normal density. Parameters to be estimated include $\beta = (\mu_1, \mu_2, u_3)$ and $\Omega = (\sigma^2_m, \sigma^2_f, \sigma^2_{mf}, \sigma^2_g, \sigma^2_e)$. Two commonly used methods in linkage analysis, the maximum likelihood (ML) method and the restricted maximum likelihood (REML) method, may be applied to estimate parameters. It is commonly recognized that the REML method gives less biased estimation compared to the ML

method (Corbeil and Searle 1976). Here we adopted the REML method with the Fisher scoring algorithm to obtain the REML estimates of the parameters (see Li and Cui 2009a for details of the algorithm).

The conditional QTL IBD-sharing values vary at different testing positions. The amount of support for a QTL at a particular map position can be displayed graphically through the use of likelihood ratio profiles, which reflect the variation of the testing position of putative QTLs. The significant QTLs are detected by the peaks of the profile plot that pass certain significant threshold (see section 2.5 for more details).

2.2.5 Hypothesis testing

In iQTL mapping, we are interested in testing whether there is any significant genetic effect at a test position and would like to further quantify the imprinting effect if any. The hypothesis for testing the existence of a QTL can be expressed as

$$\begin{cases} H_0 : \sigma_m^2 = \sigma_f^2 = \sigma_{mf}^2 = 0 \\ \\ H_1 : \text{at least one parameter is not zero} \end{cases} \tag{2.2.6}$$

The LRT is applied for this purpose. Define $\widetilde{\Omega}$ and $\widehat{\Omega}$ to be the estimates of the unknown parameters under H_0 and H_1, respectively. The LRT statistic can be calculated as

$$\text{LR} = -2[\log L(\widetilde{\Omega}|\boldsymbol{y}) - \log L(\widehat{\Omega}|\boldsymbol{y})] \tag{2.2.7}$$

Let $\boldsymbol{\theta} = (\mu_1\ \mu_2\ \mu_3\ \theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5)^T = (\mu_1\ \mu_2\ \mu_3\ \sigma_m^2\ \sigma_f^2\ \sigma_{mf}^2\ \sigma_g^2\ \sigma_e^2)^T \in \Omega = \mathbb{R}^3 \times [0,\infty) \times [0,\infty) \times [0,\infty) \times (0,\infty) \times (0,\infty)$ be the parameters to be estimated. Noted that the polygene variance is bounded away from zero if we assume there are more than one QTL in the genome. Let the true parameters under the null hypothesis be $\boldsymbol{\theta}_0 = (\mu_{10}\ \mu_{20}\ \mu_{30}\ \sigma_{m_0}^2\ \sigma_{f_0}^2\ \sigma_{mf_0}^2\ \sigma_{g_0}^2\ \sigma_{e_0}^2)^T = (\mu_{10}\ \mu_{20}\ \mu_{30}\ 0\ 0\ 0\ \sigma_{g_0}^2\ \sigma_{e_0}^2)^T \in \Omega_0 = \mathbb{R}^3 \times \{0\} \times \{0\} \times \{0\} \times (0,\infty) \times (0,\infty)$. The three tested genetic variance components under the null hypothesis lie on the boundaries of the parameter space Ω. Thus, the standard conditions for obtaining the asymptotic χ^2 distribution of the LRT are not satisfied (Self and Liang 1987). Following the results from Chernoff (1954), Shapiro (1985) and Self & Liang (1987), the following theorem shows that the LR statistic follows a mixture chi-square distribution, whereby the mixture proportions depend on the estimated Fisher information matrix.

Theorem 2.2.1. *Let C_{Ω_0} and C_Ω be closed convex cones with vertex at $\boldsymbol{\theta}_0$ to approximate Ω_0 and Ω, respectively. Let $\boldsymbol{Y}$ be a random variable with a multivariate normal distribution with mean $\boldsymbol{\theta}_0$, and variance-covariance matrix $I^{-1}(\boldsymbol{\theta}_0)$. Under the assumptions given in the Appendix, the LR statistic in (3.2.10) is asymptotically distributed as a mixture chi-square distribution with the form $\omega_3\chi_3^2 : \omega_2\chi_2^2 : \omega_1\chi_1^2 : \omega_0\chi_0^2$, where $\omega_3 = \frac{1}{4\pi}[2\pi - cos^{-1}\rho_{12} - cos^{-1}\rho_{13} - cos^{-1}\rho_{23}]$, $\omega_2 = \frac{1}{4\pi}[3\pi - cos^{-1}\rho_{12|3} - cos^{-1}\rho_{13|2} - cos^{-1}\rho_{23|1}]$, $\omega_1 = \frac{1}{4\pi}(cos^{-1}\rho_{12} + cos^{-1}\rho_{13} + cos^{-1}\rho_{23})$, and $\omega_0 = \frac{1}{2} - \frac{1}{4\pi}[3\pi - cos^{-1}\rho_{12|3} - cos^{-1}\rho_{13|2} - cos^{-1}\rho_{23|1}]$; ρ_{ab} is the correlation between the variance terms a and b calculated from the Fisher information matrix, and $\rho_{ab|c} = \frac{(\rho_{ab} - \rho_{ac}\rho_{bc})}{(1-\rho_{ac}^2)^{1/2}(1-\rho_{bc}^2)^{1/2}}$.*

Note that the symbol π in the above theorem is the irrational number (a mathematical constant) not the IBD sharing probability. The proof of the theorem is given in Appendix.

Remark: When the random parameter estimators are uncorrelated or the correlation is extremely small, i.e., the Fisher information matrix is close to diagonal, the mixture proportions for the χ^2_k components are reduced to the binomial form with $\binom{3}{k}2^{-3}$, which is consistent with the results (Case 9) given in Self and Liang (1987).

Once a QTL is identified at a genomic position, we can further assess its imprinting property. To evaluate whether a QTL shows imprinting effect, the hypotheses can be formulated as

$$\begin{cases} H_0 : \sigma_f^2 = \sigma_m^2 \\ H_1 : \sigma_f^2 \neq \sigma_m^2 \end{cases} \tag{2.2.8}$$

Again, the likelihood ratio test can be applied which asymptotically follows a χ^2 distribution with 1 degree of freedom since the tested parameter under the null is nonnegative and does not lie on the boundary of the parameter space. Rejecting H_0 indicates genomic imprinting, and the QTL can be called an iQTL. We denote this imprinting test as LR_{imp}. If the null is rejected, one would be interested in testing whether the detected iQTL is completely maternally or paternally imprinted with the corresponding null hypothesis expressed as $H_0 : \sigma_m^2 = 0$ and $H_0 : \sigma_f^2 = 0$, respectively. The LRT statistic for the two tests asymptotically follows a mixture χ^2 distribution with the form $\frac{1}{2}\chi_0^2 : \frac{1}{2}\chi_1^2$. Rejection of complete imprinting indicates partial imprinting.

Maternal effects can be tested by formulating hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$. Note that these three parameters do not represent the true maternal effects as they are confounded with the main genetic effects. But a test of pairwise differences can be applied to detect the significance of any maternal contribution.

2.2.6 Multiple iQTL model

In practice, there may be several QTLs to reflect the phenotypic variation in the whole genome. When testing QTL effects at one chromosome, the effects from QTLs located at other chromosomes are absorbed by the polygenic effect (g). In some case, two or more QTLs may located at the same chromosome, which are termed background QTLs in comparison to the tested one. When this happens, it is essential to adjust for the background QTLs' effects. Otherwise, it may lead to biased estimation for the putative QTL caused by the interference of QTLs close to the tested interval (Zeng 1994).

In the previous work of Li and Cui (2009a), the authors proposed a multiple iQTL model following the idea of next-to-flanking markers proposed by Xu and Atchley (1995). We adopted a similar strategy in the current study. Briefly, assume there are S (i)QTLs in one chromosome, the multiple iQTL model considering parent-specific allele effect can be expressed as

$$y_{ki} = \mu_k + \sum_{s=1}^{S} 2a_{kmis} + \sum_{s=1}^{S} a_{kfis} + g_{ki} + e_{ki}, \quad k = 1, \cdots, K;\ i = 1, \cdots, n_k$$

where each (i)QTL effect is partitioned as two separate terms to reflect the contribu-

tion of the maternal and paternal alleles. In reality, the exact number and location of the QTLs in a chromosome is generally unknown before the genome-wide search. This problem can be eased by applying the next-to-flanking markers idea proposed by Xu and Atchley (1995).

Denote a test interval with two flanking markers as $\mathcal{M}_l$—$\mathcal{M}_r$. The markers next to these two markers are denoted as $\mathcal{M}_L$ on the left of $\mathcal{M}_l$, and $\mathcal{M}_R$ on the right of $\mathcal{M}_r$ ($L = l - 1$ and $R = r + 1$). Conditional on the two markers, $\mathcal{M}_L$ and $\mathcal{M}_R$, we expect the effects of QTLs located outside of the tested interval can be absorbed by the IBD values calculated from the two next-to-flanking markers (Xu and Atchley 1995). Thus, the calculation of (i)QTL covariance conditional on these two markers will avoid the requirement for the position of QTLs outside of the tested interval. The phenotypic covariance between two individuals i and j can be expressed as

$$
\begin{aligned}
&Cov(y_{ki}, y_{kj} | \pi_{L|k}, \hat{\pi}_{i_m j_m}, \hat{\pi}_{i_m/j_f}, \hat{\pi}_{i_f j_f}, \pi_{R|k}) \\
=&\sum_{l=1}^{L} K(\theta_{lL}, \pi_{L|k})\sigma_l^2 + \hat{\pi}_{i_m j_m}\sigma_m^2 + \hat{\pi}_{i_m/j_f|k}\sigma_{mf}^2 + \hat{\pi}_{i_f j_f}\sigma_f^2 \\
&+ \sum_{r=1}^{R} K(\theta_{lR}, \pi_{R|k})\sigma_r^2 + \phi_{ij}\sigma_g^2 + I_{ij}\sigma_e^2 \\
=&\pi_{L|k}\sigma_L^2 + \hat{\pi}_{i_m j_m}\sigma_m^2 + \hat{\pi}_{i_m/j_f|k}\sigma_{mf}^2 + \hat{\pi}_{i_f i_f}\sigma_f^2 + \pi_{R|k}\sigma_R^2 + \phi_{ij}\sigma_g^2 + I_{ij}\sigma_e^2
\end{aligned}
$$

where $\pi_{L|k}$ is the IBD sharing value at marker L, and σ_L^2 is a composite variance component which reflects the variation of (i)QTL effects on the left side of the tested interval (see Li and Cui 2009a for details). $\pi_{R|k}$ and σ_R^2 are defined similarly. The

calculation of $\pi_{L|k}$ and $\pi_{R|k}$ reflect the triploid structure of the endosperm genome. Testing (i)QTL effects can then be focused on a tested interval while adjusting for the background QTLs' effects located in other place.

2.3 SIMULATION

2.3.1 Single iQTL simulation

Six evenly spaced markers are simulated with a total length of 100cM. For simplicity, we assume equal family size (i.e., $n_k = n$). A putative iQTL is simulated at 48cM away from the first marker. The effect of the putative iQTL is simulated by assuming different imprinting modes (i.e., no imprinting, completely imprinting and partial imprinting). Phenotypes are simulated by randomly drawing multivariate normal distribution with the covariance structure given in model (2.4) with different parameter combinations.

Table 2.2: The power, REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL with no imprinting effect under different sampling designs. The square roots of the mean squared errors of the parameter estimates are given in parentheses.

	Position	$\mu 1$	$\mu 2$	$\mu 3$	σ_m^2	σ_f^2	σ_{mf}^2	σ_g^2	σ_e^2			
$n_F \times n_k$	48cM	13	12	10	0.75	0.75	1	1	2.5	Power	Error^1	Error^2
4×100	47.98	12.94	12.03	9.62	0.329	0.381	1.339	1.363	2.517	0.58	0.1	0.15
	(19.631)	(1.93)	(1.32)	(2.12)	(0.671)	(0.779)	(2.224)	(1.02)	(0.36)			
8×50	48.21	13.03	12.05	10.08	0.611	0.665	1.16	1.08	2.545	0.85	0.1	0.11
	(15.322)	(1.52)	(0.93)	(1.55)	(0.585)	(0.908)	(1.026)	(0.799)	(0.337)			
20×20	48.20	13.06	11.99	9.94	0.751	0.761	1.27	0.94	2.506	0.88	0.03	0.05
	(14.924)	(0.85)	(0.56)	(0.88)	(0.488)	(0.618)	(0.958)	(0.576)	(0.293)			
100×4	47.5	13.00	12.01	9.96	0.647	0.881	0.74	1.15	2.477	0.46	0.09	0.06
	(23.281)	(0.35)	(0.27)	(0.37)	(0.600)	(0.935)	(0.739)	(0.604)	(0.306)			

Note n_F and n_k refer to the number of families and offspring size for each family, respectively. The locations of the the QTL is described by the map distances (in cM) from the first marker of the linkage group (100 cM long); The true QTL is located at 48cM; Error^1 and Error^2 refer to the type I errors for testing $H_0 : \sigma_m^2 = \sigma_f^2 = \sigma_{mf} = 0$ and $H_0 : \sigma_m^2 = \sigma_f^2$, respectively. Note that Error^1 is estimated when data are simulated assuming $\sigma_m^2 = \sigma_f^2 = \sigma_{mf} = 0$.

For experimental crosses, the number of families and the offspring size can be easily controlled. We simulate data assuming different family and offspring size combinations to evaluate the effect of family and offspring size on testing power and parameter estimation. For a fixed total sample size of 400, we vary the family and offspring size with different combinations, i.e., 4×100, 8×50, 20×20 and 100×4. The first number for each combination indicates the family size and the second number indicates the offspring size. Without loss of generality, we assume equal offspring size for all families in each simulation. Results with 100 Monte Carlo repetitions are recorded for each simulation.

Table 2.2 tabulates the results assuming no imprinting (i.e., $\sigma_m^2 = \sigma_f^2$). The simulated parameter values are listed underneath each parameter. The REML estimates as well as the root mean squared errors (RMSEs) (given in the parenthesis) are recorded for each simulation. n_F denotes the number of families and n_k denotes the number of offsprings in each family. Overall, the 20 × 20 combination produces the smallest RMSE and bias for QTL position estimation, high QTL detection power and reasonable type I error rate among the four designs. The 100 × 4 design gives the most accurate estimates for the maternal effects, but with small power to detect QTL effect. The 4 × 100 design gives very biased parameter estimates for the main maternal and paternal variance terms. The 20× 20 design also produces the most reasonable imprinting type I error. Thus, a balance of the family and offspring size is necessary in achieving optimal power and estimation precision for the QTL position and genetic effects. In reality, one should always try to avoid designs with extremely

large or small family size.

Since the 20 × 20 design outperforms the others, we focus this design and conduct additional simulations under different imprinting mechanisms. The results are summarized in Table 2.3. Four imprinting action modes are assumed: complete paternal imprinting ($\sigma_m^2 = 1.5$ and σ_f^2=0), complete maternal imprinting ($\sigma_m^2 = 0$ and σ_f^2=1.5), partial maternal imprinting ($\sigma_m^2 = 0.5$ and σ_f^2=1) and partial paternal imprinting ($\sigma_m^2 = 1$ and σ_f^2=0.5). Overall, all parameters can be properly estimated with reasonable precision under different scenarios. The complete maternal imprinting has the lowest overall QTL testing power (62%) compared to others. Since the majority of the total variance comes from the maternal alleles (two copies), this result is expected. Also noted that the imprinting power is low in the four cases. Since the size of the real data analyzed in section 4 is close to 400, we focus our simulation with a total size of 400. As the total sample size increases, we do observe increased imprinting power (data not shown).

The imprinting power is listed in the last column of Table 2.3, which varies a lot under different imprinting cases. Note the imprinting power is calculated only when a simulated QTL is significant. Simulations are not counted when calculating the imprinting power when no QTLs are detected. Thus, we expect low imprinting power under the current simulation design given that the overall QTL detection power is less than 90%. The observed low imprinting power might be due to small sample size. When sample size is increased, we do observe increased imprinting power (data not shown). For more explanations, see the Discussion section.

Table 2.3: The power, REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL showing different imprinting effects under the 20×20 sampling design. The square roots of the mean squared errors of the parameters are given in parentheses.

Position	$\mu 1$	$\mu 2$	$\mu 3$	σ_m^2	σ_f^2	σ_{mf}^2	σ_g^2	σ_e^2		
48cM	13	12	10			0.5	0.5	2	Power	iPower
				1.5	0					
47.52	13.02	12.09	9.93	0.95	0.17	0.52	1.16	1.93	0.82	0.28
(13.69)	(0.863)	(0.529)	(0.891)	(0.994)	(0.295)	(0.633)	(1.252)	(0.332)		
				0	1.5					
46.70	13.06	12.05	9.99	0.06	1.21	0.51	0.62	1.98	0.62	0.50
(19.81)	(0.580)	(0.404)	(0.648)	(0.110)	(1.148)	(0.537)	(0.291)	(0.205)		
				0.5	1					
47.72	12.96	12.11	10.05	0.45	0.85	0.62	0.52	2.03	0.88	0.14
(15.45)	(0.684)	(0.495)	(0.736)	(0.320)	(0.773)	(0.607)	(0.377)	(0.227)		
				1	0.5					
48.34	12.97	11.98	10.04	0.74	0.44	0.61	0.76	1.98	0.81	0.11
(14.26)	(0.881)	(0.465)	(0.756)	(0.681)	(0.556)	(0.628)	(0.645)	(0.266)		

Power refers to the overall QTL detection power; iPower refers to the imprinting power for testing $H_0 : \sigma_m^2 = \sigma_f^2$. The imprinting test is only conducted at the position where hypothesis (2.6) is rejected. See Table 2.2 for explanations of other parameters.

In summary, the results show that both the 4×100 and the 100×4 designs yield lower QTL detection power and higher RMSE (root mean squared error) for QTL position estimation than the other two designs do. The 20×20 design slightly beats the 8×50 design with smaller imprinting type I error and higher QTL detection power. These results indicate that it is necessary to maintain a balance between the family size and the offspring size, in order to achieve optimal power and good effects estimation precision. For a given budget with a fixed total sample size, one should always try to avoid extreme designs with large (or small) number of families, each with small (or large) number of offsprings. Focusing on the 20 × 20 design, additional simulation shows that the performance of the imprinting model depends on the underlying degree of imprinting. High imprinting power is observed when an iQTL is maternally imprinting compared to the case when an iQTL is paternally imprinting.

2.3.2 Multiple iQTL simulation

When multiple (i)QTLs occur in one chromosome, especially when they show linkage effects, the inference of a tested QTL will be biased if other QTLs' effects are not corrected. In the simulation, the same setup as described in single iQTL simulation is adopted, except that two putative iQTLs are simulated, one located at 28cM and the other one located at 72cM. Data are simulated assuming two iQTLs located at the two genomic positions and are subject to both single iQTL and multiple iQTL analysis. Figure 2.2 plots the LR profiles averaged out of 100 replications for both analyses.

The dotted and solid curves represent the LR profiles calculated from the single and multiple iQTL models, respectively. Results indicate that the single iQTL analysis produces three clear LR peaks. The highest peak corresponds to the wrong QTL position, which is often termed as "ghost" QTL (Zeng 1994). On the contrary, the multiple iQTL model can correctly target the two QTL positions with high precision as indicated by two distinct LR peaks.

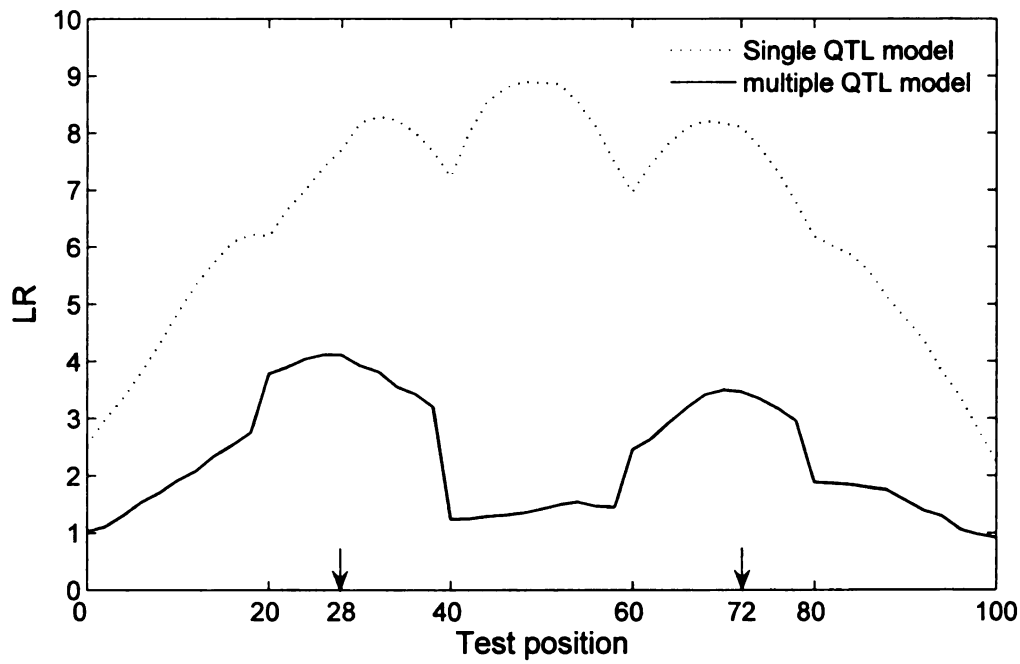

Figure 2.2: The LR profile plot for the single iQTL and multiple iQTL analyses. The true iQTL positions are simulated at $28cM$ and $72cM$ (see the arrow signs). The dotted and the solid curves represent the LR profiles by the single iQTL and multiple iQTL analyses, respectively.

In summary, the results indicate a clear benefit of analysis by fitting a multiple iQTL model than fitting a single iQTL model. While the single iQTl analysis detect one "ghost" QTL located between the two simulated QTLs, the multiple iQTL anal-

ysis can clearly separate the two QTLs with high precision. Note that the multiple iQTL analysis normally generates low LR values than the single iQTL analysis does. The distribution of the LR value under the multiple iQTL analysis is not clear, and permutation should be used to assess significance of any (i)QTLs in multiple iQTL analysis (Xu and Atchley 1995).

2.4 A CASE STUDY

We apply our method to a real data set which have two endosperm traits of interests: mean ploidy level (denoted as Mploidy) and percentage of endoreduplicated nuclei (denoted as Endo). The two traits describe the level of endoreduplication in maize endosperm, which is thought to be genetically controlled by imprinted genes (Dilkes et al. 2002). Four backcross segregation populations, initiated with two inbred lines, Sg18 and Mo17, were sampled. The four populations were obtained from a reciprocal backcross design as illustrated in Table 2.1. The data show large degree of variation for endoreduplication among the four backcross populations, and ten linkage groups were constructed from the observed marker data (Coelho et al. 2007). For more details about the data, readers are referred to Coelho et al. (2007). The two traits were analyzed with our multiple iQTL model aimed to identify iQTLs across the ten linkage groups.

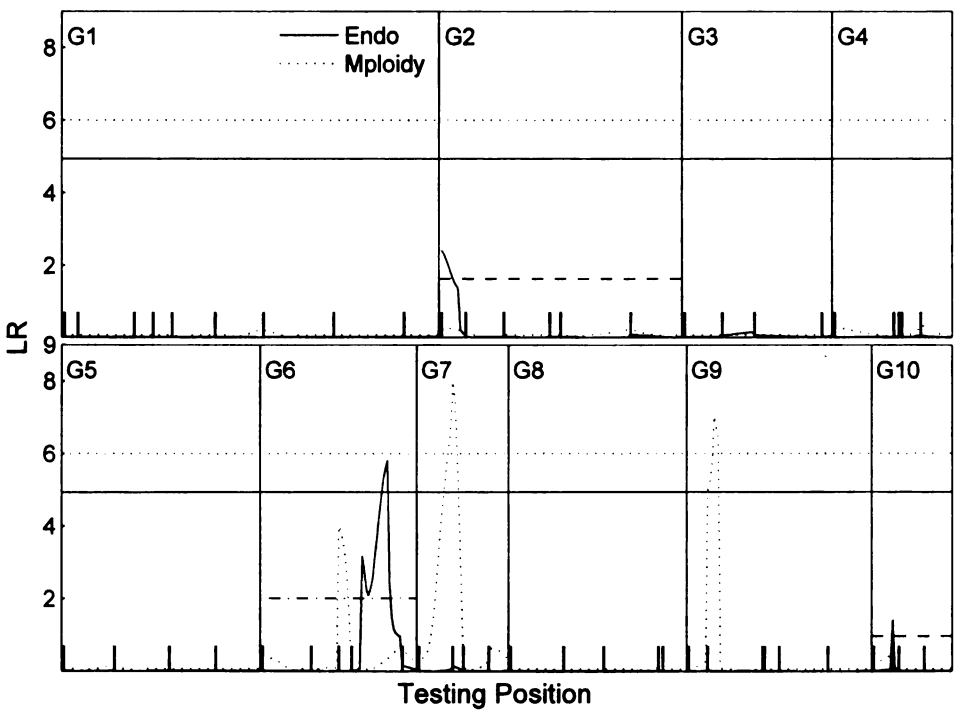

Figure 2.3: The profile of the log-likelihood ratios (LR) for testing the existence of QTLs underlying the two endosperm traits across the 10 maize linkage groups ($G_1, \cdots, G_{10}$). The genome-wide LR profiles for the percentage of endoreduplication (Endo) and mean ploidy (Mploidy) traits are indicated by solid and dotted curves, respectively. The threshold values for claiming the existence of QTLs are given as the horizonal solid and dotted line for the genome-wide threshold, dashed and dash-dotted line for the chromosome-wide threshold, for the two traits Endo and Mploidy, respectively. The genomic positions corresponding to the peak of the curves that pass the corresponding thresholds are the MLEs of the QTL location. The positions of markers on the linkage groups (Coelho et al. 2007) are indicated at ticks.

