Abstract

LIU, WENLEI. DEVELOPMENT OF LINKAGE AND ASSOCIATION METHODS TO MAP DISEASE GENES. (Advisor: Dr. Bruce S. Weir)

Identification of disease susceptibility genes is one of the primary aims of contemporary genetic research. With the recent development in molecular biology techniques, large-scale gene mapping with a dense genome-spanning set of markers becomes a reality. The availability of markers throughout the genome has made linkage and association studies more feasible. In the first chapter, we review many linkage and association methods and point out the potential problems with current linkage and association analysis. In the second chapter, we modify two identity-by-state (IBS) test statistics of Lange (Lange K. 1986a, A test statistic for the affected-sib-set method. *Annals of Human Genetics* **50,** 283–290; Lange K. 1986b, The affected sib-pair method using identity by descent relations. *American Journal of Human Genetics* **39,** 148–150.) to allow for inbreeding in the population. We evaluate the power and false positive rates of the modified tests under three disease models using simulated data. When the population inbreeding coefficient is large, both the false

positive rates and power are reduced when the modified test statistics were applied, although power remained high under a recessive disease model. Allowing for inbreeding is therefore appropriate at least for diseases known to be recessive. In the third chapter, we compute the proportions of affected sib pairs sharing 0, 1 and 2 marker alleles identity-by-decent (IBD) in an inbred population and express them in terms of higher order decent measures. We perform two consistency checks on the identity state probabilities and the two consistency checks verify our calculations. We did the same thing for affected sib pairs from first cousin marriage in an inbred population. In the fourth chapter, we study linkage and linkage disequilibrium (LD) simultaneously for single QTL using family data in an attempt to increase mapping resolution and reduce false positive rates. We estimate QTL allele frequencies, LD and recombination factions between the marker loci and the QTL locus and the QTL model parameters using an EM algorithm. After performing single analysis, we extend our model to study two marker loci simultaneously so that we can increase the accuracy of the estimations. Our simulation results show that our EM algorithm can give consistent estimates of all the parameters considered.

DEVELOPMENT OF LINKAGE AND ASSOCIATION METHODS TO MAP DISEASE GENES

by
WENLEI LIU

A dissertation submitted to the Graduate Faculty of

North Carolina State University

in partial fulfillment of the

requirements for the Degree of

Doctor of Philosophy

BIOINFORMATICS

Raleigh

2002

APPROVED BY:

Bruce S. Weir CHAIR OF ADVISORY COMMITTEE	Zhao-Bang Zeng
Dahlia M. Nielsen	Gregory C. Gibson

Dedication

To my husband, parents, and sister

Biography

Wenlei Liu was born in 1977, in Jinan, Shangdong Province, the People's Republic of China. In 1983, she moved to Tianjin with her parents and grew up there. In June 1998, she received her B.S. in Biochemistry from Jilin University in China. She went to the Genetics Department at North Carolina State University in August 1998. She worked there as a R.A. for one year. During that time, she met Wei Zhao, a student in North Carolina State University. In August 1999, she switched to Bioinformatics Program at North Carolina State University to pursue a Ph.D. degree and became the first student in that program. She got married with Wei Zhao during the summer of 2000.

Acknowledgements

First, I would like to thank all my committee members for the kind help they gave me, especially Dr. Bruce S. Weir and Dr. Zhao-Bang Zeng. They gave me lots of guidance and encouragement during my research. Second, I would like to thank Dr. Dennis Boos for his helpful discussions. Third, I would like to thank my friends, Dr. Tao Wang, Dr. Shupang Huang, Dr. Kuejun Xin, Wenli Tao and Kejun Liu for their kind help. Finally, I would like to thank my husband, parents and sister for their continuous support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Review of linkage and association Methods

1.1 Introduction

In the 20th century geneticists began to unravel some aspects of inherited human diseases and try to find the disease genes. The primary tools for gene discovery are linkage and association studies. With the development in recent decades of molecular biology techniques, large-scale gene mapping with a dense genome-spanning set of markers became a reality. The availability of markers throughout the genome has made linkage and association studies more feasible. Markers are loci of DNA sequences that are polymorphic between individuals. These loci are normally known sequences and the variation can be easily detected. One important method for locating genes that influence a particular disease is to study the possible genetic linkage between a putative disease susceptibility locus and a marker locus. Linkage refers to two loci being physically close to each other on a chromosome. Showing that a marker locus is not transmitted randomly in affected individuals implies that the marker locus is physically close to the disease gene locus on the chromosome. Identifying markers

that are linked to a disease locus is the initial step to locating disease gene. Molecular biology methods are needed to further refine the position of the disease locus on the chromosome.