Figures 2.3 plots the LR profiles across the ten linkage groups for the two traits. The solid and dotted curves represent LR profiles for traits Endo and Mploidy, respectively. To adjust for the genome-wide error rate across the entire linkage group, permutation tests are applied in which the critical threshold value is empirically calculated on the basis of repeatedly shuffling the relationships between marker genotypes and phenotypes (Churchill and Doerge 1994). The corresponding genome-wide significance thresholds (at 5% level) for the two traits are denoted by the horizontal solid (for Endo) and dotted (for Mploidy) lines. The 5% level chromosome-wide thresholds are denoted by the dashed (for Endo) and dash-dotted (for Mploidy) lines. QTLs that are significance at the chromosome-wide level are called suggestive QTLs. It can be seen that two QTLs (on G7 and G9) associated with Mploidy and one QTL (on G6) associated with Endo are detected at the 5% genome-wide significance level (denoted by "*" in Table 3.3). Two suggestive QTLs (on G2 and G10) associated with Endo and one suggestive QTL (on G6) associated with Mploidy are also indetified. The detailed QTL location and effect estimates as well as the test results for imprinting are tabulated in Table 3.3. For the trait Mploidy, the identified three QTLs are all imprinted ($p_{imp} < 0.05$) and all show completely maternal imprinting, i.e., the maternal copies do not express. They are thus termed iQTLs. The cytoplasmic maternal effect does not show any evidence of significance for all the three iQTLs ($p_M > 0.05$). For the trait Endo, only the QTL detected on G6 shows imprinting effect ($p_{imp} < 0.05$) and it shows completely paternal imprinting ($p_{if} < 0.05$). The other two QTLs does not show evidence of imprinting ($p_{imp} > 0.05$). For this trait,

significant maternal effects are detected ($p_M < 0.01$).

In our study, one maternally controlled iQTL was detected for trait Endo, which is consistent with the result given by Dilkes et al. (2002). Meanwhile, according to the genetic conflict theory proposed by Haig and Westoby (1991), in which maternally derived alleles tend to trigger a negative effect on the increase of endosperm growth, whereas paternally derived alleles tend to play an opposite effect to increase seed size. The identified iQTLs shwoing maternal imprinting for trait Mploidy can be well explained by the genetic conflict theory. Both empirical evidence and theoretical hypothesis support the current finding.

Table 2.4: The estimated parameters for the three maternal effects and the variance components for two endosperm traits: mean ploidy (Mploidy) and percent of the endoreduplicated nuclei (Endo).

Trait	Ch	Maternal effects			Genetic effects									
		μ_1	μ_2	μ_3	σ^2_m	σ^2_f	σ^2_{mf}	σ^2_L	σ^2_R	σ^2_g	σ^2_e	p_M	p_{imp}	p_m	p_f
Mploidy	6*	13.13	11.88	9.78	0.01	0.30	0.03	≈ 0	0.22	1.25	2.59	0.34	0.045	0.023	0.31
	7	11.78	11.19	9.16	0.15	0.60	0.94	≈ 0	0.12	1.07	2.69	0.31	0.048	0.024	0.49
	9	13.84	12.08	10.01	≈ 0	0.94	0.71	≈ 0	0.01	1.59	2.55	0.12	0.013	0.021	0.48
Endo	2*	72.23	62.40	52.86	0.43	0.83	2.41	0.99	≈ 0	5.10	37.49	<0.01	0.67	-	-
	6	68.37	63.18	54.92	2.92	≈ 0	7.14	1.42	0.92	1.28	38.91	<0.01	0.02	0.28	0.01
	10*	70.78	62.28	50.67	0.58	0.03	1.52	≈ 0	0.17	3.24	39.20	<0.01	0.29	-	-

The three QTLs for trait Mploidy are located at marker umc1805, marker dupssr9 and umc1040+5.76cM on chromosome 6, 7 and 9, respectively. The three QTLs for trait Endo are located at marker umc2094, bnlg345+33.49cM and MMC501+18cM on chromosome 2, 6 and 10, respectively. QTLs showing significance at the genome-wide significant level are indicated by "*". p_M, p_{imp}, p_m and p_f are the p-values for testing maternal effect ($H_0 : \mu_1 = \mu_2 = \mu_3$), imprinting effect ($H_0 : \sigma^2_m = \sigma^2_f$), complete maternal imprinting ($H_0 : \sigma^2_m = 0$) and complete paternal ($H_0 : \sigma^2_f = 0$), respectively.

In the case study, we also fit a Mendelian model to the data to see if the imprinting model and the Mendelian model produce any different results. The Mendelian model for family k assumes the form

$$\boldsymbol{y}_k = X_k\beta + \boldsymbol{a}_k + \boldsymbol{g}_k + \boldsymbol{e}_k, \quad k = 1, \cdots, K \tag{2.4.1}$$

where $\boldsymbol{a}_k$ is a random vector for the main genetic effect without partitioning it into allelic specific effects. See model (2) in the main paper for an explanation of other parameters.

Figures 2.4 and 2.5 plot the results for the two traits Mean Ploidy Level (Mploidy) and Percentage of Endoreduplication (Endo), respectively. Figure 2.4 indicates that the imprinting and the Mendelian models agree with two QTLs detected, one on G7 and one on G9. Both QTLs are significant at the genome-wide significant level by fitting the imprinting model. But the Mendelian model only detects the one on G7 at the genome-wide level. Each model detects one QTL at the chromosome-wide level on G6. But the two QTLs do not overlap. Further experimental investigation is needed to confirm which model is more robust for this QTL.

The results for fitting the Endo trait is summarized in Figure 2.5. The imprinting model detects three QTLs, on G2, G6 and G10. The one on G6 is significant at the genome-wide level. The other two are only significant at the chromosome-wide level. In contrast, the Mendelian model only detects one QTL on G6 which overlaps with the one identified by the imprinting model. In fact, the two models produce quite similar LR values at QTLs on G2 and G10. Due to high threshold for the Mendelian

model, it fails to detect the two QTLs.

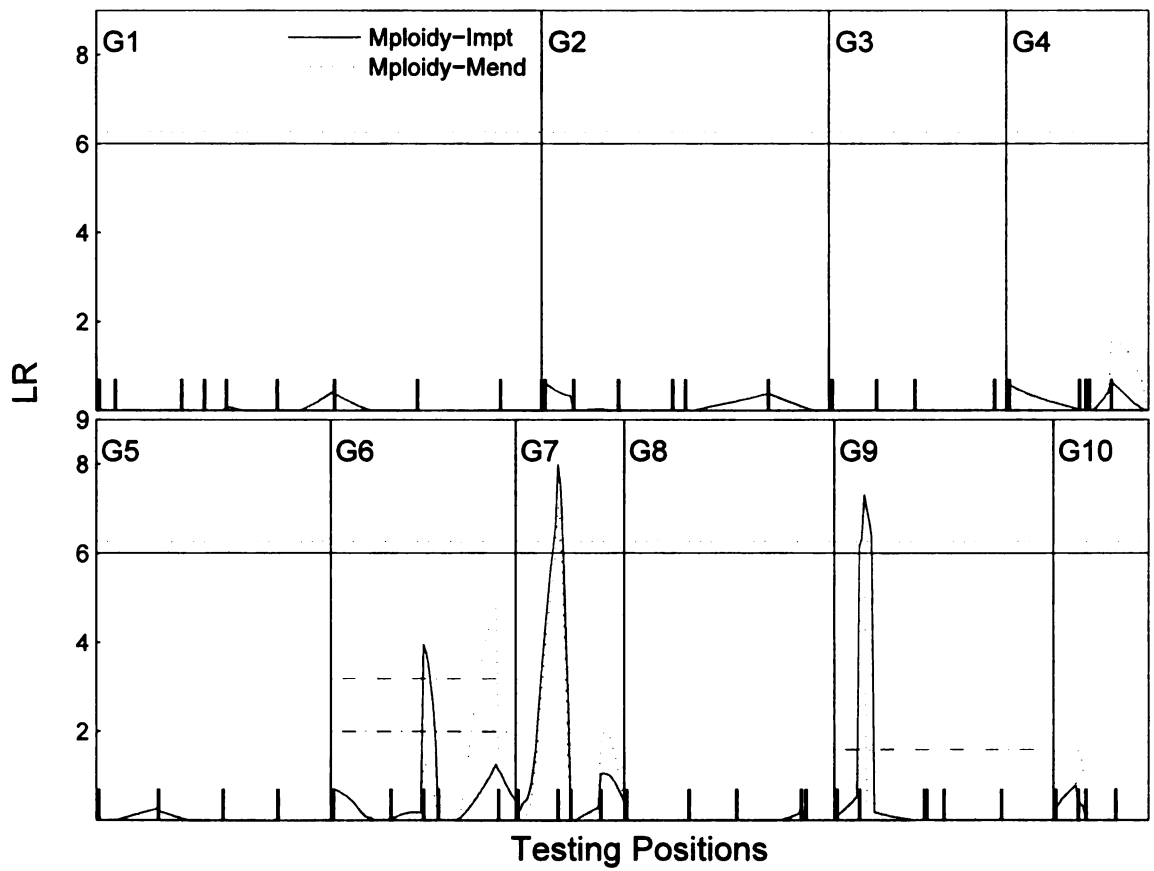

Figure 2.4: The profile of the log-likelihood ratios (LR) for testing the existence of QTLs underlying the trait Mean Ploidy Level across the 10 maize linkage groups ($G_1, \cdots, G_{10}$).

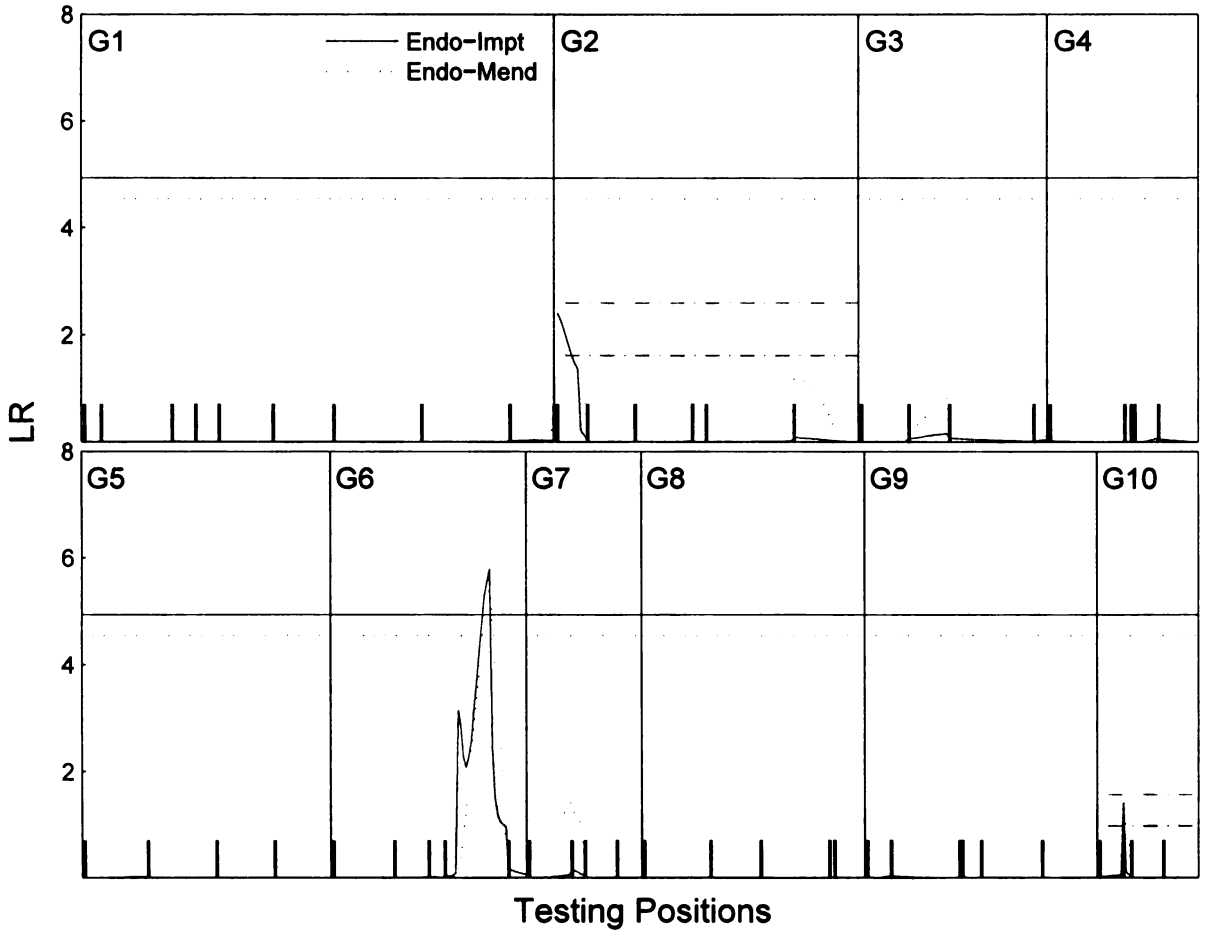

Figure 2.5: The profile of LR values for testing the existence of QTLs underlying the trait Percentage of Endoreduplication across the 10 maize linkage groups ($G_1, \cdots, G_{10}$). See Figure 2.4 for more explanations of the figure.

2.5 DISCUSSION

The role of genomic imprinting in endosperm development has been commonly recognized (Dilkes et al. 2002; Kinkshita et al 1999; Chaudhuri and Messing 1994). But little is known about the exact location and effect size of imprinted genes in endosperm. As endosperm in cereal provides the most nutrition for human being, it is important to identify imprinted genes that govern seed development, particularly endosperm development. In this article, we develop a variance components linkage analysis method with an experimental cross design, aimed to identify iQTLs for endosperm development. Our method is motivated by real applications and is evaluated through Monte Carlo simulations.

The proposed method is based on a particular genetic design (reciprocal backcross design) with inbreeding populations. We treat iQTL effects as random, different from a fixed-effect iQTL model (e.g., Cui 2007). Variance components linkage analysis with partial inbreeding human population was previously proposed (see Abney et al. 2000). However, extending the VC model to a completely inbreeding population is challenging. In our previous work, we proposed a VC-based iQTL mapping framework for an inbreeding diploid mapping population (Li and Cui 2009a). Extending the previous work, we propose a novel IBD partitioning approach to calculate allelic sharing in an inbreeding endosperm population. Extension to mapping multiple iQTLs is provided. Simulations indicate good performance of the multiple iQTL analysis compared to a single iQTL model. Meanwhile, to obtain a good balance of iQTL position and effect estimation and detection power, we have to avoid extreme sample designs. For a fixed

total sample size, extremely large or small families should always be avoided.

In an application to two endosperm traits, we identified three iQTLs for trait Mploidy. All show paternal expression. We also identified one iQTL for trait Endo, which shows a maternal expression. According to the parental conflict theory proposed by Haig and Westoby (1991), maternally derived alleles trigger a negative effect on endosperm cell growth and inhibit endosperm development because the extra maternal copy could slower nuclear division in endosperm. On the contrary, paternally derived alleles tend to increase seed size. Thus, the three iQTL identified for Mploidy can be explained by the genetic conflict theory. The occurrence of parental conflict theory explains parent-of-origin effects as an ubiquitous mechanism for the control of early seed development (Grossniklaus et al. 2001; Kinoshita et al. 1999).

In a VC-based linkage analysis, likelihood ratio test (LRT) has been commonly applied in assessing QTL significance. The LRT statistic asymptotically follows a mixture χ^2 distribution and many investigators often apply the result (Case 9) in Self and Liang (1987) with binomial mixture coefficients. In a recent investigation, we found that the LRT in a regular VC-based linkage analysis without considering imprinting follows a mixture χ^2 distribution with mixture proportions depending on the estimated Fisher information matrix (Li and Cui 2009b). The modified calculation of mixture proportion does give more reasonable type I error rate than the one with binomial coefficients. When imprinting is considered, we show that the limiting distribution of the LRT also follows a mixture χ^2 distribution, and we adopt the new criterion for power evaluation. Simulations show that the new criterion gives

type I error more closer to the nominal level than the one using binomial coefficients, and produces power as good as the later one (data not shown). We recommend investigators to adopt the new criterion in their analysis.

Increasing evidences have suggested that for correlated traits, multivariate approaches can increase the power and precision to identify genetic effects in genetic linkage analyses (e.g., Boomsma and Dolan 1998; Amos et al. 2001; Evans 2002). Also, the joint analysis of multivariate traits can provide a platform for testing a number of biologically interesting hypotheses, such as testing pleiotropic effects of QTL, testing pleiotropic vs close linkage. Moreover, if the putative QTL has pleiotropic effects on several traits, the joint analysis may perform better than mapping each trait separately (Jiang and Zeng 1995). Multivariate traits appear frequently in genetic mapping studies. For example, the two endosperm traits evaluate in this study are highly correlated (Colho et al. 2006). We expect joint analysis may provide high mapping resolution and power for iQTL detection. This will be explored in our future investigation. A computer code written in R is available upon request.

APPENDIX

In standard human linkage analysis with variance components model, many authors declare that the likelihood ratio statistic follows a mixture χ^2 distribution with binomial coefficient for each mixture component (e.g., Amos and Andrade 2001; Hanson et al. 2001; Shete et al. 2003). Following Chernoff (1954), Shapiro (1985) and Self & Liang (1987), in the following we show that the mixture proportion actually depends

on the estimated Fisher information matrix.

For a random variable $\boldsymbol{Y}$ with density function $f(\boldsymbol{y};\boldsymbol{\theta})$, following Chernoff (1954) and Self & Liang (1987), assume that:

(1) When any true parameter (θ_0) is on the boundary, the neighborhood centered at θ_0, i.e., $(\theta_0 - \delta, \theta_0 + \delta)$, is closed, and the intersection between this closure and Ω is also a closed set.

(2) The first three derivatives of $\sum_i logf(\boldsymbol{y}_i;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ on the intersection of neighborhoods of θ_0 and Ω almost surely exist. Moreover, $|\frac{\partial^3 \sum logf}{\partial\theta_i\theta_j\theta_k}| < K(\boldsymbol{y})$ for all θ on the intersection, and $E_\theta[K(\boldsymbol{y})] < \infty$.

(3) The information matrix $\mathcal{I}(\boldsymbol{\theta})$ is positive definite on neighborhoods of $\boldsymbol{\theta}_0$.

Assuming the above assumptions, the consistency and weak convergence of the estimators can be proven (see Chernoff 1954, Self & Liang 1987, Shapiro 1985). Here we cite the main results from Chernoff (1954), Shapiro (1985) and Self & Liang (1987) to show the asymptotic distribution of the LRT in our case.

Defining two closed polyhedral convex cones C_{Ω_0} and C_{Ω_1} to approximate Ω_0 and Ω_1 at $\boldsymbol{\theta}_0$. The parameter space under the null hypothesis is approximated as $C_{\Omega_0} = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^3 \times \{0\} \times \{0\} \times \{0\} \times (0,\infty) \times (0,\infty)\}$, against $C_{\Omega_1} = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^3 \times [0,\infty) \times [0,\infty) \times [0,\infty) \times (0,\infty) \times (0,\infty)\}$ under the alternative. Following Chernoff (1954, Theorem 1), the asymptotic distribution of the LRT in (3.2.10) is equivalent to the following quadratic approximation

$$LR^* = \inf_{\boldsymbol{\theta}\in C_{\Omega_0}} (\boldsymbol{Y}-\boldsymbol{\theta})' I(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}\in C_{\Omega_1}} (\boldsymbol{Y}-\boldsymbol{\theta})' I(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{\theta}) \quad (2.5.1)$$

where $\boldsymbol{Y} \sim N(\boldsymbol{\theta}_0, I^{-1}(\boldsymbol{\theta}_0))$.

Subtracting $\boldsymbol{\theta}_0$ from $\boldsymbol{Y}$ and $\boldsymbol{\theta}$, the expression in (2.5.1) is given by

$$LR^* = \inf_{\boldsymbol{\theta} \in C_{\Omega_0} - \boldsymbol{\theta}_0} (\boldsymbol{Y}-\boldsymbol{\theta})' I(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta} \in C_{\Omega_1} - \boldsymbol{\theta}_0} (\boldsymbol{Y}-\boldsymbol{\theta})' I(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{\theta}) \quad (2.5.2)$$

and $\boldsymbol{Y} \sim \mathrm{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$ under the linear transformation.

Let $C^{\ddagger} = (C_{\Omega_1} - \boldsymbol{\theta}_0) \cap (C_{\Omega_0} - \boldsymbol{\theta}_0)^c = \{\boldsymbol{\theta} : \theta_1 > 0, \theta_2 > 0, \theta_3 > 0\}$, which is a closed polyhedral convex cone with 3 dimensions. By the Pythagoras theorem, the statistic in (2.5.2) can be expressed as

$$LR^* = \inf_{\boldsymbol{\theta} \in C^{\ddagger}} (\boldsymbol{Y} - \boldsymbol{\theta})' I(\boldsymbol{\theta}_0)(\boldsymbol{Y} - \boldsymbol{\theta}) \quad (2.5.3)$$

Let $\mathcal{F}(C^{\ddagger})$ represent the set of all faces of $C^{\ddagger}$, and let $C^{\ddagger 0} = \{\gamma \in \mathbb{R}^3 : \gamma' \boldsymbol{\theta} \leq 0, \ \forall \ \boldsymbol{\theta} \in C^{\ddagger}\}$ be a polar cone which is also a polyhedral convex cone such that $(C^{\ddagger 0})^0 = C^{\ddagger}$. Following Shapiro (1985), we can select a face $\nu \in \mathcal{F}(C^{\ddagger})$ corresponding to a polar face $\nu^0 \in \mathcal{F}(C^{\ddagger 0})$ such that the linear spaces generated by ν and ν^0 are orthogonal to each other. For one face ν (or ν^0), a projection T_ν (or $T_{\nu 0}$) (a symmetric idempotent matrix giving projection onto the space generated by ν (or ν^0)) can be found and $T_\nu = I - T_{\nu_0}$ since they are orthogonal. Then $T_\nu \boldsymbol{Y}$ (or $T_{\nu 0} \boldsymbol{Y}$) is a projection of $\boldsymbol{Y}$ onto $C^{\ddagger}$ (or $C^{\ddagger 0}$). For a given $\boldsymbol{Y}$, let $g(\boldsymbol{Y})$ be the minimizer to achieve the infimum in (2.5.3), such that $LR^* = (\boldsymbol{Y} - g(\boldsymbol{Y}))' I(\boldsymbol{\theta}_0)(\boldsymbol{Y} - g(\boldsymbol{Y}))$. Define $\psi_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y} \in \mathbb{R}^3 : g(\boldsymbol{Y}) \in \nu\})$ so that $g(\boldsymbol{Y}) \in \nu$ if and only if $T_\nu \boldsymbol{Y} \in C^{\ddagger}$

and $T_{\nu 0}\boldsymbol{Y} \in C^{\ddagger 0}$. By Shapiro (1985), $g(\boldsymbol{Y})= T_{\nu}\boldsymbol{Y} \in C^{\ddagger}$, $\forall\ \boldsymbol{Y} \in \psi_{\nu|\boldsymbol{Y}}$.

Note that the set $\psi_{\nu|\boldsymbol{Y}}$ is composed of 2^3 disjoint sets in $\mathbb{R}^3$. All these disjoint sets can be classified into four categories as

1). $\psi^1_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 > 0, g(\boldsymbol{Y}) \in \nu\}$

2). $\psi^2_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \leq 0, g(\boldsymbol{Y}) \in \nu\}$; $\psi^3_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_3 > 0, g(\boldsymbol{Y}) \in \nu\}$; $\psi^4_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_3 > 0, g(\boldsymbol{Y}) \in \nu\}$

3). $\psi^5_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_3 > 0, g(\boldsymbol{Y}) \in \nu\}$; $\psi^6_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_3 \leq 0, g(\boldsymbol{Y}) \in \nu\}$; $\psi^7_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_3 \leq 0, g(\boldsymbol{Y}) \in \nu\}$

4). $\psi^8_{\nu|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_3 \leq 0, g(\boldsymbol{Y}) \in \nu\}$

Define $C^* = \{\boldsymbol{\theta}^* : \boldsymbol{\theta}^* = \Lambda^{1/2}P'\boldsymbol{\theta}, \forall\ \boldsymbol{\theta} \in C^{\ddagger}\}$ to be also a polyhedral closed convex cone. Then 2.5.3 can be further expressed as

$$LR^* = \inf_{\boldsymbol{\theta}^* \in C^*} \|\boldsymbol{z} - \boldsymbol{\theta}^*\|^2 \tag{2.5.4}$$

where $\boldsymbol{z}$=$\Lambda^{1/2}P'\boldsymbol{Y}$ ($P\Lambda P^T = I(\boldsymbol{\theta}_0)$) has a multivariate normal distribution with mean $\mathbf{0}$ and identify covariance matrix.

Let C^{*0} be a polar cone of C^* and $(C^{*0})^0 = C^*$. Two faces ν^* and ν^{*0} can be defined with respect to $\mathcal{F}(C^*)$ and $\mathcal{F}(C^{*0})$. The relevant orthogonal projections T_{ν^*} and $T_{\nu^{*0}}$ corresponding to ν^* and ν^{*0} can be found. Suppose $h(\boldsymbol{z})$ is the minimizer

to achieve the infimum in (2.5.4). Following Shapiro (1985), we can have $h(\boldsymbol{z})=$ $T_{\nu^*}\boldsymbol{z} \in C^*, \forall\, \boldsymbol{z} \in \psi_{\nu^*|\boldsymbol{z}}$. The set $\psi_{\nu^*|\boldsymbol{z}}$ is defined similarly as $\psi_{\nu|\boldsymbol{Y}}$, and it satisfies the conditions of Lemma 3.1 (Shapiro 1985). Then we have

$$LR^* = \|\boldsymbol{z}-h(\boldsymbol{z})\|^2 = \|\boldsymbol{z}-T_{\nu^*}\boldsymbol{z}\|^2 = \boldsymbol{z}'(I-T_{\nu^*})\boldsymbol{z} = \boldsymbol{z}'T_{\nu^*0}\boldsymbol{z}\ \forall\, z \in\ \psi_{\nu^*|\boldsymbol{z}} \quad (2.5.5)$$

Thus the distribution of LR^* in 2.5.3 can be expressed as

$$\begin{aligned} Pr(LR^* > c^2) =& Pr((\boldsymbol{Y}-g(\boldsymbol{Y}))'I(\boldsymbol{\theta}_0)(\boldsymbol{Y}-g(\boldsymbol{Y})) > c^2, \boldsymbol{Y} \in \bigcup_{i=1}^{2^3} \psi^i_{\nu|\boldsymbol{Y}}) \\ =& \sum_{i=1}^{2^3} Pr(\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}}) \\ & Pr((\boldsymbol{Y}-g(\boldsymbol{Y}))'I(\boldsymbol{\theta}_0)(\boldsymbol{Y}-g(\boldsymbol{Y})) > c^2 | \boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}}) \\ =& \sum_{i=1}^{2^3} Pr(\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}}) Pr(\boldsymbol{z}'T_{\nu^*0}\boldsymbol{z} > c^2 | \boldsymbol{z} \in \psi^i_{\nu^*|\boldsymbol{z}}) \end{aligned} \quad (2.5.6)$$

where conditional on $\boldsymbol{z} \in \psi^i_{\nu^*|\boldsymbol{z}}$, $\boldsymbol{z}'T_{\nu^*0}\boldsymbol{z}$ is a chi-square distribution. By Bayes' theorem, the distribution of LR^* follows a mixture chi-square distribution with mixing proportions $Pr(\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}})$ (i=1,...,2^3) and $\sum_{i=1}^{2^3} Pr(\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}}) = 1$.

The calculation of the mixture proportions follows Plackett (1954). Specifically, when $\boldsymbol{Y} \in \psi^1_{\nu|\boldsymbol{Y}}$, $LR^* \sim \chi^2_3$, and the corresponding mixture proportion w_3 =Pr($\boldsymbol{Y} \in$ $\psi^1_{\nu|\boldsymbol{Y}}$) =$\frac{1}{4\pi}[2\pi - cos^{-1}\rho_{12} - cos^{-1}\rho_{13} - cos^{-1}\rho_{23}]$. For category (2), $LR^* \sim \chi^2_2$ for $\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}}$, $i = 2,3,4$ with the corresponding mixture probability calculated by $w_2 = \sum_{j=2}^{4} Pr(\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}})$=$\frac{1}{4\pi}[3\pi - cos^{-1}\rho_{12|3} - cos^{-1}\rho_{13|2} - cos^{-1}\rho_{23|1}]$.

Correspondingly, $LR^* \sim \chi^2_1$ for $\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}}$, $i = 5, 6, 7$ with the corresponding mixture probability calculated as $w_1 = \sum_{j=5}^{7} Pr(\boldsymbol{Y} \in \psi^i_{\nu|\boldsymbol{Y}}) = \frac{1}{2} - w_3$ in category (3). For the last category, $LR^* \sim \chi^2_0$ for $\boldsymbol{Y} \in \psi^8_{\nu|\boldsymbol{Y}}$ with the mixture probability $w_0 = Pr(\boldsymbol{Y} \in \psi^8_{\nu|\boldsymbol{Y}}) = \frac{1}{2} - w_2$.

Chapter 3

Bivariate quantitative trait linkage analysis in mapping imprinted quantitative trait loci underlying endosperm traits in flowering plant

3.1 Introduction

The availability of multiple phenotypic traits allows researchers to associate genetic effects with the joint information of multivariate traits. Comparing with a univariate phenotypic trait, multivariate traits can provide more information in explaining the variation resulted from few particular genes or QTL, especially when correlations of these traits are observed to measure the related levels of multivariate phenotypes. By

accurately modeling the relative correlations of different phenotypes, the multivariate traits analysis significantly improves the power to detect QTL, and the degree of accuracy of position estimation for true QTL (Williams et al. 1999; Almasy et al. 1997).

Initiated by the double fertilization, a unique reproductive process in angiosperm plants, endosperm is developed from the fuse of the two polar nuclei and a sperm cell, ended up with a triploid tissue with two identical chromosomes inherited from maternal and one chromosome from paternal parent. Surrounding the embryo, the endosperm supplies main nutrition to the embryo (Brink and Cooper 1947). In cereal, it serves as the major source of food for human being. But the function of the endosperm is far more complicated and is beyond simple nutrient delivery to the embryo. Meanwhile, it is frequent to observe multivariate endosperm traits in maize, for instance, two highly correlated maize endosperm traits were collected: endoreduplication and mean ploidy (Cintia et al. 2006). To reveal the association of genetic effects with the variation of correlated endosperm traits, the multivariate traits analysis provides an essential channel to extract the maximum information to identify important genes or QTL.

In terms of the association of gene expression with variations of the phenotype, genomic imprinting is defined as the epigenetic phenomena that cause uniparental gene expression (Wolf et al. 2008). Under genomic imprinting, the expression of the same allele A from different heterozygote genotypes Aa and aA depends on the origin of inheritance of this allele. Then the maternally derived allele A (from Aa) functions

differently from that of paternally derived allele A (inherited from aA). A number of studies have illustrated that many endosperm traits are controlled by genomic imprinting. In maize, several paternal imprinting genes have been identified: the r gene in the regulation of anthocyanin (Kermicle 1970), the seed storage protein regulatory gene *dsrl* (Chaudhuri and Messing 1994), the *MEA* gene in seed development (Kinoshita et al. 1999) and some *α-tubulin* genes (Lund et al 1995). In the contrary direction, endoreduplication expresses a maternally controlled parent-of-origin effect (Dilkes et al. 2002). Endoreduplication, commonly occurring in angiosperms, is crucial for the endosperm development. By amplifying the genome to result in larger cells, endoreduplicaiton plays a critical role in process about the terminal differentiation and specialized function of given tissues. However, to our best knowledge, no study has been conducted to map iQTL with multivariate traits underlying potential imprinting process. It is the purpose of this study to develop an efficient multivariate iQTL mapping procedure incorporating the nature of the imprinting characteristic.

One important merit of the multivariate traits analysis is to make a number of biologically interesting hypotheses tests, such as testing pleiotropic effects of (i)QTL, testing pleiotropic against close linkage. These tests can not be accomplished under a univariate trait analysis. Generally, one phenotype is affected by one gene, but in a few cases, the same gene may govern several phenotypes simultaneously. This unique phenomenon is termed as pleiotropy. In real experiments, this special event may be confused with close linkage, another exceptional phenomenon during which variations of several phenotypic traits are influenced by multiple closely linked genes. Although

several genes are located closely in close linkage event, each gene only controls one trait. The discrepancy between pleiotropy and close linkage is simply distinguished by the number of traits one gene could control. It is practically important in distinguishing these two phenomena.

In maize, some vital genes displaying pleiotropic effects are revealed. For example, maize zfl regulatory genes in genetic backgrounds have pleiotripic effects on structure traits in branching and inflorescence (Bomblies and Doebley 2006); the tb1 gene with the intergenic sequences illustrates the pleiotropic effects on maize morphology (Clark et al. 2006); the early phase change (epc) gene has effects on maize development in several aspects (Vega et al. 2002); a maize gene GLOSSY1 (GL1) expresses its effects on trichome size and cutin structure during epidermis development (Sturaro et al. 2005); encoding with a transcription factor, a maize gene Glossy15 (Gl15) functioning like APETALA2 gene controls the development of ovule and identity of floral organ (Moose and Sisco 1996). It is known that endoreduplication and mean ploidy are two highly correlated endosperm traits in maize (Cintia et al. 2006). The identification of genes with pleiotropic effect based on these two phenotypic traits is practically important.

Variance components model is a powerful tool in multi-trait linkage analysis for an outbred or human population (Almasy et al. 1997; Williams et al. 1999). Due to the special inbreeding structure and unique genetic make-up of the endosperm genome, the current multi-trait linkage analysis methods can not be applied directly to endosperm phenotypes. In an extension to our previously proposed variance components

model in mapping iQTL underlying endosperm trait, in this work we will propose a bivariate iQTL mapping method to track down iQTL with possible pleiotropic effect, and to further distinguish potential close linkage signals.

3.2 Statistical method

3.2.1 The model

We will follow the same genetic design as illustrated in Chapter 2. Let $\boldsymbol{y}_{1_k} = (y_{1_1}, ..., y_{1_{n_k}})^T$ be a vector of the 1st phenotypic trait value in the kth family, and $\boldsymbol{y}_{2_k} = (y_{2_1}, ..., y_{2_{n_k}})^T$ be the 2nd phenotype within the same family. We assume multivariate normality for the joint distribution of $\boldsymbol{y}_{1_k}$ and $\boldsymbol{y}_{2_k}$. In the kth family, n_k individuals are randomly selected for each quantitative trait. The total K families are collected through four distinct reciprocal backcross populations. In bivariate trait analysis, the genotype-specific cytoplasmic maternal effect ($\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$), additive genetic effect at the QTL (a_{1_k}, a_{2_k}), polygenic additive effect (g_{1_k}, g_{2_k}), and random environmental effect (e_{1_k}, e_{2_k}) are considered. The parent-of-origin effect is further partitioned into effects due to the expression of the maternal allele with respect to each phenotypic trait (denoted as $a_{1_{mk}}$, $a_{2_{mk}}$), and due to the expression of the paternal allele (denoted as $a_{1_{kf}}$, $a_{2_{kf}}$). Hence, the genetic model underlying bivariate endosperm traits is represented in a vector form:

$$(y_{1_k}, y_{2_k}) = X_k^*(\beta_1, \beta_2) + 2(a_{1_{mk}}, a_{2_{mk}}) + (a_{1_{fk}}, a_{2_{fk}}) + (g_{1_k}, g_{2_k}) + (e_{1_k}, e_{2_k}) \tag{3.2.1}$$

where $k = 1, \cdots, K$. According to the reciprocal backcross design, three maternal genotypes AA, Aa and aa are observed, thus β_1 and β_2 denote mean parameters of two phenotypic traits with respect to three maternal genotypes, i.e., $\beta_1 = (\mu_1, \mu_2, \mu_3)$, $\beta_2 = (\mu_4, \mu_5, \mu_6)$. The design matrix X_k^* is an $n_k \times 3$ matrix with one column of ones and two columns of zeros. The random effects corresponding to the 1st trait are $a_{1_{mk}}$, $a_{1_{fk}}$, g_{1_k} and e_{1_k}, and each of these random components is distributed as normal distribution i.e., $a_{1_{mk}} \sim N(0, \Pi_{m|k}\sigma^2_{m_1})$, $a_{1_{fk}} \sim N(0, \Pi_{f|k}\sigma^2_{f_1})$, $g_{1_k} \sim N(0, \Phi_k\sigma^2_{g_1})$ and $e_{1_k} \sim N(0, I_k\sigma^2_{e_1})$, where $\sigma^2_{m_1}$ and $\sigma^2_{f_1}$ are the additive genetic variances at the QTL for maternal and paternal sides respectively; $\Pi_{m|k}$ and $\Pi_{f|k}$ are IBD sharing matrices that are derived from the similarities of maternal and paternal alleles among siblings, respectively; $\sigma^2_{g_1}$ and $\sigma^2_{e_1}$ are the additive polygenic variance and the residual environmental variance, respectively; $\mathbf{\Phi_k}$ is the expected proportion of alleles shared IBD; and $\mathbf{I_k}$ is the identity matrix. Correspondingly, $a_{2_{mk}}$, $a_{2_{fk}}$, g_{2_k} and e_{2_k} are random effects with normal distribution for the 2nd phenotypic trait i.e., $a_{2_{mk}} \sim N(0, \Pi_{m|k}\sigma^2_{m_2})$, $a_{2_{fk}} \sim N(0, \Pi_{f|k}\sigma^2_{f_2})$, $g_{2_k} \sim N(0, \Phi_k\sigma^2_{g_2})$ and $e_{2_k} \sim N(0, I_k\sigma^2_{e_2})$.

In addition, when bivariate phenotypic traits are involved in the model, the covariances of two phenotypes are expressed in terms of each random effect, i.e.,

$\text{Cov}(a1_{mk}, a2_{mk}) = \Pi_{m|k}\sigma m_{12}$ together with $\text{Cov}(a1_{fk}, a2_{fk}) = \Pi_{f|k}\sigma f_{12}$ are the covariances of the additive genetic effects at QTL; the covariance of the polygenic effects is $\text{Cov}(g1_k, g2_k) = \Phi_k \sigma g_{12}$; the covariance of the environmental effects is $\text{Cov}(e1_k, e2_k) = I_k \sigma e_{12}$.

3.2.2 Parent-specific allele sharing & genomewide linkage scan

The variance components model is built upon the basis of IBD sharing at the QTL. In triploid inbreeding population, a unique decomposition of parent-specific allele sharing pattern is illustrated in Figure 2.1. In the kth backcross family, the phenotypic variance-covariance corresponding to the 1st phenotype is denoted as: $\mathbf{\Sigma_{1k}} = \mathbf{\Pi}_{m|k}\sigma^2_{m_1} + \mathbf{\Pi}_{m/f|k}\sigma^2_{mf_1} + \mathbf{\Pi}_{f|k}\sigma^2_{f_1} + \mathbf{\Phi}_{g|k}\sigma^2_{g_1} + \mathbf{I}\sigma^2_{e_1}$, where $\mathbf{\Pi}_{m/f|k}$ is the IBD sharing matrix that the shared alleles are derived from different parents. Similarly, the phenotypic variance-covariance for the 2nd phenotype is given as $\mathbf{\Sigma_{2k}} = \mathbf{\Pi}_{m|k}\sigma^2_{m_2} + \mathbf{\Pi}_{m/f|k}\sigma^2_{mf_2} + \mathbf{\Pi}_{f|k}\sigma^2_{f_2} + \mathbf{\Phi}_{g|k}\sigma^2_{g_2} + \mathbf{I}\sigma^2_{e_2}$. The covariance of two phenotypic traits is expressed as $\mathbf{\Sigma_{12k}} = \mathbf{\Pi}_{m|k}\sigma m_{12} + \mathbf{\Pi}_{m/f|k}\sigma mf_{12} + \mathbf{\Pi}_{f|k}\sigma f_{12} + \mathbf{\Phi}_{g|k}\sigma g_{12} + \mathbf{I}\sigma e_{12}$. Therefore, the phenotypic variance-covariance of two phenotypic traits within the kth backcross family is expressed in a matrix form:

$$\mathbf{\Sigma_k} = \begin{pmatrix} \mathbf{\Sigma_{1k}} & \mathbf{\Sigma_{12k}} \\ \mathbf{\Sigma_{12k}} & \mathbf{\Sigma_{2k}} \end{pmatrix} \tag{3.2.2}$$

Where

- $\mathbf{\Sigma_{1k}} = \mathbf{\Pi}_{m|k}\sigma^2_{m_1} + \mathbf{\Pi}_{m/f|k}\sigma^2_{mf_1} + \mathbf{\Pi}_{f|k}\sigma^2_{f_1} + \mathbf{\Phi}_{g|k}\sigma^2_{g_1} + \mathbf{I}\sigma^2_{e_1}$

- $\mathbf{\Sigma_{12k}} = \mathbf{\Pi}_{m|k}\sigma_{m12} + \mathbf{\Pi}_{m/f|k}\sigma_{mf12} + \mathbf{\Pi}_{f|k}\sigma_{f12} + \mathbf{\Phi}_{g|k}\sigma_{g12} + \mathbf{I}\sigma_{e12}$

- $\mathbf{\Sigma_{2k}} = \mathbf{\Pi}_{m|k}\sigma^2_{m_2} + \mathbf{\Pi}_{m/f|k}\sigma^2_{mf_2} + \mathbf{\Pi}_{f|k}\sigma^2_{f_2} + \mathbf{\Phi}_{g|k}\sigma^2_{g_2} + \mathbf{I}\sigma^2_{e_2}$

The calculation of the IBD sharing probability is based on the marker positions. Unless each marker interval is dense, the QTL may be anywhere in the interval bracketed by two flanking markers. To acquire the accurate position of QTL, we need to search the putative QTL at every 1 or 2 cM throughout the entire genome (see Chapter 2 for more details).

3.2.3 Likelihood function and parameter estimation

In the kth family, two phenotype vectors are expressed as $\boldsymbol{y}_{1_k} = (y_{1_1}, ..., y_{1_{n_k}})^T$ and $\boldsymbol{y}_{2_k} = (y_{2_1}, ..., y_{2_{n_k}})^T$. Let $\boldsymbol{y}_k$=$(y_{1_1}, ..., y_{1_{n_k}}, y_{2_1}, ..., y_{2_{n_k}})^T$. Assuming the multivariate normality of $\boldsymbol{y}_k$ and different families are independent, the overall log likelihood function is given by:

$$\ell = \sum_{k=1}^{K} \log[f(\mathbf{y}_k; \boldsymbol{\beta}, \mathbf{\Sigma}_k)] \tag{3.2.3}$$

where $\boldsymbol{\beta}$ is a mean vector of both phenotypes in terms of $\boldsymbol{\beta}_1$ (denoted as the mean vector of the 1st phenotype) and $\boldsymbol{\beta}_2$ (as the mean vector for the 2nd phenotype) i.e., $\boldsymbol{\beta}_{(6\times1)} = \binom{\boldsymbol{\beta}_1}{\boldsymbol{\beta}_2}$. With respect to three maternal genotypes, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ include three different mean values for each trait. The random parameters in $\mathbf{\Sigma}_k$ are of main interest and defined as $\Theta = (\sigma^2_{m_1}, \sigma_{m12}, \sigma^2_{m_2}, \sigma^2_{f_1}, \sigma_{f12}, \sigma^2_{f_2}, \sigma^2_{mf_1}, \sigma_{mf12}, \sigma^2_{mf_2}, \sigma^2_{g_1}, \sigma_{g12}, \sigma^2_{g_2}, \sigma^2_{e_1}, \sigma_{e12}, \sigma^2_{e_2})$. To estimate these parameters, two commonly used

approaches are applied, the maximum likelihood (ML) method and the restricted maximum likelihood (REML) method.

3.2.3.1 The ML estimation

Defined the parameters as $\Omega = (\boldsymbol{\beta}, \boldsymbol{\Theta})$ where $\boldsymbol{\beta}$=(μ_1, μ_2, μ_3, μ_4, μ_5, μ_6) and $\boldsymbol{\Theta}$=($\sigma^2_{m_1}$, $\sigma_{m_{12}}$, $\sigma^2_{m_2}$, $\sigma^2_{f_1}$, $\sigma_{f_{12}}$, $\sigma^2_{f_2}$, $\sigma^2_{mf_1}$, $\sigma_{mf_{12}}$, $\sigma^2_{mf_2}$, $\sigma^2_{g_1}$, $\sigma_{g_{12}}$, $\sigma^2_{g_2}$, $\sigma^2_{e_1}$, $\sigma_{e_{12}}$, $\sigma^2_{e_2}$). The log-likelihood function to be maximized is in the form:

$$\ell^*(\Omega) = \sum_{k=1}^{K} \log[f(\mathbf{y}_k|\Omega)] = -\frac{1}{2}\sum_{k=1}^{K}\left\{\log|\boldsymbol{\Sigma}_k| + (\mathbf{y}_k - X_k\beta)'\Sigma_k(\mathbf{y}_k - X_k\beta)\right\} \tag{3.2.4}$$

where $\boldsymbol{y}_k$=($\boldsymbol{y}_{1_k}$, $\boldsymbol{y}_{2_k}$)^T is the phenotypic vector for both phenotypes; Σ_k is the variance-covariance matrix of $\mathbf{y}_k$ with dimension $2n_k \times 2n_k$, and the elements of this matrix are explained in section 3.2.3; the mean effect of the kth backcross family is denoted as $X_k\beta = \begin{pmatrix} \mu_{k|1}\mathbf{1}_{n_k} \\ \mu_{k|2}\mathbf{1}_{n_k} \end{pmatrix}$, and X_k is a design matrix.

We applied the Fisher scoring algorithm to estimate the parameters given in Ω,

$$\Theta^{(t+1)} = \Theta^{(t)} + \mathcal{I}(\Theta^{(t)})^{-1}\frac{\partial\ell^*(\Theta)}{\partial\Theta}|\Theta^{(t)}$$

Define

$$\Pi^{(1)}_{m|k} = \begin{pmatrix} \Pi_{m|k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{matrix } \mathbf{0} \text{ with dimension } n_k \times n_k,\ k = 1, ..., K$$

$$\Pi^{(2)}_{m|k} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Pi_{m|k} \end{pmatrix} \quad \text{matrix } \mathbf{0} \text{ with dimension } n_k \times n_k,\ k = 1, ..., K \qquad (3.2.5)$$

$$\Pi^{(3)}_{m|k} = \begin{pmatrix} \mathbf{0} & \Pi_{m|k} \\ \Pi_{m|k} & \mathbf{0} \end{pmatrix} \quad \text{matrix } \mathbf{0} \text{ with dimension } n_k \times n_k,\ k = 1, ..., K$$

Replacing the matrix $\Pi_{m|k}$ in the above equation (3.2.5) by $\Pi_{f|k}$, $\Pi_{mf|k}$, Φ_k, and I_k individually, we will obtain matrices $\Pi^{(s)}_{f|k}$, $\Pi^{(s)}_{mf|k}$, $\Phi^{(s)}_k$ and $I^{(s)}_k$, s=1,2,3, k=1,...,K.

The first-derivative of the log-likelihood function ℓ^* with respective to each variance component is given by

$$\frac{\partial \ell^*}{\partial \sigma^2_{m_1}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(1)}_{m|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(1)}_{m|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{m_2}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(2)}_{m|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(2)}_{m|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma_{m_{12}}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(3)}_{m|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(3)}_{m|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{f_1}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(1)}_{f|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(1)}_{f|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{f_2}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(2)}_{f|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(2)}_{f|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma_{f_{12}}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(3)}_{f|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(3)}_{f|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{mf_1}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(1)}_{mf|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(1)}_{mf|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{mf_2}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(2)}_{mf|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(2)}_{mf|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma_{mf_{12}}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Pi^{(3)}_{mf|k}) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Pi^{(3)}_{mf|k}\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{g_1}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Phi^{(1)}_k) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Phi^{(1)}_k\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{g_2}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Phi^{(2)}_k) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Phi^{(2)}_k\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma_{g_{12}}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}\Phi^{(3)}_k) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}\Phi^{(3)}_k\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{e_1}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}I^{(1)}_k) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}I^{(1)}_k\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{e_2}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}I^{(2)}_k) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}I^{(2)}_k\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

$$\frac{\partial \ell^*}{\partial \sigma_{e_{12}}} = -\frac{1}{2}\sum_{k=1}^{K}(tr(\hat{\Sigma}_k^{-1}I^{(3)}_k) - (\mathbf{y}_k - X_k\hat{\beta})^T\hat{\Sigma}_k^{-1}I^{(3)}_k\hat{\Sigma}_k^{-1}(\mathbf{y}_k - X_k\hat{\beta})),$$

Taking the expectation of the negative 2nd derivative of log-likelihood function with respect to $\Theta^{(t)}$, we obtain the Fisher information matrix ($\mathcal{I}(\Theta^{(t)})$).

Taking the 1st derivative of log-likelihood function with respect to $\boldsymbol{\beta}$, the maximum likelihood estimation of $\boldsymbol{\beta}$ is written as,

$$\hat{\boldsymbol{\beta}} = \sum_{k=1}^{K} (X_k^T \hat{\Sigma}_k^{-1} X_k)^{-1} X_k^T \hat{\Sigma}_k^{-1} Y_k$$

3.2.3.2 The REML Estimation

Comparing with the performance of maximum likelihood estimators (MLE), the REML approach reduces the biases of the parameters. The log-likelihood function to be maximized is given by

$$\ell^*(\Theta) = \log[f(\mathbf{y}|\Theta)] = -\frac{1}{2}\left\{\log|\boldsymbol{\Sigma}| + \log(|X'\boldsymbol{\Sigma}^{-1}X|) + \mathbf{y}'P\mathbf{y}\right\} \qquad (3.2.6)$$

where $\mathbf{y}$ is the phenotypic vector for both phenotypes with dimension $N \times 1$ ($N = 2 * \sum_{k=1}^{K} n_k$); Σ is the variance-covariance matrix of $\mathbf{y}$ with dimension $N \times N$ and is composed of Σ_k (k=1,...,K); P is a matrix denoted as $P = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}X(X'\boldsymbol{\Sigma}^{-1}X)^{-1}X'\boldsymbol{\Sigma}^{-1}$. X is an design matrix with dimension $N \times 6$ and consists of all X_k (k=1,...,K).

The vector $\mathbf{y}$ can be decomposed into three vectors with respect to three maternal genotypes (i.e., $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)^T$). The overall log-likelihood function is re-expressed as,

$$\ell^*(\Theta) = \sum_{r=1}^{3} \log[f(\mathbf{y_r}|\Theta)] = -\frac{1}{2}\sum_{r=1}^{3}\left\{\log|\mathbf{\Sigma_r}| + \log(|X_r'\mathbf{\Sigma_r}^{-1}X_r|) + \mathbf{y_r}'P_r\mathbf{y_r}\right\} \tag{3.2.7}$$

For r=1, the vector $\mathbf{y}_1$ is distributed as multivariate normal

$$\mathrm{N}\left(\begin{pmatrix}\mu_1\mathbf{1}_{\sum_{k=1}^{l_1} n_k}\\ \mu_4\mathbf{1}_{\sum_{k=1}^{l_1} n_k}\end{pmatrix}, \Sigma_1 = \begin{pmatrix}\Sigma_1 & & \\ & \cdots & \\ & & \Sigma_{l_1}\end{pmatrix}\right).$$

$$\text{For r=2, } \mathbf{y}_2 \sim \mathrm{N}\left(\begin{pmatrix}\mu_2\mathbf{1}_{\sum_{k=l_1+1}^{l_1+l_2} n_k}\\ \mu_5\mathbf{1}_{\sum_{k=l_1+1}^{l_1+l_2} n_k}\end{pmatrix}, \Sigma_2 = \begin{pmatrix}\Sigma_{l_1+1} & & \\ & \cdots & \\ & & \Sigma_{l_1+l_2}\end{pmatrix}\right).$$

$$\text{And } \mathbf{y}_3 \sim \mathrm{N}\left(\begin{pmatrix}\mu_3\mathbf{1}_{\sum_{k=l_1+l_2+1}^{K} n_k}\\ \mu_6\mathbf{1}_{\sum_{k=l_1+l_2+1}^{l_1+l_2+l_3} n_k}\end{pmatrix}, \Sigma_3 = \begin{pmatrix}\Sigma_{l_1+l_2+1} & & \\ & \cdots & \\ & & \Sigma_K\end{pmatrix}\right)$$

where $l_1 + l_2 + l_3 = K$. Note that Σ_1, Σ_2, and Σ_3 are also block diagonal matrices individually, P_r (r=1,2,3) is defined as $P_r = \mathbf{\Sigma}_r^{-1} - \mathbf{\Sigma}_r^{-1}X_r(X_r'\mathbf{\Sigma}_r^{-1}X_r)^{-1}X_r'\mathbf{\Sigma}_r^{-1}$. With this combination, we develop the following REML estimation procedure using the Fisher scoring algorithm.

Define

$$\Pi^{(1)}_{m|r} = \begin{pmatrix} \Pi_{m|r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{matrix } \mathbf{0} \text{ with dimension } \sum_{k=1}^{l_1} n_k \times \sum_{k=1}^{l_1} n_k$$

$$\Pi^{(2)}_{m|r} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Pi_{m|r} \end{pmatrix} \quad \text{matrix } \mathbf{0} \text{ with dimension } \sum_{k=l_1+1}^{l_1+l_2} n_k \times \sum_{k=l_1+1}^{l_1+l_2} n_k$$

$$\Pi^{(3)}_{m|r} = \begin{pmatrix} \mathbf{0} & \Pi_{m|r} \\ \Pi_{m|r} & \mathbf{0} \end{pmatrix} \quad \text{matrix } \mathbf{0} \text{ with dimension } \sum_{k=l_1+l_2+1}^{l_1+l_2+l_3} n_k \times \sum_{k=l_1+l_2+1}^{l_1+l_2+l_3} n_k \tag{3.2.8}$$

Replacing the matrix $\Pi_{m|r}$ in the above equation (3.2.8) by $\Pi_{f|r}$, $\Pi_{mf|r}$, Φ_r, and I_r individually, we obtain matrices $\Pi^{(s)}_{f|r}$, $\Pi^{(s)}_{mf|r}$, $\Phi^{(s)}_r$ and $I^{(s)}_r$ (s,r=1, 2, 3; $l_1 + l_2 + l_3$=K).

The 1st derivative of the log-likelihood function ℓ^* (3.2.7) with respective to each variance component is given by

$$\frac{\partial \ell^*}{\partial \sigma^2_{m_1}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(1)}_{m|r} - \mathbf{y}_r^T\hat{P}_r\Pi^{(1)}_{m|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{m_2}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(2)}_{m|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(2)}_{m|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma_{m_{12}}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(3)}_{m|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(3)}_{m|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{f_1}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(1)}_{f|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(1)}_{f|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{f_2}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(2)}_{f|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(2)}_{f|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma_{f_{12}}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(3)}_{f|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(3)}_{f|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{mf_1}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(1)}_{mf|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(1)}_{mf|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{mf_2}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(2)}_{mf|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(2)}_{mf|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma_{mf_{12}}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Pi^{(3)}_{mf|r}) - \mathbf{y}_r^T\hat{P}_r\Pi^{(3)}_{mf|r}\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{g_1}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Phi^{(1)}_r) - \mathbf{y}_r^T\hat{P}_r\Phi^{(1)}_r\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{g_2}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Phi^{(2)}_r) - \mathbf{y}_r^T\hat{P}_r\Phi^{(2)}_r\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma_{g_{12}}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r\Phi^{(3)}_r) - \mathbf{y}_r^T\hat{P}_r\Phi^{(3)}_r\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{e_1}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r I^{(1)}_r) - \mathbf{y}_r^T\hat{P}_r I^{(1)}_r\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma^2_{e_2}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r I^{(2)}_r) - \mathbf{y}_r^T\hat{P}_r I^{(2)}_r\hat{P}_r\mathbf{y}_r),$$

$$\frac{\partial \ell^*}{\partial \sigma_{e_{12}}} = -\frac{1}{2}\sum_{r=1}^{3}(tr(\hat{P}_r I^{(3)}_r) - \mathbf{y}_r^T\hat{P}_r I^{(3)}_r\hat{P}_r\mathbf{y}_r),$$

The Fisher information matrix ($\mathcal{I}(\Theta^{(t)})$) in the REML procedure is obtained by taking the expectation of the negative 2nd derivative of log-likelihood function with respect to each variance component.

The REML estimator of $\boldsymbol{\beta}$ is the generalized least squares estimator, that is,

$$\hat{\boldsymbol{\beta}} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} Y$$

3.2.4 Hypothesis testing

In the bivariate traits analysis, the existence of quantitative trait loci (QTL) is tested by the following hypothesis

$$\begin{cases} H_0 : \sigma^2_{m_1} = \sigma^2_{m_2} = \sigma^2_{m_{12}} = \sigma^2_{f_1} = \sigma^2_{f_2} = \sigma^2_{f_{12}} = \sigma^2_{mf_1} = \sigma^2_{mf_2} = \sigma^2_{mf_{12}} = 0 \\ H_1 : \ H_0 \text{ is not true} \end{cases} \tag{3.2.9}$$

The significance of the above test is assessed through the likelihood ratio test (LRT). Let $\widetilde{\Omega}$ and $\widehat{\Omega}$ be estimates of the unknown parameters with respect to H_0 and H_1, respectively, then the likelihood ratio statistic is evaluated by,

$$\text{LR} = -2[\log L(\widetilde{\Omega}|\boldsymbol{y}) - \log L(\widehat{\Omega}|\boldsymbol{y})] \tag{3.2.10}$$

which, under the null hypothesis, is distributed with a mixture chi-square distribution with the form $\frac{\binom{6}{0}}{2^6}\chi^2_9 : \frac{\binom{6}{1}}{2^6}\chi^2_7 : \frac{\frac{1}{5}\binom{6}{2}}{2^6}\chi^2_6 : \frac{\frac{4}{5}\binom{6}{2}}{2^6}\chi^2_5 : \frac{\frac{3}{5}\binom{6}{3}}{2^6}\chi^2_4 : \frac{\frac{2}{5}\binom{6}{3}+\frac{1}{5}\binom{6}{4}}{2^6}\chi^2_3 : \frac{\frac{4}{5}\binom{6}{4}}{2^6}\chi^2_2 : \frac{\binom{6}{5}}{2^6}\chi^2_1 : \frac{\binom{6}{6}}{2^6}\chi^2_0$.

Once a QTL is identified at a genomic position, its imprinting property for each phenotypic trait is assessed by the following two imprinting hypotheses formulated by,

$$\begin{cases} H_0 : \sigma^2_{f_1} = \sigma^2_{m_1} \\ H_1 : \sigma^2_{f_1} \neq \sigma^2_{m_1} \end{cases} \tag{3.2.11}$$

and

$$\begin{cases} H_0 : \sigma^2_{f_2} = \sigma^2_{m_2} \\ H_1 : \sigma^2_{f_2} \neq \sigma^2_{m_2} \end{cases}$$

Again, the likelihood ratio test is applied and the test statistic (denoted as LR_{imp}) follows a chi-square distribution with 1 degree of freedom. If the null is rejected at the tested QTL position, imprinting effect is claimed. We further assess whether the imprinted genetic effect is completely derived from the maternal allele or from the paternal allele. Two hypotheses are formulated for this purpose to assess completely maternally imprinting by

$$\begin{cases} H_0 : \sigma^2_{m_t} = 0 & t = 1,\ 2 \\ H_1 : \sigma^2_{m_t} \neq 0 & t = 1,\ 2 \end{cases} \tag{3.2.12}$$

and to assess completely paternally imprinting by

$$\begin{cases} H_0 : \sigma^2_{f_t} = 0 & t = 1,\ 2 \\ H_1 : \sigma^2_{f_t} \neq 0 & t = 1,\ 2 \end{cases}$$

The likelihood ratio test statistic (LR_{cimp}) corresponding to the above tests follows a mixture chi-square distribution with $\frac{1}{2}\chi_1^2 : \frac{1}{2}\chi_0^2$.

If the test in (3.2.9) is rejected, we can further test if the QTL controls the 1st trait by testing

$$\begin{cases} H_0 : \sigma_{m_1}^2 = \sigma_{f_1}^2 = \sigma_{mf_1}^2 = 0 \\ H_1 : \ H_0 \text{ is not true} \end{cases} \tag{3.2.13}$$

or the 2nd trait by testing

$$\begin{cases} H_0 : \sigma_{m_2}^2 = \sigma_{f_2}^2 = \sigma_{mf_2}^2 = 0 \\ H_1 : \ H_0 \text{ is not true} \end{cases}$$

The likelihood ratio statistic corresponding to the above tests is denoted as LR_{plei} and follows a mixture chi-square distribution under the null with the form $\frac{1}{4\pi}[2\pi - cos^{-1}\rho_{12} - cos^{-1}\rho_{13} - cos^{-1}\rho_{23}]\chi_3^2$: $\frac{1}{4\pi}[3\pi - cos^{-1}\rho_{12|3} - cos^{-1}\rho_{13|2} - cos^{-1}\rho_{23|1}]\chi_2^2$: $\frac{1}{4\pi}(cos^{-1}\rho_{12} + cos^{-1}\rho_{13} + cos^{-1}\rho_{23})\chi_1^2$: $[\frac{1}{2} - \frac{1}{4\pi}(3\pi - cos^{-1}\rho_{12|3} - cos^{-1}\rho_{13|2} - cos^{-1}\rho_{23|1})]\chi_0^2$ where the correlation between the variance terms a and b is calculated from the Fisher information matrix, and $\rho_{ab|c} = \frac{(\rho_{ab} - \rho_{ac}\rho_{bc})}{(1-\rho_{ac}^2)^{1/2}(1-\rho_{bc}^2)^{1/2}}$ (see Chapter 2 for details).

Rejecting the null for the above two tests indicates pleiotropic effect (i.e., one gene has effect on two traits). But if two genes are closely linked at the detected QTL (i.e., close linkage), we still get the same testing result. To further distinguish close linkage against pleiotropic effect, we develop the following two tests

$$\begin{cases} H_0 : \rho m_{12} = \rho f_{12} = \rho m f_{12} = 1 \\ H_1 : \ H_0 \ is \ not \ true \end{cases} \tag{3.2.14}$$

for testing pleiotropic effect and

$$\begin{cases} H_0 : \rho m_{12} = \rho f_{12} = \rho m f_{12} = 0 \\ H_1 : \ H_0 \ is \ not \ true \end{cases} \tag{3.2.15}$$

for testing close linkage. The null hypothesis in test (3.2.14) indicates that the additive effects for the two traits are perfectly correlated and they are possibly controlled by a single gene. On the contrary, the null hypothesis in test (3.2.15) indicates two closely linkage genes at one QTL location. The likelihood ratio test is denoted by LR_{p} for test (3.2.14) and LR_{c} for test (3.2.14). The null distribution of LR_{p} has a mixture chi-square distribution (since 1 is a boundary point for correlation ρ) with the form $\frac{1}{4\pi}[2\pi - cos^{-1}\rho_{12} - cos^{-1}\rho_{13} - cos^{-1}\rho_{23}]\chi^2_3$: $\frac{1}{4\pi}[3\pi - cos^{-1}\rho_{12|3} - cos^{-1}\rho_{13|2} - cos^{-1}\rho_{23|1}]\chi^2_2$: $\frac{1}{4\pi}(cos^{-1}\rho_{12}+cos^{-1}\rho_{13}+cos^{-1}\rho_{23})\chi^2_1$: $[\frac{1}{2}-\frac{1}{4\pi}(3\pi-cos^{-1}\rho_{12|3}-cos^{-1}\rho_{13|2} - cos^{-1}\rho_{23|1})]\chi^2_0$.

The null distribution of LR_{c} follows a classical chi-square distribution with 3 degrees of freedom, i.e., $LR_{\mathrm{c}} \sim \chi^2_3$.

We can also test the maternal main effect on each trait by

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \\ H_1 : \ H_0 \text{ is not true} \end{cases} \tag{3.2.16}$$

for the 1st trait, and

$$\begin{cases} H_0 : \mu_4 = \mu_5 = \mu_6 \\ H_1 : \ H_0 \text{ is not true} \end{cases}$$

for the 2nd trait.

3.3 Simulation

3.3.1 Simulation design

A simulation study was conducted to evaluate the performance of the proposed method. Five equally-spaced markers ($\mathcal{M}_1 - \mathcal{M}_5$) are simulated for one linkage group assuming a backcross design. This linkage group covers a length of 40cM with 10cM for each marker interval. Haldane map function is used to convert map distance to recombination rate. Assume one QTL is at 22cM away from the first marker, and has effects on two phenotypic traits. Phenotypic values of both traits are generated from a multivariate normal distribution with variance-covariance given in (3.2.2) in terms of different parameter settings. Backcross families are simulated following the structure of the real data described in Chapter 2 (i.e., 4 families with 100 sibs within each family). In each simulation scenario, the IBD value of any two siblings is evaluated at every 2cM along the linkage group. The REML method is adopted to estimate

unknown parameters of interests, and 100 simulation replicates are recorded.

3.3.2 Results

The simulated results without imprinting effect (i.e., $\sigma^2_{m_1} = \sigma^2_{f_1}$ and $\sigma^2_{m_2} = \sigma^2_{f_2}$) are tabulated in Table 3.1. Estimations of the QTL position, observed statistical power and type I errors are compared between bivariate traits analysis (T1+T2) and each univariate trait analysis (T1 and T2). Overall, the bivariate traits analysis gives more precise QTL position estimate, larger statistical power and reasonable type I error rate. The results indicate that the joint mapping incorporating bivariate phenotypic information can capture information of QTLs with small or moderate effects that could be easily missed by univariate trait analysis. In addition, the bivariate trait analysis provides less biased parameter estimates for additive QTL variance terms derived from maternal and paternal parents. For example, variance term for σ^2_{m1} is estimated as 0.428(SMSE=0.11) for the joint analysis, while it is 0.25 (SMSE=0.44) for the single trait analysis.

Table 3.1: The power, REMLs of the QTL position estimated based on 100 simulation replicates for a QTL with no imprinting effect under 4 × 100 design. The square roots of the mean squared errors of the parameter estimates are given in parentheses.

Quantitative traits	heritability H^2	Position 22cM	$\sigma^2_{m_1}$ 0.5	$\sigma^2_{f_1}$ 0.5	$\sigma^2_{mf_1}$ 0.25	$\sigma^2_{m_2}$ 1	$\sigma^2_{f_2}$ 1	$\sigma^2_{mf_2}$ 0.5	$\sigma^2_{g_1}$ 1.5	$\sigma^2_{g_2}$ 3	Power	Type I error
Trait 1 + Trait 2		21.58	0.428	0.447	0.44	0.732	0.758	0.833	2.192	5.03	0.86	0.09
		(9.89)	(0.11)	(0.15)	(0.68)	(0.42)	(0.63)	(1.15)	(1.17)	(6.20)		
Trait 1	0.217	19.66	0.250	0.451	0.708				1.626		0.47	
		(10.21)	(0.44)	(1.96)	(1.36)				(1.35)			
Trait 2	0.05	20.12				0.255	0.929	1.371		4.207	0.04	0.07
		(13.46)				(1.01)	(5.63)	(4.51)		(8.15)		

The locations of the QTL is described by the map distances (in cM) from the first marker of the linkage group (40 cM long); The true QTL is located at 22cM; Power is calculated using the empirical distribution through simulation.

Table 3.1 also shows that the results for trait 1 is better than trait 2, due to high heritability ($H^2 = 0.20$) for T1 than that ($H^2 = 0.05$) for T2. When the heritability for both traits are increased, we observe better performance (data not shown). The type I error for the bivariate traits analysis and single trait with T2 is reasonably controlled. The type I error for T1 is a little inflated. But the joint analysis gives much larger power than that for both single trait analysis.

To demonstrate the imprinting property of an iQTL in the bivariate traits analysis, additional simulations under different imprinting mechanisms are conducted and the results are listed in Table 3.2. We design four imprinting modes such as: complete paternal imprinting $(\begin{pmatrix} \sigma^2_{m_1}=1 \\ \sigma^2_{m_2}=2 \end{pmatrix}, \begin{pmatrix} \sigma^2_{f_1}=0 \\ \sigma^2_{f_2}=0 \end{pmatrix})$, complete maternal imprinting $(\begin{pmatrix} \sigma^2_{m_1}=0 \\ \sigma^2_{m_2}=0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{f_1}=1 \\ \sigma^2_{f_2}=2 \end{pmatrix})$, partial maternal imprinting $(\begin{pmatrix} \sigma^2_{m_1}=0.25 \\ \sigma^2_{m_2}=0.5 \end{pmatrix}, \begin{pmatrix} \sigma^2_{f_1}=0.75 \\ \sigma^2_{f_2}=1.5 \end{pmatrix})$, and partial paternal imprinting $(\begin{pmatrix} \sigma^2_{m_1}=0.75 \\ \sigma^2_{m_2}=1.5 \end{pmatrix}, \begin{pmatrix} \sigma^2_{f_1}=0.25 \\ \sigma^2_{f_2}=0.5 \end{pmatrix})$. With respect to the imprinting test (3.2.11), the largest imprinting power is achieved by complete maternal (paternal) imprinting for both phenotypic traits (T1 and T2). No significant difference in power (Power) under different imprinting mechanisms is observed for the joint analysis. The low power for the imprinting test for single trait T2 is due to its low heritability.

Table 3.2: The power, REMLs of the QTL position and effect parameters estimated based on 100 simulation replicates for a QTL showing different imprinting effects under the 4×100 sampling design. The square roots of the mean squared errors of the parameters are given in parentheses.

Position	$\sigma_{m_1}^2$	$\sigma_{f_1}^2$	$\sigma_{m_2}^2$	$\sigma_{f_2}^2$	$\sigma_{m_{12}}^2$	$\sigma_{f_{12}}^2$	$\sigma_{mf_1}^2$	$\sigma_{mf_{12}}^2$	$\sigma_{mf_2}^2$	$\sigma_{g_1}^2$	$\sigma_{g_{12}}^2$	$\sigma_{g_2}^2$	$\sigma_{e_1}^2$	$\sigma_{e_{12}}^2$	$\sigma_{e_2}^2$			
22cM							0.25	0.3	0.5	1.5	1.7	3	3	10	40	Power	iPower[1]	iPower[2]
	0	1	0	2														
19.38	0.082	0.515	0.108	0.909	0.047	0.503	0.321	0.369	0.812	1.586	1.989	5.467	3.362	9.226	35.966	0.74	0.71	0.35
(11.52)	(0.08)	(0.55)	(0.19)	(1.44)	(0.065)	(0.91)	(0.21)	(0.35)	(1.06)	(0.85)	(2.32)	(7.63)	(0.63)	(1.53)	(6.19)			
	1	0	2	0														
20.54	0.49	0.097	0.881	0.271	0.469	0.091	0.311	0.331	0.751	2.639	3.248	7.275	3.046	8.981	36.031	0.87	0.63	0.34
(8.18)	(0.55)	(0.11)	(1.19)	(0.62)	(0.85)	(0.18)	(0.39)	(0.51)	(0.95)	(1.71)	(3.21)	(8.74)	(0.46)	(1.62)	(5.99)			
	0.25	0.75	0.5	1.5														
19.82	0.116	0.505	0.178	0.845	0.089	0.485	0.337	0.397	0.874	1.785	2.32	6.013	3.282	9.251	36.272	0.9	0.64	0.31
(9.64)	(0.14)	(0.31)	(0.44)	(1.03)	(0.25)	(0.60)	(0.30)	(0.46)	(0.96)	(0.92)	(2.49)	(8.12)	(0.57)	(1.38)	(5.53)			
	0.75	0.25	1.5	0.5														
20.66	0.478	0.127	0.836	0.36	0.456	0.14	0.448	0.541	0.946	2.713	3.218	6.335	3.106	9.088	36.308	0.86	0.55	0.30
(8.32)	(0.29)	(0.14)	(0.73)	(0.55)	(0.53)	(0.25)	(0.57)	(0.80)	(1.28)	(1.64)	(2.83)	(7.55)	(0.45)	(1.43)	(5.55)			

Power refers to the overall QTL detection power; iPower[1], iPower[2] refer to the imprinting power for testing $H_0 : \sigma_m^2 = \sigma_f^2$ for trait T1 and T2, respectively. The imprinting test is only conducted at the position where hypothesis (2.2.6) is rejected.

3.4 Real Data Analysis

Empirical study shows that imprinted genes affect variations of maize endosperm traits (Dilkes et al. 2002). Two endosperm traits, mean ploidy level (denoted as ploidy) and percentage of endoreduplicated nuclei (denoted as endo), are studied. Four backcross segregation populations are generated from two inbred line (Sg18 and Mo17). The details of this genomic data were explained by Coelho et al (2007), and the imprinting effect analyzed with the univariate trait analysis was reported in Chapter 2. To examine the pleiotropic effect of the imprinted genes, we conducted a joint analysis.

LR profiles across ten linkage groups in bivariate traits analysis (endo+ploidy) and univariate trait analysis (endo and ploidy) are plotted in figure 3.1. The genome-wide significance threshold (horizontal dotted line, at 5% level) is determined by permutations based on repeatedly shuffling the relationship between marker genotypes and phenotypes (Churchill and Doerge 1994). Six QTL are detected at the 5% genome-wide significance level on G2, G4, G6, G9 and G10. In contrast with previous QTL detected in the univariate trait analysis (see Chapter 2), more QTL are detected in the bivariate joint analysis.

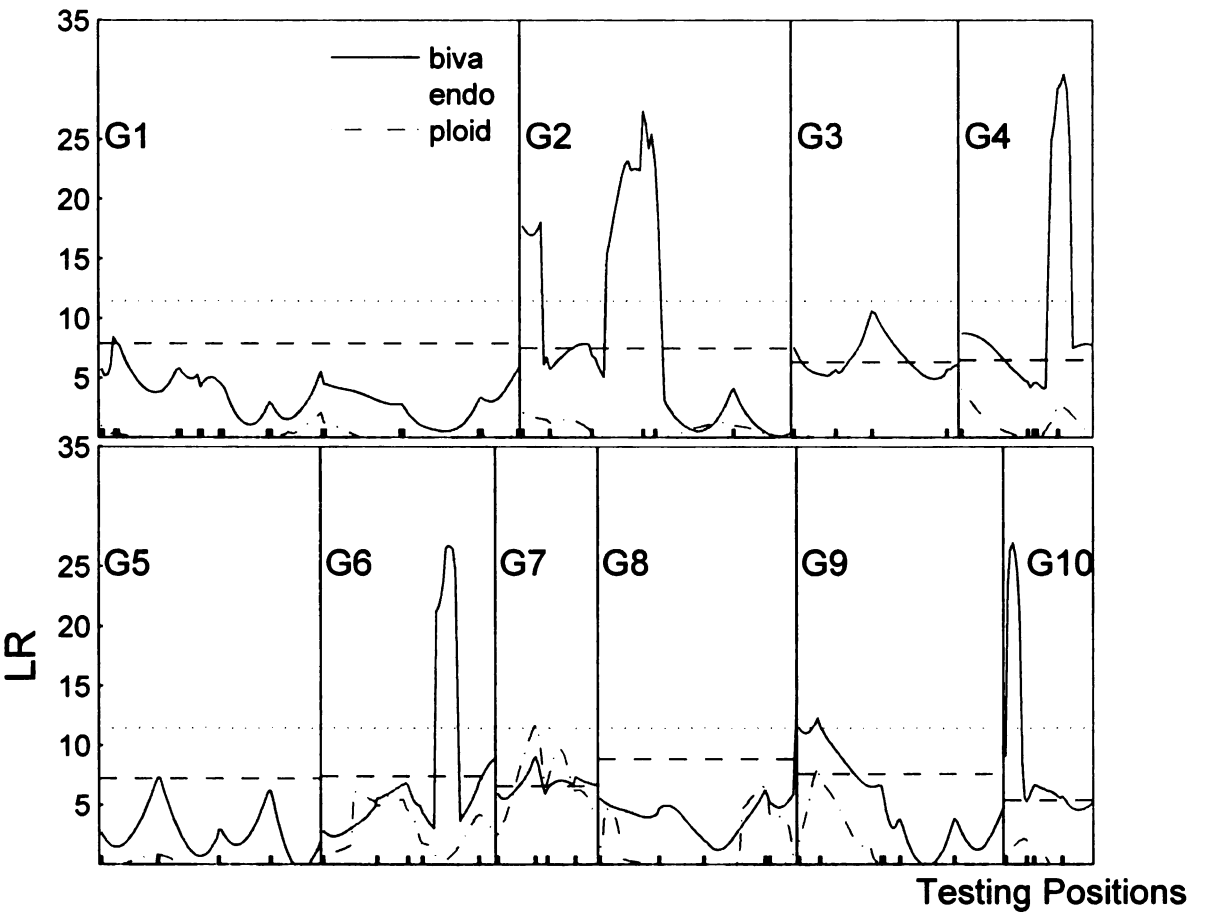

Figure 3.1: The profile of the log-likelihood ratios (LR) for testing the existence of QTLs underlying the two endosperm traits across the 10 maize linkage groups ($G_1, \cdots, G_{10}$).

Table 3.3: The estimated parameters for the variance components for joint bivariate endosperm traits: mean ploidy (Mploidy) and percent of the endoreduplicated nuclei (Endo).

Ch	Position	Genetic effects																	
		$\sigma^2_{m_1}$	$\sigma_{m_{12}}$	$\sigma^2_{m_2}$	$\sigma^2_{f_1}$	$\sigma_{f_{12}}$	$\sigma^2_{f_2}$	$\sigma^2_{mf_1}$	$\sigma_{mf_{12}}$	$\sigma^2_{mf_2}$	$\sigma^2_{g_1}$	$\sigma_{g_{12}}$	$\sigma^2_{g_2}$	$\sigma^2_{e_1}$	$\sigma_{e_{12}}$	$\sigma^2_{e_2}$	p_{im_1}	p_{im_2}	p_{if_1}	p_{if_2}	p_{plei}	p_{co-in}
2	umc-1.89cM	0.30	0.38	0.83	0.17	0.08	0.39	0.09	-0.14	0.22	1.26	0.30	4.15	2.81	8.91	36.77						
2	bmc-4.53cM	0.39	0.47	0.63	0.31	0.28	0.36	0.16	-0.41	1.08	0.71	-0.37	3.96	3.08	9.24	37.02						
4*	umc+3.62cM	0.57	1.06	2.27	0.30	0.62	1.30	0.02	0.15	1.62	0.82	-0.81	0.94	2.84	9.31	38.48					0.075	0.044
6*	bnlg+21.48cM	0.26	0.49	1.71	0.42	1.20	3.41	0.19	-0.56	4.46	0.88	-0.46	1.24	2.93	9.35	38.44			0.051	0.001	0.363	≈ 0
9	pOi	0.29	0.18	0.16	0.28	0.41	1.80	0.31	-0.06	0.01	1.01	0.01	3.90	3.02	9.20	37.19						
10*	MMC+6cM	0.33	0.56	0.95	0.29	0.65	1.45	0.03	-0.02	0.51	0.71	-0.78	1.78	2.96	9.50	38.70						

Note: Six QTLs for joint traits Mploidy and Endo are located at chromosome 2, 4, 6, 9 and 10. QTLs showing significance at the genomewide level are indicated by "$*$". p_{imp_t}, p_{im_t}, p_{if_t}, p_{plei} and p_{co_in} are the p-values for testing imprinting effect ($H_0 : \sigma^2_{m_t} = \sigma^2_{f_t}$), complete maternal imprinting ($H_0 : \sigma^2_{m_t} = 0$), complete paternal ($H_0 : \sigma^2_{f_t} = 0$) t=1, 2, pleiotropic effect ($H_0 : \rho_1 = \rho_2 = \rho_3 = 1$) and co-incidence linkage ($H_0 : \rho_1 = \rho_2 = \rho_3 = 0$) respectively.

Table 3.3 lists the QTL location, variance components estimation, and test outcomes for imprinting and pleiotropic effects. In the bivariate analysis, the identified QTL on G6 is imprinted for T2 ($p_{imp_2} < 0.05$) but not for T1 $p_{imp_1} > 0.05$). Further test shows that this QTL shows completely paternal imprinting on T2. From the parameter estimation, it is clear that correlations of genetic variance for two phenotypic traits are strong, and further tests to detect pleiotropic effects vs close linkage are meaningful in the bivariate traits analysis. Results show that two iQTLs on G4 and G6 indicate strong pleiotropic effects ($p_{plei} > 0.05$, $p_{co-in} < 0.05$).

In our study, multivariate approaches for genetic linkage analysis increase the power and precision to identify genetic effect, especially when a QTL has pleiotropic effect on several traits. In accordance with the finding about the strong correlation between endoreduplication and mean ploidy in maize endosperm (Cintia et al. 2006), the pleiotropic effects of iQTLs on two endosperm traits are detected in our analysis. As shown in the simulation study, the joint analysis provides larger power and resolution for iQTL detection compared to the single trait analysis, which explains the additional QTLs detected by the joint analysis.

3.5 Discussion

A number of studies have shown that for correlated traits, multivariate approaches for genetic linkage analysis can increase the power and precision to identify genetic effects (Evans 2002), especially when a QTL has the pleiotropic effect on several traits (Jiang and Zeng 1995). Considering the importance of imprinted genes in

endosperm development and the relative merit of multi-trait analysis, we developed a bivariate variance components model based on a reciprocal backcross design to identify imprinted QTLs while incorporating the special genetic make-up of triploid inbreeding population. Both simulation and real data analysis show the efficiency of the approach.

In simulation studies, we compared the outcomes of bivariate traits analysis with those of univariate trait analyses. The bivariate trait analysis can greatly improve the performance in QTL position estimation, testing power, and type I error rate. Simulation study also shows that when a trait has low heritability (i.e., T2), the joint analysis can also identify the gene by borrowing information from other traits with high heritability (i.e., T1) given that they are correlated. Our results are consistent with other multivariate traits studies (e.g., Jiang and Zeng 1995; Almasy et al. 1997; Williams et al. 1999).

We applied our joint model to a real data set with two highly correlated endosperm traits, i.e., endoreduplication and mean ploidy. Six QTLs are detected on G1, G2, G4, G6, G8, G10 across the genome. One maternally imprinted QTL on G6 for T2 and three paternally imprinted QTLs on G4, G6, G8 for T1 are also identified. The results of the imprinting tests are consistent with that of univariate trait analysis and can be explained by the genetic conflict theory proposed by Haig and Westoby (1991). Comparing with univariate trait analysis, additional QTLs are mapped in the bivariate joint analysis. These additional QTLs are those showing small genetic effect in the single trait analysis. This demonstrates the power of the joint analysis

for correlated traits.

Another advantage of the joint analysis is the test of pleiotropic effect and close linkage. We proposed a set of hypothesis tests to detect the existence of QTLs and genomic imprinting in bivariate traits analysis, and moreover, to distinguish the pleiotropic and close linkage effect. For the real data, one iQTL on G6 displays a strong pleiotropic effect, which controls both the endoreduplication trait and the mean ploidy trait (Table 3.3). Our method provides a testable framework in iQTL mapping with multivariate traits.

Chapter 4

Assessing statistical significance in genetic linkage analysis with the variance components model

4.1 Introduction

Variance components (VC) model is a powerful tool for quantitative trait loci (QTL) mapping in human linkage analysis. In a VC analysis, genetic effects are often partitioned as additive, dominance and polygene effects whereby each one is treated as random. Thus, we are interested in testing whether the variance of a genetic effect is significantly deviated from zero. Likelihood ratio test (LRT) is often applied for the the testing purpose. Due to irregular conditions (i.e., parameter boundary problem), the asymptotic distribution of the LRT does not follow a regular chi-square distribu-

tion, rather a mixture χ^2 distribution, where the mixture proportions are calculated with standard binomial coefficients, a special case in Self and Liang (1987).

A number of studies have showed the asymptotic distribution of LRT under irregular conditions, see for example, Self and Liang (1987), Chernoff (1954) and Shapiro (1988). Chernoff (1954) showed that the limiting distribution of the LRT has a mixture chi-square distribution when parameters of interest are on one side of a hyperplane, or in the first quadrant within a $\mathbb{R}^2$ space. Self and Liang (1987) extended the Chernoff's comment to boundary cases. In linkage analysis with variance components model, the result displayed in case 9 in Self and Liang (1987) has been commonly applied for a threshold determination (e.g., Amos 1994; Hanson et al. 2001). This result is based on the assumption of a diagonal variance-covariance matrix of unknown parameters. In reality, this assumption could be easily violated. This consequently leads to conservative hypothesis tests (e.g., Allison et al. 1999).

For a bivariate linkage analysis, Amos et al. (2001) proposed an approach to approximate the null distribution of the LRT (see section 4.2 for more details). But, their derivation assumes a diagonal Fisher information matrix. Moreover, they assume that the genetic correlation between two traits is either positively correlated ($\rho = 1$) or negatively correlated ($\rho = -1$). This is unrealistic in reality. Corresponding to the VC model, Morris et al. (2009) define the constrained likelihood ratio test (CLRT) with respect to this model. They try to apply Geyer's regularity (1994) to show the asymptotic distribution of the constrained CLRT, but can not make sure that the global M-maximizer is definitely attained. Because of this limitation, a new simulation

method is developed. However, it is quite difficult to express the predominance of this method comparing with others.

In this chapter we rigorously show that the LRT statistic in testing variance components in linkage analysis follows a mixture chi-square distribution and the mixture proportions depend on the estimated Fisher information matrix. The rest of this chapter is organized as follows. Section 4.2 introduce three classical VC models with both univariate and multivariate trait analysis. The main result is illustrated in section 4.3. Section 4.4 shows the performance of the new approximation by a few simulation examples.

4.2 Motivating models

4.2.1 Model I

Assume K families are collected and the phenotype for the k^{th} family is denoted by y_k with n_k offsprings. Under the variance components model mapping framework, the genetic effect is partitioned into several components expressed as

$$y_k = \mu + a_k + d_k + g_k + e_k \tag{4.2.1}$$

where μ is the overall mean; $a_k \sim N(0, \sigma_a^2)$ and $d_k \sim N(0, \sigma_d^2)$ are the additive and dominant effect of a genetic variant; $g_k \sim N(0, \sigma_g^2)$ is the polygenic effect (i.e., the effect of QTLs not located on the same chromosome as the tested one); and $e_k \sim N(0, \sigma_e^2)$ is the residual term. When a testing QTL is not on a marker position,

the variance-covariance of the phenotype for a pair of sibpairs y_{ki} and y_{kj} in the kth family can be expressed as

$$\text{cov}(y_{ki}, y_{kj}|\pi_{ij}, \varphi_{ij}) = \begin{cases} \sigma_a^2 + \sigma_d^2 + \sigma_g^2 + \sigma_e^2 & \text{if } i = j \\ b_{ij}(\theta, \pi_{ij})\sigma_a^2 + c_{ij}(\theta, \pi_{ij}, \varphi_{ij})\sigma_d^2 + \phi_{ij}\sigma_g^2 & \text{if } i \neq j \end{cases}$$

where π_{ij} is the proportion of marker alleles shared identical by descent (IBD) between two sibs; φ_{ij} is the probability of sharing two alleles IBD between any pair of sibs; ϕ_{ij} is the kinship coefficient; θ is the recombination fraction between a trait locus and a marker. When a trait locus is not at the marker, $b_{ij}(\theta, \pi_{ij}) = \frac{1}{2} + (1-2\theta)^2(\pi_{ij} - \frac{1}{2})$ and $c_{ij}(\theta, \pi_{ij}, \varphi_{ij}) = 4\theta^2(1-\theta)^2 + (1-2\theta)^2\pi_{ij} + (1-2\theta)^4\varphi_{ij}$ (Amos et al. 2001).

In matrix notation, the phenotypic variance-covariance matrix among individuals in family k can be expressed as

$$\mathbf{\Sigma_k} = \mathbf{\Pi_k}\sigma_a^2 + \mathbf{\Delta_k}\sigma_d^2 + \mathbf{\Phi_k}\sigma_g^2 + \mathbf{I_k}\sigma_e^2$$

where $\mathbf{\Pi}_k$ is the matrix of marker alleles shared IBD for the pedigree, and $\mathbf{\Delta}_k$ is the matrix of the proportion of marker alleles shared two alleles IBD in the pedigree, $\mathbf{\Phi}_k$ is a matrix of the expected proportion of alleles shared IBD, and $\mathbf{I}_k$ is an identity matrix.

The quantitative trait loci is tested by the genetic linkage test defined as

$$\begin{cases} H_0 : \sigma_a^2 = \sigma_d^2 = 0 \\ \\ H_1 : \text{at least one parameter is not zero.} \end{cases} \tag{4.2.2}$$

Define $\theta_1 = \mu, \theta_2 = \sigma_a^2, \theta_3 = \sigma_d^2, \theta_4 = \sigma_g^2$ and $\theta_5 = \sigma_e^2$. Let $\boldsymbol{\theta} = (\theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5)^T$ $\in \Omega = \mathbb{R} \times [0,\infty) \times [0,\infty) \times (0,\infty) \times (0,\infty)$ be the true parameter space. Under the null, the parameter space is reduced to $\theta_0 = (\theta_{10}\ \theta_{20}\ \theta_{30}\ \theta_{40}\ \theta_{50})^T =$ $(\mu_0\ 0\ 0\ \sigma_{g0}^2\ \sigma_{e0}^2)^T \in \Omega_0 = \mathbb{R} \times \{0\} \times \{0\} \times (0,\infty) \times (0,\infty)$. Thus two parameters under the null are on the boundary of the true parameter space (Ω). In current applications, the LRT statistic for the above test has been commonly claimed to be a mixture chi-square distribution with the form $\frac{1}{4}\chi_2^2 : \frac{1}{2}\chi_1^2 : \frac{1}{4}\chi_0^2$, a special case discussed in Self and Liang (1987). We will give a new approximation and illustrate by simulation that this approximation produces conservative results.

4.2.2 Model II

When a QTL has a pleiotropic effect on several traits or several QTLs are closed linked, multivariate approaches for the genetic linkage analysis are more powerful than a single trait linkage analysis (Jiang and Zeng, 1995; Evans, 2002). Considering a bivariate trait analysis assuming only additive effect, the VC model for family k can be expressed as

$$(y_{k_1}, y_{k_2}) = (\mu_1, \mu_2) + (a_{k_1}, a_{k_2}) + (g_{k_1}, g_{k_2}) + (e_{k_1}, e_{k_2}), \quad k = 1, \cdots, K \tag{4.2.3}$$

where y_{k_t} is the tth ($t = 1, 2$) phenotypic vector for the kth family; μ_t (t=1,2) is the overall mean for the tth phenotypic trait, (a_{k_1}, a_{k_2}) is the random additive effect of a major gene for two phenotypic traits, respectively; g_{k_t} and e_{k_t} are the random polygene and residual effects. All random terms are assumed to be normally distributed with 0 means. The phenotypic variance-covariance matrix for family k is given as

$$Cov\begin{pmatrix} y_{k_1} \\ y_{k_2} \end{pmatrix} = \begin{pmatrix} \sigma^2_{a_1} & \sigma_{a12} \\ \sigma_{a12} & \sigma^2_{a_2} \end{pmatrix} \otimes \Pi_k + \begin{pmatrix} \sigma^2_{g_1} & \sigma_{g12} \\ \sigma_{g12} & \sigma^2_{g_2} \end{pmatrix} \otimes \Phi_k + \begin{pmatrix} \sigma^2_{e_1} & \sigma_{e12} \\ \sigma_{e12} & \sigma^2_{e_2} \end{pmatrix} \otimes I_k$$

where $\otimes$ is the kronecker product; σ_{a12}, σ_{g12}, and σ_{e12} are the covariances between the additive, polygene and the residual effects for the two traits, respectively. All the others are defined similarly as in Scenario 1.

The hypothesis test to detect major gene under a bivariate model is formulated as

$$\begin{cases} H_0 : \sigma^2_{a_1} = \sigma^2_{a_2} = \sigma_{a12} = 0 \\ H_1 : \sigma^2_{a_1} > 0 \text{ or } \sigma^2_{a_2} > 0 \end{cases} \tag{4.2.4}$$

Under the alternative, when either one of the variance terms is zero, the covariance term is restricted to zero.

4.2.3 Model III

Now consider the above bivariate trait model, but assuming both additive and dominant effect. The variance component model will be changed to

$$(y_{k_1}, y_{k_2}) = (\mu_1, \mu_2) + (a_{k_1}, a_{k_2}) + (d_{k_1}, d_{k_2}) + (g_{k_1}, g_{k_2}) + (e_{k_1}, e_{k_2}), \quad k = 1, \cdots, K \tag{4.2.5}$$

where (d_{k_1}, d_{k_2}) is the random dominant effect of major gene at the quantitative trait locus for two phenotypic traits. The variance-covariance matrix between two sibs is changed to

$$Cov \begin{pmatrix} y_{k_1} \\ y_{k_2} \end{pmatrix} = \begin{pmatrix} \sigma^2_{a_1} & \sigma_{a12} \\ \sigma_{a12} & \sigma^2_{a_2} \end{pmatrix} \otimes \Pi_k + \begin{pmatrix} \sigma^2_{d_1} & \sigma_{d12} \\ \sigma_{d12} & \sigma^2_{d_2} \end{pmatrix} \otimes \Delta_k + \begin{pmatrix} \sigma^2_{g_1} & \sigma_{g12} \\ \sigma_{g12} & \sigma^2_{g_2} \end{pmatrix} \otimes \Phi_k + \begin{pmatrix} \sigma^2_{e_1} & \sigma_{e12} \\ \sigma_{e12} & \sigma^2_{e_2} \end{pmatrix} \otimes I_k$$

The hypothesis to test major gene under this genetic model will be

$$\begin{cases} H_0 : \sigma^2_{a_1} = \sigma^2_{a_2} = \sigma_{a12} = \sigma^2_{d_1} = \sigma^2_{d_2} = \sigma_{d12} = 0 \\ H_1 : \sigma^2_{a_1} > 0 \text{ or } \sigma^2_{a_2} > 0 \text{ or } \sigma^2_{d_1} > 0 \text{ or } \sigma^2_{d_2} > 0 \end{cases} \tag{4.2.6}$$

Similarly, under the alternative, when either one of the variance terms (additive or dominant) is zero, the corresponding covariance term is restricted to zero.

4.3 Main results

For a random sample $X_1, X_2, ..., X_n$ of size n with a common density function $f(x, \boldsymbol{\theta})$, let $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)^T \in \Omega$ be the parameter vector, and $\boldsymbol{\theta}_0$ be the true population parameter vector. Let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} log f(x_i, \boldsymbol{\theta})$ be the log-likelihood function.

Assumption 1. *Following Chernoff (1954), the following assumptions are assumed:*

- *I. for every $\boldsymbol{\theta} \in G$ where G is a closure neighborhood centered at $\boldsymbol{\theta}_0$, the first three derivatives of $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist for almost all x.*
- *II. for every $\boldsymbol{\theta} \in G$, $|\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_i}|$ and $|\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}|$ are bounded by a finite integrable function $K(\boldsymbol{x})$, and $|\frac{\partial^3 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k}| < K(\boldsymbol{x})$ where $E[K(\boldsymbol{x})] < \infty$.*
- *III. The information matrix M ($M_{ij} = -\frac{1}{n} E(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j})$) is nonsingular for $\boldsymbol{\theta} \in G$, and $\|M\| < \infty$.*

Proposition 1. *Under Assumption 1, there is a vector $\hat{\boldsymbol{\theta}}_\varsigma$ in Ω, such that $\hat{\boldsymbol{\theta}}_\varsigma \longrightarrow \boldsymbol{\theta}_0$ in probability, and $(\hat{\boldsymbol{\theta}}_\varsigma - \boldsymbol{\theta}_0) = O_p(n^{-\frac{1}{2}})$.*

Proof: In terms of the arguments of Lehmann and Casella (1998), it is possible to search a sequence of points $\hat{\boldsymbol{\theta}}_\varsigma$ in the closed set G about $\boldsymbol{\theta}_0$ to locally maximize $\ell(\boldsymbol{\theta})$. Following Lemma 1 in Chernoff (1954), the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_\varsigma$ can be proved.

Denote a local maximum estimator by $\tilde{\boldsymbol{\theta}}_\varsigma$. Since the regularity conditions of Chernoff (1954) on the parameter set only derive the asymptotic distribution based on a global maximum estimator (Geyer 1994; Shapiro 2000), additional conditions are required to achieve the asymptotic equivalence of local estimators.

Assumption 2. *Considering more conditions as follows:*

- *IV.* $\hat{\boldsymbol{\theta}}_\varsigma$ *and* $\tilde{\boldsymbol{\theta}}_\varsigma$ *are* $\sqrt{n}$*-consistent optimizers.*

- *V. the parameter set* Ω *is a nearly convex set at* $\boldsymbol{\theta}_0$.

- *VI. Condition vi in Theorem 3.2 (Shapiro, 2000).*

Proposition 2. *If the above two assumptions (1 and 2) are satisfied,* $\hat{\boldsymbol{\theta}}_\varsigma$ - $\tilde{\boldsymbol{\theta}}_\varsigma = o_p(n^{-\frac{1}{2}})$.

Proof: See the proof of Theorem 3.2 in Shapiro (2000). In brief, two key steps are involved. First, the parameter set is nearly convex at $\boldsymbol{\theta}_0$. Comparing with convexity, near convexity is a loose condition. In particular, near convexity can be achieved by some smooth constraints in real application. When the fitted function is monotonically nondecreasing and twice continuously differentiable on a given interval, the parameter set is nearly convex at $\boldsymbol{\theta}_0$ under the Mangasarian-Fromovitz constraints. Next, Lipschitz continuous function $F_n(\hat{\boldsymbol{\theta}}_\varsigma)$ and $F_n(\tilde{\boldsymbol{\theta}}_\varsigma)$ are defined as minus $\frac{1}{n}$ time log-likelihood function in terms of $\hat{\boldsymbol{\theta}}_\varsigma$ and $\tilde{\boldsymbol{\theta}}_\varsigma$, respectively. When $F_n(\hat{\boldsymbol{\theta}}_\varsigma)$ and $F_n(\tilde{\boldsymbol{\theta}}_\varsigma)$ satisfy conditions 3.8 and 3.9 of Theorem 3.2 in Shapiro (2000), the asymptotic equivalence of $\hat{\boldsymbol{\theta}}_\varsigma$ and $\tilde{\boldsymbol{\theta}}_\varsigma$ is achieved by the property of the near convexity (condition A in Shapiro 2000).

It is well known that a cone contains several desirable properties that may simplify the optimization problem. According to the arguments of Chernoff (1954) and Self and Liang (1987), a cone is defined as

Definition 1. *The set* $\Omega \subset \mathbb{R}^m$ *is approximated by a cone* C_Ω *at* $\boldsymbol{\theta}_0$*, if*

$$\inf_{\boldsymbol{s}\in C_\Omega} \|\boldsymbol{s}-\boldsymbol{t}\| = o(\|\boldsymbol{t}-\boldsymbol{\theta}_0\|) \text{ for all } \boldsymbol{t} \in \Omega; \quad \inf_{\boldsymbol{t}\in\Omega} \|\boldsymbol{s}-\boldsymbol{t}\| = o(\|\boldsymbol{s}-\boldsymbol{\theta}_0\|) \text{ for all } \boldsymbol{s} \in C_\Omega$$

Note that the cone C_Ω is positively homogeneous if $\boldsymbol{s} \in C_\Omega$, $c(\boldsymbol{s}-\boldsymbol{\theta}_0)+\boldsymbol{\theta}_0 \in C_\Omega$ when c≥0. Moreover, $C_\Omega - \boldsymbol{\theta}_0$ with vertex at the origin is acquired by translating the cone C_Ω with vertex at $\boldsymbol{\theta}_0$. Thus, Ω can be approximated by a closed convex cone C_Ω with vertex at $\boldsymbol{\theta}_0$.

Proposition 3. *When* $\boldsymbol{\theta} = \mathbf{0}$*, F is the distribution of the MLE* $\hat{\boldsymbol{\theta}}_\varsigma$ *based on one observation* $\boldsymbol{Y}$ *with the population distribution* $N(\boldsymbol{\theta},M^{-1})$ *where* $\boldsymbol{\theta} \in C_\Omega - \boldsymbol{\theta}_0$*. If all previous conditions hold,* $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_\varsigma - \boldsymbol{\theta}_0)$ *weakly converges to F, a multivariate normal distribution with mean zero and covariance matrix* M^{-1}.

Proof: see the proof of Theorem 2 in Self and Liang (1987).

Assumption 3. *VII. Let* C_{Ω_0} *and* C_{Ω_1} *be two closed convex cones with vertex at* $\boldsymbol{\theta}_0$ *to approximate* Ω_0 *and* Ω_1*. Then* $C_{\Omega_0} - \boldsymbol{\theta}_0$ *with vertex at origin is also a closed convex cone by translating* C_{Ω_0} *at* $\boldsymbol{\theta}_0$.

Theorem 4.3.1. *If above Assumptions 1-3 hold and when* $\boldsymbol{\theta}=\boldsymbol{\theta}_0$*, the large sample distribution of the likelihood ratio (LR) is the same as that of the test* $\boldsymbol{\theta} \in C_{\Omega_0}$ *against* $\boldsymbol{\theta} \in C_{\Omega_1}$ *based on one observation* $\boldsymbol{Y}$ *generated from population distribution* $N(\boldsymbol{\theta},M^{-1})$*. Moreover, the LR is distributed as a mixture chi-square distributions with the form:*

$$Pr(LR > c^2) = \sum_{i=1}^{l(q)} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}) Pr(\chi^2_{r(T_{\nu^*0})} > c^2)$$

where $Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}})$ *is the mixing proportion corresponding to the chi-square components with* $\sum_{i=1}^{l(q)} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}) = 1$, *and* $r(T_{\nu^*0})$=*rank*(T_{ν^*0}).

Proof: Two arguments need to be proved. First, the LR can be approximated as the difference of two quadratic forms with respect to Ω_0 and Ω_1 (see Theorem 1 in Chernoff 1954). Follow the $\sqrt{n}$-consistency of the optimizer and the property of the approximating cone, the large sample distribution of the LR is the same as that of testing $\boldsymbol{\theta} \in C_{\Omega_0}$ against $\boldsymbol{\theta} \in C_{\Omega_1}$. Next, we prove that LR asymptotically follows a mixture chi-square distribution.

Following Chernoff (1954, Theorem 1), the asymptotic distribution of the LR is equivalent to the following quadratic approximation

$$LR = \inf_{\boldsymbol{\theta} \in C_{\Omega_0}} (\boldsymbol{Y} - \boldsymbol{\theta})' M (\boldsymbol{Y} - \boldsymbol{\theta}) - \inf_{\boldsymbol{\theta} \in C_{\Omega_1}} (\boldsymbol{Y} - \boldsymbol{\theta})' M (\boldsymbol{Y} - \boldsymbol{\theta}) \qquad (4.3.1)$$

Where $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\theta}, M^{-1})$. Subtracting $\boldsymbol{\theta}_0$ from $\boldsymbol{Y}$ and $\boldsymbol{\theta}$, we get an equivalent form of 4.3.1

$$LR = \inf_{\boldsymbol{\theta} \in C_{\Omega_0} - \boldsymbol{\theta}_0} (\boldsymbol{Y} - \boldsymbol{\theta})' M (\boldsymbol{Y} - \boldsymbol{\theta}) - \inf_{\boldsymbol{\theta} \in C_{\Omega_1} - \boldsymbol{\theta}_0} (\boldsymbol{Y} - \boldsymbol{\theta})' M (\boldsymbol{Y} - \boldsymbol{\theta}) \qquad (4.3.2)$$

with $\boldsymbol{Y} \sim \mathrm{N}(\mathbf{0}, M^{-1})$ and M is the Fisher information matrix.

Let $C^\vee = (C_{\Omega_1} - \boldsymbol{\theta}_0) \bigcap (C_{\Omega_0} - \boldsymbol{\theta}_0)^\perp$, where $(C_{\Omega_0} - \boldsymbol{\theta}_0)^\perp$ is the orthogonal

complement of $(C_{\Omega_0} - \boldsymbol{\theta}_0)$. Following the Pythagoras theorem, the statistic LR (in 4.3.2) can be expressed as

$$LR = \inf_{\boldsymbol{\theta} \in C^{\vee}} (\boldsymbol{Y} - \boldsymbol{\theta})' M (\boldsymbol{Y} - \boldsymbol{\theta}) \tag{4.3.3}$$

It can be seen that $C^{\vee}$ is also a closed polyhedral convex cone with q ($q \leq m$) dimension because $C^{\vee}$ is the intersection of convex cones. Thus a polar cone $C^{\vee 0}$ is defined as $C^{\vee 0} = \{\gamma \in \mathbb{R}^q; \gamma'\theta \leq 0, \ \forall\, \theta \in C^{\vee}\}$, and $(C^{\vee 0})^0 = C^{\vee}$ by the basic property of the polar cone.

Let $\mathbb{F}(C^{\vee})$ represent the set of all faces of $C^{\vee}$. Following Shapiro (1985), we can select a face $\nu^{\vee} \in \mathbb{F}(C^{\vee})$ corresponding to a polar face $\nu^{\vee 0} \in \mathbb{F}(C^{\vee 0})$ such that the linear spaces generated by $\nu^{\vee}$ and $\nu^{\vee 0}$ are orthogonal to each other. For one face $\nu^{\vee}$ (or $\nu^{\vee 0}$), we can find a projection $T_{\nu^{\vee}}$ (or $T_{\nu^{\vee 0}}$) (a symmetric idempotent matrix giving projection onto the space generated by $\nu^{\vee}$ (or $\nu^{\vee 0}$)) and $T_{\nu^{\vee}}$=I-$T_{\nu^{\vee 0}}$ since they are orthogonal. Then $T_{\nu^{\vee}}\boldsymbol{Y}$ (or $T_{\nu^{\vee 0}}\boldsymbol{Y}$) is a projection of a random vector $\boldsymbol{Y}$ onto $C^{\vee}$ (or $C^{\vee 0}$). For a given $\boldsymbol{Y}$, let g($\boldsymbol{Y}$)=$(g_1(\boldsymbol{Y}), g_2(\boldsymbol{Y}), ..., g_q(\boldsymbol{Y}))^T$ be the minimizer to achieve the infimum in (4.3.3). Define $\psi_{\nu^{\vee}|\boldsymbol{Y}} = \{\boldsymbol{Y} \in \Re^q : g(\boldsymbol{Y}) \in \nu^{\vee}\}$ so that $g(\boldsymbol{Y}) \in \nu^{\vee}$ if and only if $T_{\nu^{\vee}}\boldsymbol{Y} \in C^{\vee}$ and $T_{\nu^{\vee 0}}\boldsymbol{Y} \in C^{\vee 0}$. By Shapiro(1985), $\psi_{\nu^{\vee}|\boldsymbol{Y}}$ can also be defined by the inequalities as $\psi_{\nu^{\vee}|\boldsymbol{Y}} = \{\boldsymbol{Y} \in \Re^q : e'T_{\nu^{\vee}}\boldsymbol{Y} \leq 0, e \in C^{\vee 0}, f'T_{\nu^{\vee 0}}\boldsymbol{Y} \leq 0, f \in C^{\vee}\}$. Thus, g($\boldsymbol{Y}$)= $T_{\nu^{\vee}}\boldsymbol{Y} \in C^{\vee}$, for all $\boldsymbol{Y} \in \psi_{\nu^{\vee}|\boldsymbol{Y}}$.

Consequently, the likelihood ratio statistic in (4.3.3) is expressed as:

$$LR = (\boldsymbol{Y} - g(\boldsymbol{Y}))' M (\boldsymbol{Y} - g(\boldsymbol{Y})) \quad \textit{for all } \boldsymbol{Y} \in \psi_{\nu^{\vee}|\boldsymbol{Y}} \tag{4.3.4}$$

Note that the set $\psi_{\nu^{\vee}|\boldsymbol{Y}}$ is composed of several almost disjoint sets $\psi^{i}_{\nu^{\vee}|\boldsymbol{Y}}, i = 1, ..., l(q)$. The total number of these disjoint subsets ($l(q)$) are counted by the general form of binomial theorem, i.e., $l(q) = 2^{q-o}$, q is the number of parameters in $C^{\vee}$ and o is the number of covariance terms in $C^{\vee}$. Moreover, All these subsets are classified into $q - o + 1$ categories. To display these subsets, we start from the simple case that no covariance term is in $\psi_{\nu^{\vee}|\boldsymbol{Y}}$ ($o = 0$). The subsets of $\psi_{\nu^{\vee}|\boldsymbol{Y}}$ are given as the following table 4.1.

Table 4.1: The whole possible subsets of $\psi_{\nu^\vee|\boldsymbol{Y}}$ without covariance terms.

set	subsets	number
$\psi_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y};\boldsymbol{Y}\in\mathbb{R}^q, g(\boldsymbol{Y})\in\nu^\vee\}$	$\psi^1_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y};Y_1>0,Y_2>0,...,Y_q>0,g(\boldsymbol{Y})\in\nu^\vee\}$	$\binom{q}{0}$
	$\psi^2_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y};Y_1\leq 0,Y_2>0,...,Y_q>0,g(\boldsymbol{Y})\in\nu^\vee\}$	$\binom{q}{1}$
	$\psi^3_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y};Y_1>0,Y_2\leq 0,...,Y_q>0,g(\boldsymbol{Y})\in\nu^\vee\}$	
	...	
	$\psi^{q+1}_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y};Y_1>0,Y_2>0,...,Y_q\leq 0,g(\boldsymbol{Y})\in\nu^\vee\}$	
	...	
	...	
	...	
	$\psi^{2^q}_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y};Y_1\leq 0,Y_2\leq 0,...,Y_q\leq 0,g(\boldsymbol{Y})\in\nu^\vee\}$	$\binom{q}{q}$

When a covariance term occurs in $C^{\vee}$, we use one simple case to describe the relation among the parameters. For example: $C^{\vee} = \{\boldsymbol{\theta}; \theta_1 > 0, \theta_2 > 0, \theta_3 \in \mathbb{R}\}$, where θ_1 is a variance term for trait one, θ_2 is the variance term for trait 2, θ_3 is the covariance between the two traits. Because of the definition of covariance, θ_3 occurs only when $\theta_1 > 0$ and $\theta_2 > 0$, so θ_3 is represented by $\theta_3 I(\theta_1 > 0, \theta_2 > 0)$. In the corresponding way, the estimator of θ_3 is denoted as $Y_3 I(Y_1 > 0, Y_2 > 0)$, and Y_1 and Y_2 are estimators of variances for two traits. In accordance with this constrain, the set $\psi_{\nu^{\vee}|\boldsymbol{Y}}$ is denoted as $\psi_{\nu^{\vee}|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_i \in \mathbb{R}, i \in q^{\vee} \backslash o^{\vee}, Y_j I(Y_{j-2} > 0, Y_{j-1} > 0) \in \mathbb{R}, j \in o^{\vee}, g(\boldsymbol{Y}) \in \nu^{\vee}\}$, where set $q^{\vee}$ is defined as $q^{\vee} = \{1, 2, ..., q\}$ and $o^{\vee}$ is denoted as a subset of $q^{\vee}$, and is shown as $o^{\vee} = \{3, 6, ..., q\} = \{3k^{\vee}, k^{\vee} = 1, 2, ..., \frac{q}{3}\}$. Considering the property of covariance term, the partition of $\psi_{\nu^{\vee}|\boldsymbol{Y}}$ is not related to the covariance term. Thus the whole subsets under this constrained condition are shown as the table 4.2.

Table 4.2: The whole possible subsets of $\psi_{\nu^\vee|\boldsymbol{Y}}$ with covariance terms.

subsets	number
$\psi^1_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y}; Y_1>0, Y_2>0, Y_3\in\mathbb{R}, Y_4>0, Y_5>0, Y_6\in\mathbb{R}, ..., Y_{q-2}>0, Y_{q-1}>0, Y_q\in\mathbb{R}, g(\boldsymbol{Y})\in\nu^\vee\}$	$\binom{q-o}{0}$
$\psi^2_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y}; Y_1\leq 0, Y_2>0, Y_4>0, Y_5>0, Y_6\in\mathbb{R}, ..., Y_{q-2}>0, Y_{q-1}>0, Y_q\in\mathbb{R}, g(\boldsymbol{Y})\in\nu^\vee\}$	$\binom{q-o}{1}$
$\psi^3_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y}; Y_1>0, Y_2\leq 0, Y_4>0, Y_5>0, Y_6\in\mathbb{R}, ..., Y_{q-2}>0, Y_{q-1}>0, Y_q\in\mathbb{R}, g(\boldsymbol{Y})\in\nu^\vee\}$	
$\psi^4_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y}; Y_1>0, Y_2>0, Y_3\in\mathbb{R}, Y_4\leq 0, Y_5>0, ..., Y_{q-2}>0, Y_{q-1}>0, Y_q\in\mathbb{R}, g(\boldsymbol{Y})\in\nu^\vee\}$	
...	
$\psi^{q-o+1}_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y}; Y_1>0, Y_2>0, Y_3\in\mathbb{R}, Y_4>0, Y_5>0, Y_6\in\mathbb{R}, ..., Y_{q-2}>0, Y_{q-1}\leq 0, g(\boldsymbol{Y})\in\nu^\vee\}$	
...	
...	
...	
$\psi^{2^{q-o}}_{\nu^\vee\|\boldsymbol{Y}}=\{\boldsymbol{Y}; Y_1\leq 0, Y_2\leq 0, Y_4\leq 0, Y_5\leq 0..., Y_{q-2}\leq 0, Y_{q-1}\leq 0, g(\boldsymbol{Y})\in\nu^\vee\}$	$\binom{q-o}{q-o}$

The general form of the whole number of subsets is $l(q) = 2^{q-o}$, and these subsets consist of $q - o + 1$ groups. Consequently, $\psi_{\nu^\vee|\boldsymbol{Y}} = \bigcup_{i=1}^{l(q)} \psi^i_{\nu^\vee|\boldsymbol{Y}}$.

Considering a linear transformation on $\boldsymbol{Y}$ and $\boldsymbol{\theta}$, a new closed convex cone C^* is defined as $C^*=\{\boldsymbol{\theta}^\vee;\ E^{\frac{1}{2}}D'\boldsymbol{\theta}, \boldsymbol{\theta} \in C^\perp\}$, where $DED' = M$, and a new random vector $\boldsymbol{Z}$ ($\boldsymbol{Z}=E^{\frac{1}{2}}D'\boldsymbol{Y}$) is distributed with multivariate normal distribution with mean zero and an identity covariance. In terms of this random vector $\boldsymbol{Z}$ the likelihood ratio LR (in 4.3.3) is evaluated equivalently as:

$$LR = \inf_{\boldsymbol{\theta}^* \in C^*} \|\boldsymbol{Z} - \boldsymbol{\theta}^*\|^2 \tag{4.3.5}$$

In the same way, C^* is a closed convex cone and C^{*0} is denoted as the polar cone of C^* with $(C^{*0})^0 = C^*$. So there is a face $\nu^* \in \mathbb{F}(C^*)$ (or $\nu^{*0} \in \mathbb{F}(C^{*0})$) such that a symmetric idempotent matrix T_{ν^*} (or $T_{\nu^{*0}}$) giving projection onto the space generated by ν^* (or ν^{*0}) is defined. The linear transformation of $\boldsymbol{Y}$ to $\boldsymbol{Z}$ guarantees that there also exists a minimizer denoted by $\mathrm{d}(\boldsymbol{Z})=(d_1(\boldsymbol{Z}), d_1(\boldsymbol{Z}), ..., d_q(\boldsymbol{Z}))^T$ for (4.3.5), in which $\mathrm{d}(\boldsymbol{Z}) = T_{\nu^*}\boldsymbol{Z} \in C^*$, $\forall\ \boldsymbol{Z} \in \psi_{\nu^*|\boldsymbol{Z}}$, where $\psi_{\nu^*|\boldsymbol{Z}}$ can be defined by a linear transformation from $\psi_{\nu^\perp|\boldsymbol{Y}}$.

Note that the set $\psi_{\nu^*|\boldsymbol{Z}}$ is also a polyhedral convex cone by its definition and satisfies the conditions of Lemma 3.1 (Shapiro 1985), and $T_{\nu^{*0}}$ is an symmetric idempotent matrix corresponding to face ν^{*0}, then the likelihood ratio statistic (4.3.5)

is written:

$$LR = \|\boldsymbol{Z} - d(\boldsymbol{Z})\|^2 = \|\boldsymbol{Z} - T_{\nu^*}\boldsymbol{Z}\|^2 = \boldsymbol{Z}'(I - T_{\nu^*})\boldsymbol{Z} = \boldsymbol{Z}'T_{\nu^*0}\boldsymbol{Z}, \text{ for all } \boldsymbol{Z} \in \psi_{\nu^*|\boldsymbol{Z}} \tag{4.3.6}$$

It is clear that the minimum value of LR obtained for $\boldsymbol{Y} \in \psi_{\nu^\perp|\boldsymbol{Y}}$ (in 4.3.4) is equivalent to the infimum value of LR obtained for $\boldsymbol{Z} \in \psi_{\nu^*|\boldsymbol{Z}}$ (in 4.3.6).

Note that the set $\psi_{\nu^*|\boldsymbol{Z}}$ is also made up of several almost disjoint sets, i.e., $\psi_{\nu^*|\boldsymbol{z}} = \bigcup_{i=1}^{l(q)} \psi^i_{\nu^*|\boldsymbol{z}}$. Condition on $\boldsymbol{Z} \in \psi^i_{\nu^*|\boldsymbol{Z}}$, LR follows a chi-square distribution with rank(T_{ν^*0})=rank (I-T_{ν^*}) degrees of freedom. By Bayes' theorem, the distribution of LR (in 4.3.6) is derived to be a mixture chi-square distribution.

To control the significance of hypothesis test in α level, The probability that LR rejects the null hypothesis under the null condition is evaluated. Given a positive number $c^2 > 0$ and a random vector $\boldsymbol{Y}$, the expression of this probability is written as:

$$\begin{aligned} Pr(LR > c^2) &= Pr((\boldsymbol{Y} - g(\boldsymbol{Y}))'M(\boldsymbol{Y} - g(\boldsymbol{Y})) > c^2, \boldsymbol{Y} \in \psi_{\nu^\perp|\boldsymbol{Y}}) \\ &= Pr((\boldsymbol{Y} - g(\boldsymbol{Y}))'M(\boldsymbol{Y} - g(\boldsymbol{Y})) > c^2, \boldsymbol{Y} \in \bigcup_{i=1}^{l(q)} \psi^i_{\nu^\perp|\boldsymbol{Y}}) \end{aligned} \tag{4.3.7}$$

Applying the Distributive law of sets and the union rule for these almost disjoint sets, the representation of (4.3.7) is changed to be:

$$
\begin{aligned}
Pr(LR > c^2) =& Pr(\bigcup_{i=1}^{l(q)} \{(\boldsymbol{Y} - g(\boldsymbol{Y}))' M(\boldsymbol{Y} - g(\boldsymbol{Y})) > c^2, \boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}\}) \\
=& \sum_{i=1}^{l(q)} Pr((\boldsymbol{Y} - g(\boldsymbol{Y}))' M(\boldsymbol{Y} - g(\boldsymbol{Y})) > c^2, \boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}) \\
=& \sum_{i=1}^{l(q)} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}) \\
& Pr((\boldsymbol{Y} - g(\boldsymbol{Y}))' M(\boldsymbol{Y} - g(\boldsymbol{Y})) > c^2 | \boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}})
\end{aligned}
\tag{4.3.8}
$$

According to the resemblance between the result in (4.3.4) and comment in (4.3.6), the representation of the probability is changed to be:

$$
\begin{aligned}
Pr(LR > c^2) &= \sum_{i=1}^{l(q)} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}) Pr(\boldsymbol{Z}' T_{\nu^*0} \boldsymbol{Z} > c^2 | \boldsymbol{Z} \in \psi^i_{\nu^*|\boldsymbol{Z}}) \\
&= \sum_{i=1}^{l(q)} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}) Pr(\chi^2_{r(T_{\nu^*0})} > c^2)
\end{aligned}
\tag{4.3.9}
$$

where $Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}})$ is the mixing proportion corresponding to the chi-square components with $\sum_{i=1}^{l(q)} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\perp|\boldsymbol{Y}}) = 1$, and r($T_{\nu^*0}$)=rank($T_{\nu^*0}$). This completes the proof.

Following Theorem 4.3.1, we now evaluate the distribution of the LR statistic for the three models in linkage analysis we mentioned in Section 4.2.

Model I: The parameters of this model are given as $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\} = \{\mu, \sigma^2_a, \sigma^2_d, \sigma^2_g, \sigma^2_e\}$, and the approximating cone under the null hypothesis is defined

as C_{Ω_0}=$\{\boldsymbol{\theta}; \theta_1 \in \mathbb{R}, \theta_2 = 0, \theta_3 = 0, \theta_4 > 0, \theta_5 > 0\}$ against C_{Ω_1}=$\{\boldsymbol{\theta}; \theta_1 \in \mathbb{R}, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 > 0\}$ under the alternative. The number of parameters to be tested for q is 2 and for o is 0, that is, there is no covariance term in model I. Thus $\psi_{\nu^\vee|\boldsymbol{Y}}$ consists of 2^{q-o}=2^2=4 almost disjoint sets with $q - o + 1 = 2 - 0 + 1 = 3$ categories:

(i) $\psi^1_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(ii) $\psi^2_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 < 0, g(\boldsymbol{y}) \in \nu^\vee\}$,

$\psi^3_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(iii) $\psi^4_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$.

In the same way, $\psi_{\nu^*|\boldsymbol{Z}}$ can be divided into four almost disjoint subsets by linear transformation. When $\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}$, $LR = \boldsymbol{Z}'T_{\nu 0}\boldsymbol{Z} = Z_1^2 + Z_2^2 \sim \chi^2_2$ where $Z \sim N(0, I)$, and the corresponding mixture proportion is estimated by $Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}})$. As $\boldsymbol{Y}$ is in the 2nd category (i.e., $\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}$, i=2,3), $LR \sim \chi^2_1$ with the corresponding mixing proportion calculated by $Pr(\boldsymbol{Y} \in \psi^2_{\nu^\vee|\boldsymbol{Y}})$+$Pr(\boldsymbol{Y} \in \psi^3_{\nu^\vee|\boldsymbol{Y}})$. For the last category, $LR \sim \chi^2_0$ for $\boldsymbol{Y} \in \psi^4_{\nu^\vee|\boldsymbol{Y}}$, and the mixing proportion is $Pr(\boldsymbol{Y} \in \psi^4_{\nu^\vee|\boldsymbol{Y}})$. The calculation of the mixing proportion follows Plackett(1954) or Kendall(1941). Specifically, $Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}) = \frac{\pi - cos^{-1}\rho_{12}}{2\pi}$, $\sum_{i=2}^{3} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}) = \frac{1}{2}$, and $Pr(\boldsymbol{Y} \in \psi^4_{\nu^\vee|\boldsymbol{Y}}) = \frac{cos^{-1}\rho_{12}}{2\pi}$, and ρ_{12} is the correlation between estimator of additive gene effect and that of dominance effect. Finally, the distribution is approximated as

$$Pr(LR > c^2) = \frac{\pi - cos^{-1}\rho_{12}}{2\pi} P(\chi^2_2 > c^2) + \frac{1}{2} P(\chi^2_1 > c^2) \qquad (4.3.10)$$

Model II: For model II (in (4.2.3)), only the random additive effect of QTL (a_{k_1}, a_{k_2}) for each trait is considered. The parameters of additive major gene effect are denoted as: $\sigma^2_{a_1}, \sigma^2_{a_2}, \sigma_{a_{12}}$ where $\sigma_{a_{12}}$ is the covariance term between two traits. Similarly, two covariance terms are denoted for polygene effect and random residual effect. All parameters in this model are defined as: $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}\} = \{\mu_1, \mu_2, \sigma^2_{a_1}, \sigma^2_{a_2}, \sigma^2_{a_{12}}, \sigma^2_{g_1}, \sigma^2_{g_2}, \sigma^2_{g_{12}}, \sigma^2_{e_1}, \sigma^2_{e_2}, \sigma^2_{e_{12}}\}$. The parameter approximating cone under the null hypothesis is defined as $C_{\Omega_0} = \{\boldsymbol{\theta}; \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}, \theta_3 = 0, \theta_4 = 0, \theta_5 = 0, \theta_6 > 0, \theta_7 > 0, \theta_8 \in \mathbb{R}, \theta_9 > 0, \theta_{10} > 0, \theta_{11} \in \mathbb{R}\}$. Similarly, the cone under the alternative hypothesis is denoted as $C_{\Omega_1} = \{\boldsymbol{\theta}, \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}, \theta_3 > 0, \theta_4 > 0, \theta_5 \in \mathbb{R}, \theta_6 > 0, \theta_7 \geq 0, \theta_8 \in \mathbb{R}, \theta_9 \geq 0, \theta_{10} \geq 0, \theta_{11} \in \mathbb{R}\}$. Corresponding to the hypothesis test, the number of tested parameters q is 3 and o is 1, then the set $\psi_{\nu^\vee|\boldsymbol{Y}}$ has $2^{3-1} = 4$ almost disjoint subsets and all these subsets are classified into 3-1+1=3 groups.

(i) $\psi^1_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$;

(ii) $\psi^2_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$,

$\psi^3_{\nu^\perp|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(iii) $\psi^4_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$.

The estimator of covariance term is only observed in $\psi^1_{\nu^\vee|\boldsymbol{Y}}$, and it will vanish automatically when $Y_1 \leq 0$ or $Y_2 \leq 0$. Moreover, $\psi_{\nu^\vee|\boldsymbol{Y}} = \bigcup_{i=1}^{2^{3-1}} \psi^i_{\nu^\vee|\boldsymbol{Y}}$, and $\psi_{\nu^*|\boldsymbol{Z}}$ in terms of $\boldsymbol{\theta}^*$ is defined in a similar way, that is, $\psi_{\nu^*|\boldsymbol{Z}} = \bigcup_{i=1}^{2^{3-1}} \psi^i_{\nu^*|\boldsymbol{Z}}$. When $\boldsymbol{Y} \in \psi_{\nu^\vee|\boldsymbol{Y}}$, the LR is shown in the form:

$$LR = \begin{cases} Z_1^2 + Z_2^2 + Z_3^2 \sim \chi_3^2 & \text{with mixing prop : } Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}) \\ Z_1^2 \sim \chi_1^2 & \text{with mixing prop : } Pr(\boldsymbol{Y} \in \psi^2_{\nu^\vee|\boldsymbol{Y}}) \\ Z_2^2 \sim \chi_1^2 & \text{with mixing prop : } Pr(\boldsymbol{Y} \in \psi^3_{\nu^\vee|\boldsymbol{Y}}) \\ 0 \sim \chi_0^2 & \text{with mixing prop : } Pr(\boldsymbol{Y} \in \psi^4_{\nu^\vee|\boldsymbol{Y}}) \end{cases}$$

As $\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}$, $LR \sim \chi_3^2$, and the corresponding mixture proportion is calculated as $Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}) = Pr(Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}) = Pr(Y_1 > 0, Y_2 > 0) = \frac{\pi - cos^{-1}\rho_{12}}{2\pi}$. When $\boldsymbol{Y}$ is in the 2nd category (i.e., $\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}$ i=2,3), $LR \sim \chi_1^2$ with mixing probability $\sum_{i=2}^3 Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}) = \frac{1}{2}$. For $\boldsymbol{Y} \in \psi^4_{\nu^\vee|\boldsymbol{Y}}$, $LR \sim \chi_0^2$, the relevant mixing probability is evaluated by $Pr(\boldsymbol{Y} \in \psi^4_{\nu^\vee|\boldsymbol{Y}}) = \frac{cos^{-1}\rho_{12}}{2\pi}$. These three mixing proportions is the same as those in model I. Hence, the probability of LR under model II is in the form:

$$Pr(LR > c^2) = \frac{\pi - cos^{-1}\rho_{12}}{2\pi} P(\chi_3^2 > c^2) + \frac{1}{2} P(\chi_1^2 > c^2) \tag{4.3.11}$$

Model III: In model III, random dominant effects (d_{k_1}, d_{k_2}) are considered. The parameters under this model is denoted as: $\boldsymbol{\theta}=\{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}\}=\{\mu_1, \mu_2, \sigma^2_{a_1}, \sigma^2_{a_2}, \sigma^2_{a_{12}}, \sigma^2_{d_1}, \sigma^2_{d_2}, \sigma^2_{d_{12}}, \sigma^2_{g_1}, \sigma^2_{g_2}, \sigma^2_{g_{12}}, \sigma^2_{e_1}, \sigma^2_{e_2}, \sigma^2_{e_{12}}\}$. The approximating cone under the null hypothesis is $C_{\Omega_0}=\{\boldsymbol{\theta}; \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}, \theta_3 = 0, \theta_4 = 0, \theta_5 = 0, \theta_6 = 0, \theta_7 = 0, \theta_8 = 0, \theta_9 > 0, \theta_{10} > 0, \theta_{11} \in \mathbb{R}, \theta_{12} > 0, \theta_{13} > 0, \theta_{14} \in \mathbb{R}\}$, and the cone under the alternative hypothesis is denoted as $C_{\Omega_1}=\{\boldsymbol{\theta}, \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}, \theta_3 > 0, \theta_4 > 0, \theta_5 \in \mathbb{R}, \theta_6 > 0, \theta_7 > 0, \theta_8 \in \mathbb{R}, \theta_9 >$

$0, \theta_{10} > 0, \theta_{11} \in \mathbb{R}, \theta_{12} > 0, \theta_{13} > 0, \theta_{14} \in \mathbb{R}\}$. The number of testing parameters in model III for q is 6 and for o is 2. Then the set $\psi_{\nu^\vee|\boldsymbol{Y}}$ can be partitioned into $2^{6-2} = 16$ almost disjoint subsets that comprise 6-2+1=5 categories.

(i) $\psi^1_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$;

(ii)$\psi^2_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$,

$\psi^3_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^4_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$,

$\psi^5_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(iii)$\psi^6_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^7_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^8_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^9_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{10}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{11}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(iv)$\psi^{12}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{13}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{14}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{15}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(v)$\psi^{16}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

The set $\psi_{\nu^\vee|\boldsymbol{Y}} = \bigcup_{i=1}^{2^{6-2}} \psi^i_{\nu^\vee|\boldsymbol{Y}}$, accordingly, $\psi_{\nu^*|\boldsymbol{Z}} = \bigcup_{i=1}^{2^{6-2}} \psi^i_{\nu^*|\boldsymbol{Z}}$. Based on these almost disjoint subsets, the limiting distribution of LR with the mixing proportion is in the form:

$0, \theta_{10} > 0, \theta_{11} \in \mathbb{R}, \theta_{12} > 0, \theta_{13} > 0, \theta_{14} \in \mathbb{R}\}$. The number of testing parameters in model III for q is 6 and for o is 2. Then the set $\psi_{\nu^\vee|\boldsymbol{Y}}$ can be partitioned into $2^{6-2} = 16$ almost disjoint subsets that comprise 6-2+1=5 categories.

(i) $\psi^1_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$;

(ii)$\psi^2_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$,

$\psi^3_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^4_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$,

$\psi^5_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(iii)$\psi^6_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^7_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^8_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^9_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{10}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{11}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(iv)$\psi^{12}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 > 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{13}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 > 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{14}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 > 0, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

$\psi^{15}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 > 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

(v)$\psi^{16}_{\nu^\vee|\boldsymbol{Y}} = \{\boldsymbol{Y}; Y_1 \leq 0, Y_2 \leq 0, Y_4 \leq 0, Y_5 \leq 0, g(\boldsymbol{y}) \in \nu^\vee\}$;

The set $\psi_{\nu^\vee|\boldsymbol{Y}} = \bigcup_{i=1}^{2^{6-2}} \psi^i_{\nu^\vee|\boldsymbol{Y}}$, accordingly, $\psi_{\nu^*|\boldsymbol{Z}} = \bigcup_{i=1}^{2^{6-2}} \psi^i_{\nu^*|\boldsymbol{Z}}$. Based on these almost disjoint subsets, the limiting distribution of LR with the mixing proportion is in the form:

$$
LR = \begin{cases}
Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2 \sim \chi_6^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}) \\
Z_1^2 + Z_4^2 + Z_5^2 + Z_6^2 \sim \chi_4^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^2_{\nu^\vee|\boldsymbol{Y}}) \\
Z_2^2 + Z_4^2 + Z_5^2 + Z_6^2 \sim \chi_4^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^3_{\nu^\vee|\boldsymbol{Y}}) \\
Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 \sim \chi_4^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^4_{\nu^\vee|\boldsymbol{Y}}) \\
Z_1^2 + Z_2^2 + Z_3^2 + Z_5^2 \sim \chi_4^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^5_{\nu^\vee|\boldsymbol{Y}}) \\
Z_4^2 + Z_5^2 + Z_6^2 \sim \chi_3^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^6_{\nu^\vee|\boldsymbol{Y}}) \\
Z_1^2 + Z_2^2 + Z_3^2 \sim \chi_3^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^7_{\nu^\vee|\boldsymbol{Y}}) \\
Z_2^2 + Z_5^2 \sim \chi_2^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^8_{\nu^\vee|\boldsymbol{Y}}) \\
Z_2^2 + Z_4^2 \sim \chi_2^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^9_{\nu^\vee|\boldsymbol{Y}}) \\
Z_1^2 + Z_5^2 \sim \chi_2^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^{10}_{\nu^\vee|\boldsymbol{Y}}) \\
Z_1^2 + Z_4^2 \sim \chi_2^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^{11}_{\nu^\vee|\boldsymbol{Y}}) \\
Z_5^2 \sim \chi_1^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^{12}_{\nu^\vee|\boldsymbol{Y}}) \\
Z_4^2 \sim \chi_1^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^{13}_{\nu^\vee|\boldsymbol{Y}}) \\
Z_2^2 \sim \chi_1^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^{14}_{\nu^\vee|\boldsymbol{Y}}) \\
Z_1^2 \sim \chi_1^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^{15}_{\nu^\vee|\boldsymbol{Y}}) \\
0 \sim \chi_0^2 & \text{with mixing prop } : Pr(\boldsymbol{Y} \in \psi^{16}_{\nu^\vee|\boldsymbol{Y}})
\end{cases}
$$

Therefore, the probability of LR under model III is in the form:

$$\begin{aligned} Pr(LR > c^2) =& Pr(\boldsymbol{Y} \in \psi^1_{\nu\vee|\boldsymbol{Y}}) P(\chi^2_6 > c^2) \\ &+ \sum_{i=2}^{5} Pr(\boldsymbol{Y} \in \psi^i_{\nu\vee|\boldsymbol{Y}}) P(\chi^2_4 > c^2) \\ &+ \sum_{i=6}^{7} Pr(\boldsymbol{Y} \in \psi^i_{\nu\vee|\boldsymbol{Y}}) P(\chi^2_3 > c^2) \\ &+ \sum_{i=8}^{11} Pr(\boldsymbol{Y} \in \psi^i_{\nu\vee|\boldsymbol{Y}}) P(\chi^2_2 > c^2) \\ &+ \sum_{i=12}^{15} Pr(\boldsymbol{Y} \in \psi^i_{\nu\vee|\boldsymbol{Y}}) P(\chi^2_1 > c^2) \end{aligned} \tag{4.3.12}$$

All mixing proportions can be calculated based on previous results (Kudo p.415 1963, Shapiro p.141 1985):

(1) the mixing proportion corresponding to chi-square random variable with 4 df is $\sum_{i=2}^{5} Pr(\boldsymbol{Y} \in \psi^i_{\nu\vee|\boldsymbol{Y}})$, and this probability can be estimated as:

$$\begin{aligned} \sum_{i=2}^{5} Pr(\boldsymbol{Y} \in \psi^i_{\nu\vee|\boldsymbol{Y}}) =& \sum_{\substack{a=1,2,4,5 \\ a\neq b, a\neq c, a\neq d, a\neq e}} Pr(Y_a < 0, Y_b > 0, Y_c > 0, \\ & Y_d > 0, Y_e \in \mathbb{R}) \\ =& \sum_{\substack{a=1,2,4,5 \\ a\neq b, a\neq c, a\neq d}} Pr(Y_a < 0, Y_b > 0, Y_c > 0, Y_d > 0) \\ =& \frac{1}{8\pi}(8\pi - \sum_{a>b; a\neq c, b\neq c} cos^{-1}\rho_{ab|c}) \end{aligned}$$

where $\rho_{ab|c}$ is calculated from the equation, $\rho_{ab|c} = \frac{\rho_{ab}-\rho_{ac}\rho_{bc}}{\sqrt{\rho_{ac}^2}\sqrt{\rho_{bc}^2}}$ (a, b, c=1,...,4 and 4=q-o).

(2) With respect to the chi-square random variable with 3 df, the mixing proportion is given as $\sum_{i=6}^{7} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$, the calculation of this probability is evaluated as:

$$\begin{aligned}\sum_{i=6}^{7} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}) =& Pr(Y_1 < 0, Y_2 < 0, Y_4 > 0, Y_5 > 0, Y_6 \in \mathbb{R}) \\ &+ Pr(Y_1 > 0, Y_2 > 0, Y_3 \in \mathbb{R}, Y_4 < 0, Y_5 < 0) \\ =& Pr(Y_1 < 0, Y_2 < 0, Y_4 > 0, Y_5 > 0) \\ &+ Pr(Y_1 > 0, Y_2 > 0, Y_4 < 0, Y_5 < 0) \\ =& \frac{1}{4\pi^2}[cos^{-1}\rho_{12}(\pi - cos^{-1}\rho_{45|12}) \\ &+ cos^{-1}\rho_{45}(\pi - cos^{-1}\rho_{12|45})]\end{aligned}$$

note that $\rho_{cd|ab}$ is estimated from

$$\rho_{cd|ab} = \frac{\rho_{ab} - \frac{\rho_{ac}\rho_{bc}+\rho_{ad}\rho_{bd}-\rho_{ac}\rho_{bd}\rho_{cd}-\rho_{ad}\rho_{bc}\rho_{cd}}{1-\rho_{cd}^2}}{\sqrt{\frac{1-\rho_{cd}^2-\rho_{ac}^2-\rho_{ad}^2+2\rho_{ac}^2\rho_{ad}^2\rho_{cd}^2}{1-\rho_{cd}^2}}\sqrt{\frac{1-\rho_{cd}^2-\rho_{bc}^2-\rho_{bd}^2+2\rho_{bc}^2\rho_{bd}^2\rho_{cd}^2}{1-\rho_{cd}^2}}}.$$

(3) $\sum_{i=8}^{11} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$ is the mixing probability for the chi-square component with 2 df. This mixing proportion is evaluated as:

$$\sum_{i=8}^{11} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}) = \sum_{a=1,2;c=4,5} Pr(Y_a < 0, Y_b > 0, Y_c < 0, Y_d > 0)$$

$$= \frac{1}{4\pi^2}[cos^{-1}\rho_{14}(\pi - cos^{-1}\rho_{25|14})$$

$$+ cos^{-1}\rho_{15}(\pi - cos^{-1}\rho_{24|15})$$

$$+ cos^{-1}\rho_{24}(\pi - cos^{-1}\rho_{15|24})$$

$$+ cos^{-1}\rho_{25}(\pi - cos^{-1}\rho_{14|25})]$$

where $\rho_{cd|ab}$ is defined in the same way.

(4) Following the comment provided by Shapiro (1985), the mixing proportions are assigned equally on the even and odd places. Thus the mixing probability $\sum_{i=12}^{15} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$ for chi-square component with 1 df is calculated as:

$$\sum_{i=12}^{15} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}) = \sum_{a=1,2,4,5} Pr(Y_a > 0, Y_b < 0, Y_c < 0, Y_d < 0)$$

$$= \frac{1}{2} - \sum_{i=2}^{5} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$$

$$= \frac{1}{8\pi}\Big(\sum_{a>b;a\neq c,b\neq c} cos^{-1}\rho_{ab|c} - 4\pi\Big)$$

where there is a resemblance about calculation of $\rho_{ab|c}$ between $\sum_{i=12}^{15} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$ and $\sum_{i=2}^{5} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$.

(5) An approximated estimation of mixing proportion $Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}})$ correspond-

ing to the chi-square component with 6 df is derived according to the above comment. This mixing probability is evaluated as:

$$Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}) + Pr(\boldsymbol{Y} \in \psi^{16}_{\nu^\vee|\boldsymbol{Y}}) = \frac{1}{2} - \sum_{i=6}^{11} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$$

Suppose two mixing proportions $Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}})$ for chi-square component with 6 df and $Pr(\boldsymbol{Y} \in \psi^{16}_{\nu^\vee|\boldsymbol{Y}})$ for chi-square component with 0 df equally share the probability $\frac{1}{2} - \sum_{i=6}^{11} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}})$. Therefore, the mixing proportion $Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}})$ is approximated as:

$$\begin{aligned}
Pr(\boldsymbol{Y} \in \psi^1_{\nu^\vee|\boldsymbol{Y}}) &= \frac{1}{4} - \frac{1}{2}\sum_{i=6}^{11} Pr(\boldsymbol{Y} \in \psi^i_{\nu^\vee|\boldsymbol{Y}}) \\
&= \frac{1}{4} - \frac{1}{8\pi^2}[cos^{-1}\rho_{12}(\pi - cos^{-1}\rho_{45|12}) \\
&\quad + cos^{-1}\rho_{45}(\pi - cos^{-1}\rho_{12|45}) \\
&\quad + cos^{-1}\rho_{14}(\pi - cos^{-1}\rho_{25|14}) \\
&\quad + cos^{-1}\rho_{15}(\pi - cos^{-1}\rho_{24|15}) \\
&\quad + cos^{-1}\rho_{24}(\pi - cos^{-1}\rho_{15|24}) \\
&\quad + cos^{-1}\rho_{25}(\pi - cos^{-1}\rho_{14|25})]
\end{aligned}$$

4.4 Simulation

We designed simulations to evaluate the limiting distribution of the LRT. The results of the new approximation are compared with those from Self and Liang (1987) and Amos (2001).

We simulated 40 nuclear families each with 5 sibs. Phenotype data are generated assuming there is no main genetic effect at all under the null. For the univariate trait analysis (Model I), data are simulated with the variance of polygene effect defined as σ^2_g=2, environmental error set as σ^2_e=2.5, and 1000 replicates are recorded. For the bivariate model, data are simulated based on the parameters of polygene effect and random residual effect given by: $\begin{pmatrix} \sigma^2_{g_1} & \sigma^2_{g_{12}} \\ \sigma^2_{g_{12}} & \sigma^2_{g_2} \end{pmatrix}=\begin{pmatrix} 2 & 1.8 \\ 1.8 & 2 \end{pmatrix}$, and $\begin{pmatrix} \sigma^2_{e_1} & \sigma^2_{e_{12}} \\ \sigma^2_{e_{12}} & \sigma^2_{e_2} \end{pmatrix}=\begin{pmatrix} 2.5 & 2.2 \\ 2.2 & 2.5 \end{pmatrix}$.

The performance of the approach is illustrated at several critical values in Table 4.3. It is clear that the type I error rates with the new method are much closer to the corresponding nominal level than those of the other methods. A quantile plot of the results are shown in Figure 4.3. The current approximation method shows the best approximation for the three models.

4.5 Conclusion

The new threshold determination method provides better approximation to the distribution of the LRT under the three models evaluated. These three models represent the most widely applied models in genetic linkage analysis. We expect the new method

Table 4.3: Comparisons of the performance of different approximation methods based on 1000 simulation replicates under different models.

Model	Method	critical value			
		α=0.1	α=0.05	α=0.01	α=0.005
Model I	SF	0.063	0.032	0.005	0.002
	New	0.093	0.051	0.008	0.005
Model II	Amos	0.182	0.104	0.027	0.014
	New	0.097	0.059	0.016	0.005
Model III	Amos	0.0839	0.0462	0.0158	0.0036
	New	0.0912	0.0523	0.0158	0.0049

(SF indicates the approximation is done with the result in Self and Liang (1987), i.e., $LR \sim \frac{1}{4}\chi^2_2 : \frac{1}{2}\chi^2_1 : \frac{1}{4}\chi^2_0$; New refers to the approximation by the current method; Amos refers to the approximation given in Amos (2001), i.e., $LR \sim \frac{1}{4}\chi^2_3 : \frac{1}{2}\chi^2_1 : \frac{1}{4}\chi^2_0$ for Model II, and $LR \sim \frac{1}{16}\chi^2_6 : \frac{4}{16}\chi^2_4 : \frac{2}{16}\chi^2_3 : \frac{4}{16}\chi^2_2 : \frac{4}{16}\chi^2_1 : \frac{1}{16}\chi^2_0$ for Model III.)

can reduce false positives in determining a linkage signal, hence reduce the cost of unnecessary investigations in a lab condition due to false results. This work represents the most comprehensive evaluation of the LRT in linkage analysis with the variance components model.

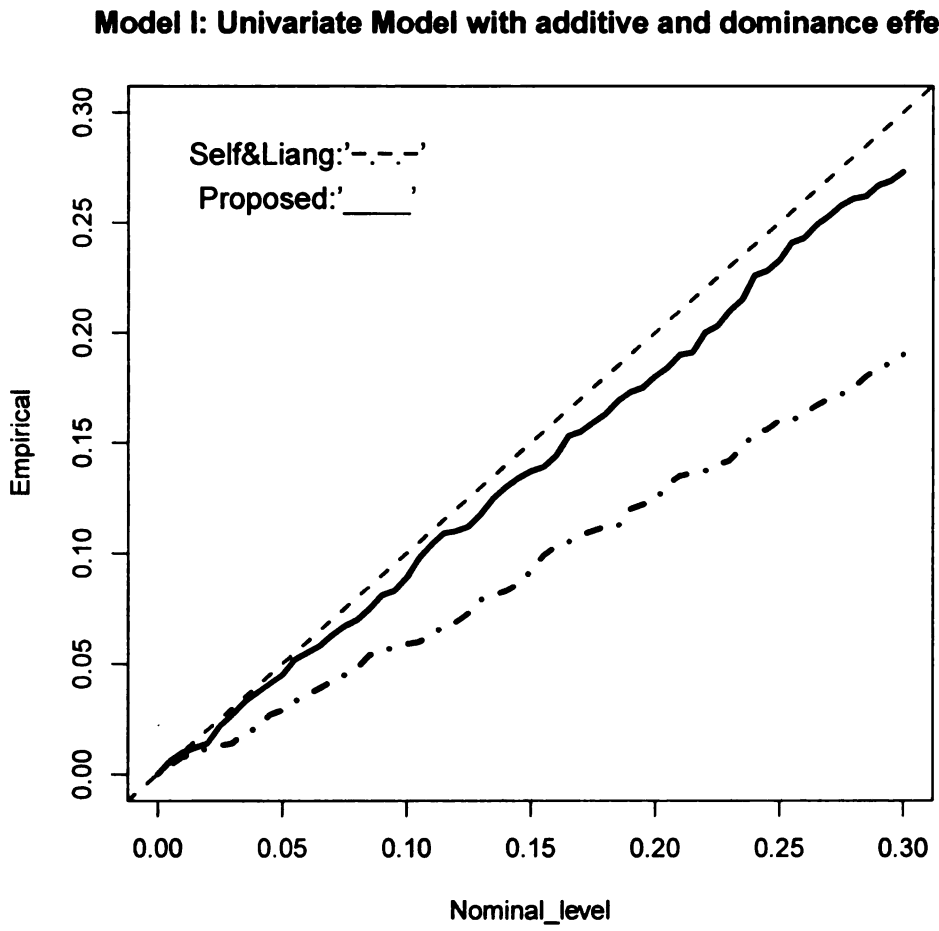

Figure 4.1: The quantile plot of the empirical p-values for Model I. For the legend: Self & Liang refers to SF; Proposed refers to the current method. See Table 4.3 for more explanation of the legend.

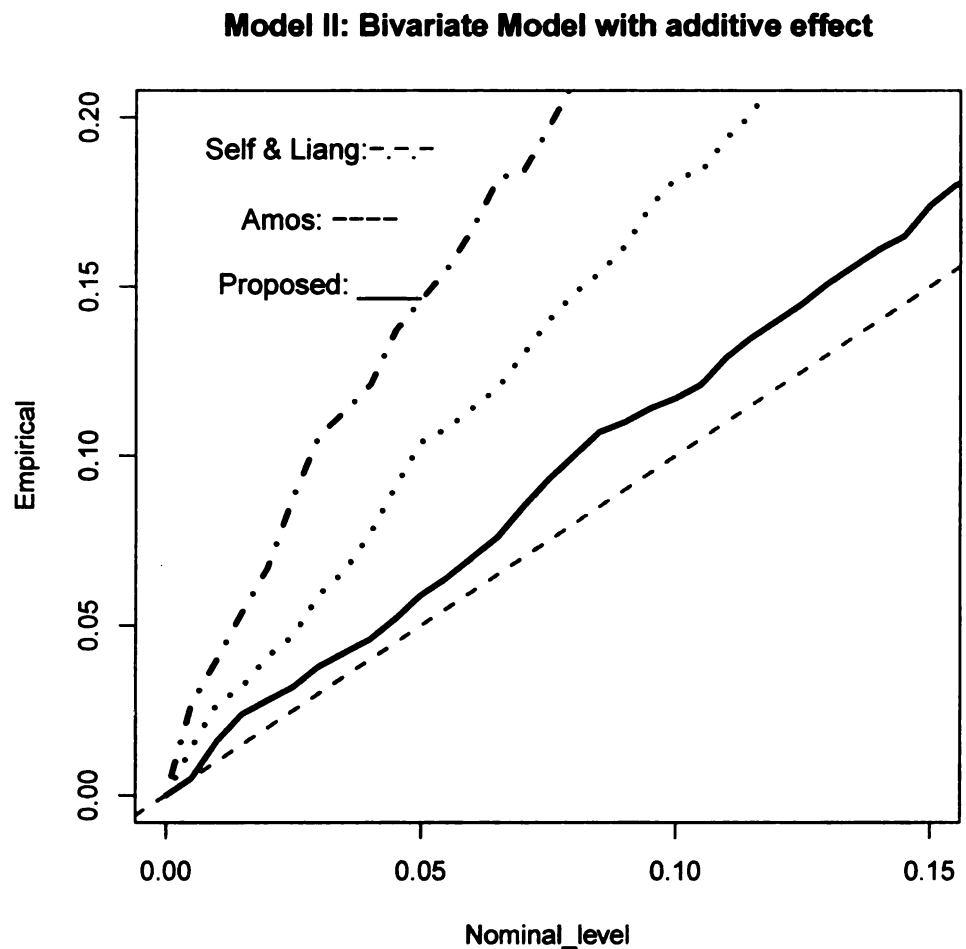

Figure 4.2: The quantile plot of the empirical p-values for Model II. For the legend: Self & Liang refers to SF; Proposed refers to the current method. See Table 4.3 for more explanation of the legend.

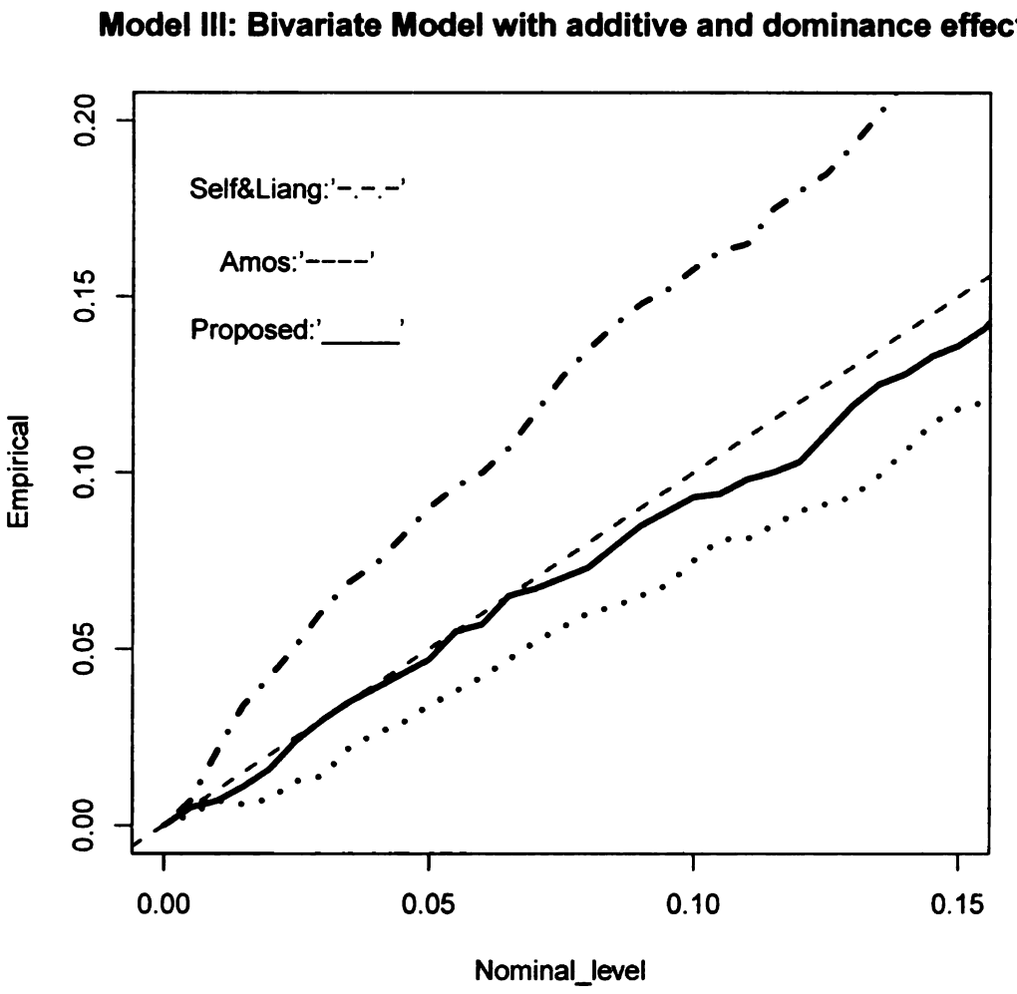

Figure 4.3: The quantile plot of the empirical p-values for Model III. For the legend: Self & Liang refers to SF; Proposed refers to the current method. See Table 4.3 for more explanation of the legend.

Chapter 5

Concluding remarks

Genomic imprinting, a unique phenomenon in multicellular organisms, is carried out in a regulated way that generally confers advantages during an organism's life cycle. Its role in controlling embryonic development and growth is not only restricted in humans and animals, but also in flowering plants. The information about how genes controlling or affecting this process is crucial for unravelling the genetic basis of many quantitative traits, which can not be explained by the traditional Mendelian inheritance theory. The identification of imprinted genes has been one of the most important and difficult tasks for genomic imprinting study. While many scientists are trying to experimentally unravel the molecular mechanism of genomic imprinting, identifying imprinting genes with statistical QTL mapping techniques is still in its infancy and therefore is in much demanding. With the abundant molecular marker information, it is now possible to detect potential imprinted genes underlying the quantitative variation of an imprinting trait. We, for the first time, developed a

series of statistical models and algorithms for detecting and characterizing specific iQTLs that are responsible for genomic imprinting under various problem settings. The developed models can make a systematic scan of iQTLs across the entire genome with a well-covered genetic linkage map.

Focusing on flowering plants, in this dissertation, I developed a series of statistical methods based on the variance components model in linkage analysis. Specifically, in chapter 2, I developed an efficient mapping approach focusing on a diploid mapping population (e.g., embryo in plants). We focused our genetic design on a reciprocal backcross design with experimental crosses. Different line crosses were combined to infer the random allelic effects under the variance components model. We partitioned the additive genetic effect into different components based on the nature of the allelic-sharing mechanism in experimental crosses. In chapter 3, we extended the idea to a triploid endosperm mapping population. The unique triploid structure in an endosperm tissue was considered. The utility of the method was demonstrated with a real data set. Important iQTLs were identified to control the endosperm development. Genomic imprinting can be explained by the genetic conflict theory proposed by Haig and Westoby (1991). Our real data analysis results are in consistent with and supported by this theory. In both chapters, we extended the single iQTL model to consider multiple iQTLs (i.e., multiple iQTL model). The multiple iQTL model can efficiently handle the problem due to the interfering of linked iQTLs on the same linkage group. Moreover, it also shows increased mapping precision as shown in the simulations studies.

When strong genetic correlations among multivariate traits occur in the QTL mapping, the multivariate analysis can largely improve the statistical power and accurate position of the genetic effect (Boomsma and Dolan 1998; Jiang and Zeng, 1995; Amos et al. 2001; Evans 2002). This motivates us to develop a multivariate iQTL mapping model, which is studied in chapter 3. Extensive simulation studies show the relative merit of multivariate analysis, especially when traits are correlated. In a multivariate linkage analysis, we also gain additional benefit by statistically quantifying pleiotropic vs close linkage effect. The real data analysis indicates that two QTLs express strong pleiotropic effect to control the two endosperm traits used in this study.

In a variance components-based linkage analysis, the likelihood ratio test has been the standard means in assessing the statistical significance of a linkage signal. However, due to irregular conditions (e.g., the restriction of the variance component terms under the hypothesis), the regular asymptotic chi-square distribution theory does not apply directly. In chapter 4, we conducted a statistical investigation of the LRT, and found that the currently applied cutoff determination method is inappropriate. This finding is in consistent with an empirical study (Allison et al., 1999). We evaluated the limiting distribution of the LRT under three model settings which are the mostly used models in a linkage analysis. Simulation study shows the superiority of the new approximation method over the currently applied ones.

Other statistical issues such as deriving the optimization algorithms for parameter estimation, and proof of the theorems have been given. Coupling with the emergence

of abundant marker information, large collection of well-phenotyped samples and high-throughput genotyping, our models provide a quantitative testable framework to assess genome-wide significance of imprinted genes. The developed models also provide a testable platform for scientists who can design their experiment accordingly and significant discoveries would be expected in the future. This dissertation contributes to the statistical methodology development in QTL mapping, to the general statistical theory in variance components model, and to the general genetic mapping community by providing statistically sound approaches and tested programs.

BIBLIOGRAPHY

Abney, M., Sara McPeek, M. and Ober, C. (2000). Estimation of variance components of quantitative traits in inbred populations. *Am. J. Hum. Genet.* 66(2): 629-650.

Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I. and Blangero, J. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.* 65, 531-544.

Almasy, L and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62, 1198-1211.

Almasy, L., Dyer, T. D., and Blangero, J. (1997). Bivariate Quantitative Trait Linkage Analysis: Pleiotropy Versus Co-incident Linkages. *Genet Epidemiol.* 14(6):953-958.

Amos, C.I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* 54, 535-543.

Amos, C. and Andrade, M. (2001). Genetic linkage methods for quantitative traits. *Stat. Methods. Med. Res.* 10: 325.

Amos, C.I. de Andrade, M and Zhu, K.D. (2001). Comparison of Multivariate Tests for Genetic Linkage. *Hum. Hered.* 51: 133-144.

Bartholomem, D.J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika.* 46, 35-48.

Bartholomem, D.J. (1959b). A test of homogeneity for ordered alternatives II. *Biometrika.* 46, 328-35.

Bartholomem, D.J. (1961a). A test of homogeneity of means under restricted alternatives. *J.R.Statist.Soc.* B,23,239-81.

Bohrer, R. and Chow, W. (1978). Weights for one-sided multivariate inference. *Appl. Statist.* 27, 100-104.

Bomblies, K. and Doebley, J.F. (2006). Pleiotropic effects of the Duplicate Maize FLORICAULA/LEAFY Genes zfl1 and zfl2 on Traits Under Selection During Maize Domestication. *Genetics.* 172: 519531.

Boomsma, D.I. and Dolan, C.V. (1998). A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores. *Behav Genet* 28: 329-340.

Brink, R. A. and Cooper, D. C. (1947). The endosperm in seed development. *Bot. Rev.* 13, 423-541.

Burt, A. and Trivers, R.L. (2006). *Genes in Conflict*, Harvard University Press, Cambridge, MA, USA.

Chant. D. (1974). On Asymptotic Tests of Composite Hypotheses in Nonstandard Conditions. *Biometrika.* 61, 291-298.

Chaudhury, A.M., Koltunow, A., Payne, T., Luo, M., Tucker, M.R., Dennis, E.S. and Peacock, W.J. (2001). Control of early seed development. *Ann. Rew. Cell Dev. Biol.* 17: 677-699.

Chaudhuri, S. and Messing, J. (1994). Allele-specific parental imprinting of *dzrl*, a post transcriptional regulator of zein accumulation. *Proc. Natl. Acad. Sci.* 91: 4867-4871.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Stat.* 25: 573-578.

Churchill, G.A. and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics.* 138: 963-971.

Clark, R.M, Wagler, T.N., Quijada, P. and Doebley, J. (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet.* doi:10.1038/ng1784.

Cockerham, C.C. (1983) Covariances of relatives from self-fertilization. *Crop. Sci.* 23: 11771180.

Corbeil, R.R. and Searle, S.R. (1976). A comparison of variance component estimators *Biometrics.* 32: 779-791.

Cui, Y.H., Cheverud, J.M. and Wu, R.L. (2007). A statistical model for dissecting genomic imprinting through genetic mapping, *Genetica.* 130, 227-239.

Cui, Y.H. (2007). A statistical framework for genome-wide scanning and testing imprinted quantitative trait loci. *J. Theo. Biol.* 244: 115-126 .

Cui, Y.H., Lu, Q., Cheverud, J.M., Littel, R.L. and Wu, R.L. (2006). Model for mapping imprinted quantitative trait loci in an inbred F_2 design. *Genomics.* 87: 543-551.

Cui, Y.H. and Wu, R.L. (2005). A statistical model for characterizing epistatic control of triploid endosperm triggered by maternal and offspring QTL. *Genet. Res.* 86: 65-76.

David,.F.N. (1953). A note on the evaluation of the multivariate normal integral. *Biometrika.* 40, 458-9.

DeChiara,T.M., Robertson, E.J. and Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. *Cell.* 64, 849-859.

Dilkes, B.P., Dante, R.A., Coelho, C. and Larkins, B.A. (2002). Genetic analysis of endoreduplication in Zea mays endosperm: evidence of sporophytic and zygotic maternal control. *Genetics.* 160: 1163-1177.

Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics.* 151, 373-386.

Evans, D.M. (2002). The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between the variables. *Am. J. Hum. Genet.* 70: 1599-1602.

Feil, R and Berger, F. (2007). Convergent evolution of genomic imprinting in plants and mammals. *Trends. Genet.* 23, 192-199.

Grime, J.P. and Mowforth, M.A. (1982). Variation in genome size: an ecological interpretation. *Nature.* 299: 151-153.

Grossniklaus, U., Spillane, C., Page, D.R., and Koehler, C. (2001). Genomic imprinting and seed development: Endosperm formation with and without sex. *Curr. Opin. Plant Biol.* 4: 2127.

Hanson, R.L., Kobes, S., Lindsay, R.S. and Kmowler, W.C. (2001). Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *Am. J. Hum. Genet.* 68(4): 951-962.

Haig, D. and Westoby, M. (1991). Genomic Imprinting in endosperm: Its effect on seed development in crosses between species, and between different ploidies of the same species, and its implications for the evolution of apomixis. *Philos. Trans. R. Soc. Lond.* 333: 1-13.

Harris, D.L. (1964). Genotypic covariances between inbred relatives. *Genetics.* 50: 1319-1348.

Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics.* 135, 205-211.

Jansen, R.C. (1994). Controlling the Type I and Type I1 errors in mapping quantitative trait loci. *Genetics.* 138, 871-881.

Jiang, C. and Zeng, Z-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics.* 140: 1111-1127.

Kao, C.H., Zeng, Z-B., and Teasdale, R.D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics.* 152, 1203-1216.

Kendall, M.G. (1941). Proof of Relations connected with the Tetrachoric Series and its Generalization. *Biometrika.* 32, 196-8.

Kermicle, J.L. (1970). Dependence of the *R*-mottled aleurone phenotype in maize on the modes of sexual transmission. *Genetics.* 66: 69-85.

Kinoshita, K., Yadegari, M., Harada, J.J., Goldberg, R.B. and Fishcher, R.L. (1999). Imprinting of the *MEDEA polycomb* gene in the *Arabidopsis* endosperm. *Pla. Cell.* 11: 1945-1952.

Knott, S.A., Marklund, L., Haley, C.S., Andersson, K., Davies, W., Ellegren, H., Fredholm, M., Hansson, I., Hoyheim, B., Lundstrm, K., Moller, M. and Andersson, L. (1998). Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics.* 149, 1069-1080.

de Koning, D-J., Rattink, A.P., Harlizius, B., van Arendonk, J.A.M., Brascamp, E.W. *et al.* (2000). Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proc. Natl. Acad. Sci.* USA 97: 7947-7950.

de Koning, D-J., Bovenhuis, H. and van Arendonk, J.A.M. (2002). On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics.* 161(2): 931-938.

Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika.* 50,

403-18.

Kudô, A and Choi, J.R. (1975). A generalized multivariate analogue of the one sided test. *Mem.Fac.Sci.,Kyushu Univ.* 29, 303-28.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics.* 121, 185-199.

Li, Y.C, Coelho, C.M, Wu, S, Zeng, Y.R, Li, Y, Hunter, B, Dante, R.A, Larkins, B.A and Wu, R. (2008). A statistical model for estimating maternal-zygotic interactions and parent-of-origin effects of QTLs for seed development. *PLoS ONE.* 3, e3131.

Li, G.X. and Cui, Y.H. (2009). A statistical variance components framework for mapping imprinted quantitative trait loci in experimental crosses. *J. Prob. Stat.* Article ID 689489, doi:10.1155/2009/689489.

Li, G.X and Cui, Y.H. (2010). A general statistical framwork for dissecting parent-of-origin effects underlying endosperm traits in flowering plants. *Ann. App. Stat.*

Lin, M., Lou, X.Y., Chang, M. and Wu, R. (2003). A general statistical framework for mapping quantitative trait loci in nonmodel systems: issue for characterizing linkage phases. *Genetics.* 165, 901-913.

Liu, T., Todhunter, R.J., Wu, S., Hou, W., Mateescu, R., Zhang, Z., Burton-Wurster, N.I., Acland, G.M., Lust,G. and Wu,R . (2007). A random model for mapping imprinted quantitative trait loci in a structured pedigree: an implication for mapping canine hip dysplasia. *Genomics.* 90, 276-284.

Lund, G., Messing, J. and Viotti, A. (1995). Endosperm-specific demethylation and activation of specific alleles of *alpha*-tubulin genes of *Zea mays* L. *Mol. Gen. Genet.* 246: 716-722.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sinauer, Sunderland, MA, USA.

Malécot, G. (1948) *Les mathématiques del'hérédité.* Masson et Cie, Paris, France.

Martinez, O. and Curnow, R.N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet.* 85, 480-488.

Moran, P.A.P. (1971). The uniform consistency of maximum-likelihood estimators. *Proc.Cambridge Philos.Soc.* 70, 435-439.

Moran, P.A.P. (1971). Maximum Likelihood Estimators in Non-Standard Conditions. *Proc. Camb. Phil. Soc.* 70, 441-450.

Moran, P.A.P. (1948). Rank correlation and product-moment correlation. *Biometrika.* 35, 203-6.

Moose, S.P. and Sisco, P.H. (1996). Glossyl5, an APETALA2-1ike gene from maize that regulates leaf epidermal cell identity. *Genes. Develop.* 10:3018-3027.

Nuesch, P. E. (1966). On the problem of testing location in multivariate populations for restricted alternatives. *Ann.Math.Statist.* 37, 113-9.

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

Pfeifer, K. (2000). Mechanisms of genomic imprinting. *Am. J. Hum. Genet.* 67: 777-787.

Plackett,R.L. (1954). A reduction formula for normal multivariate integrals. *Biometrika.* 41, 351-60.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* 82, 605-610.

Shapiro, A. (1985a). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika.* 72, 133-144.

Shapiro, A. (1988). Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis. *Internat. Statist. Rev.* 56, 49-62.

Shete, S., Zhou, X. and Amos, C.I. (2003). Genomic imprinting and linkage test for quantitative trait loci in extended pedigrees. *Am. J. Hum. Genet.* 73: 933-938.

Sturaro, M., Hartings, H., Schmelzer, E., Velasco, R., Salamini, F. and Motto, M. (2005). Cloning and Characterization of GLOSSY1, a Maize Gene Involved in Cuticle Membrane and Wax Production. *Plant Physiol.* 138, 478489.

Vega, S. H., Sauer, M., Orkwiszewski, J.A.J. and Poeth, R.S. (2002). The early phase change Gene in Maize. *Plant Cell.* 14, 133147.

Xie, C, Gessler, D.D.G and Xu, S. (1998). Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics.* 149, 1139-1146.

Xu, S. and Atchley, W.R. (1995). A random model approach to interval mapping of quantitative trait loci, *Genetics.* 141, 1189-1197.

Wilks, S.S. (1938). The large distribution of the likelihood ratio for testing composite hypotheses. *Ann.Math.Stat.* 9, 60.

Williams, J.T., Eerdewegh, P.V. Almasy, L. and Blangero, J. (1999). Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. likelihood formulation and simulation results. *Am. J. Hum. Genet.* 65, 1134-1147.

Wolf, J., Cheverud, J., Roseman, C. and Hager, R. (2008). Genome-wide analysis reveals a complex pattern of genomic imprinting in mice. *PLoS Genetics.* 4, doi: 10.1371/journal.pgen.1000091.

Wu, R.L, Casella, G, Ma, C-X. (2007). *Statistical Genetics of Quantitative Traits: Linkage, Maps and Qtl.* Springer, New York, USA.

Zeng, Z-B. (1994). Precision mapping of quantitative trait loci. *Genetics.* 136, 1457-1468.

Zeng, Z-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Nat. Acad. Sci.* USA, 90, 10972-10976.