Affected sib-pair (ASP) methods were first established by Penrose (1953) and have been among the most popular linkage methods to detect disease susceptibility genes since then. The basis of ASP methods is that affected individuals should show greater than expected concordance at markers because affected sib-pairs are likely to have inherited the same marker allele if the marker locus is linked to the disease locus. The most widely used measure of marker concordance of two siblings is the number of alleles they share identical by descent (IBD). In a non-inbreeding population, and in the absence of linkage, the expected proportions of sib-pairs with 0, 1 and 2 marker allele IBD are 0.25, 0.5 and 0.25, respectively. Affected sibs will share marker alleles IBD more often than expected by chance if there is linkage between the marker locus and the disease locus (Haseman and Elston 1972; Day and Simons 1976; Suarez 1978). This is because affected siblings will share disease alleles at the disease locus and share common alleles at nearby marker loci. The simplest approach for detecting deviations from the expected IBD proportions is to the count the number of sib pairs sharing 0, 1, and 2 alleles IBD and to compare these numbers with the expected numbers under the null hypothesis of no linkage using a usual chi-square goodness of fit test (Cudworth and Woodrow, 1975). Many other test statistics have been proposed, such as the H-E method (Haseman and Elston, 1972) in which the squared sib-pair trait differences are regressed on the proportion of IBD marker alleles; the two-allele test,

which is based on the proportion of sib pairs with two marker alleles IBD (Day and Simons, 1976; Suarez et al., 1978; Weitkamp et al., 1981); and the mean test, which is based on the mean number of marker alleles IBD (Blackwelder and Robert, 1985; Knapp M, 1994a). Methods have also been established for analyzing markers that are not highly polymorphic (Risch, 1990; Holmans, 1993) and families with more than two affected sibs (Green et al., 1983; Payami et al., 1985). Studying discordant sib-pairs enables us to detect linkage (Khoury et al., 1991) and typing unaffected siblings may substantially increase power if there is assortative mating (Sribney and Swift, 1992). Recent developments in ASP methods (Davies et al., 1994; Schwab et al., 1995; Weeks and Lathrop 1995; Field et al., 1996; Goldgar and Easton, 1997; Juo et al., 1998; Nemesure et al., 1999; Li and Reich, 2000) have led to an explosion in the number of possible statistics that people use. The power of many ASP methods were investigated by Davis and Weeks (1997). They found that the ASP mean test (Knapp et al., 1994a) is generally most powerful and theoretically sound and performs well on a variety of disease models.

Linkage analysis studies the marker segregation pattern in affected individuals through a pedigree, while association analysis, or linkage disequilibrium mapping, measures deviation from the random occurrence of alleles in a haplotype in unrelated individuals or nuclear families. Linkage disequilibrium (LD) refers to association between alleles at different loci. We call alleles at two different loci associated alleles if the haplotype frequency differs from the product of the two allele frequencies.

Association studies are an important complement to linkage analysis. The case-control test was one of the first established tests of association. The classic case control study design compares allele frequencies in a sample of unrelated affected individuals, the cases, and a sample of unrelated unaffected individuals, the controls (Terwilliger and Ott 1994, chap.24). In the simplest case of diallelic loci, a simple chi-square test of a 2×2 contingency table can be performed. Many population-based association methods have been developed recently. Hastbacka et al. (1992) performed linkage disequilibrium mapping to map the diastrophic dysplasia (DTD) gene in an isolated Finland population. They also estimated the recombination fraction between a marker locus and a disease locus and the mutation rate of marker loci by using Luria and Delbruck's classic analysis (1943). Kaplan et al. (1995) proposed a likelihood-based association method to locate disease genes in non equilibrium populations. They modeled the initial growth phase of the disease with a Poison branching process and estimated the recombination fraction between marker and disease loci based on a simulated disease population. Terwilliger (1995) also constructed a likelihood-ratio test, which has only 1 df, to test for linkage disequilibrium. His approach can be applied to multiallelic marker systems and extended to multiple marker loci simultaneously. It maintains higher power than the conventional case control test.

1.2 IBS tests

Most ASP tests require clearly determined marker identity by descent state. In many cases, the parents may be unavailable for typing, especially for diseases with late onset.

Even when parents are available, the markers may be incompletely polymorphic and thus the IBD status of the sib pairs can not be inferred unequivocally. This will cause trouble in ASP tests which are based on testing IBD sharing. These problems can be avoided if we relax IBD relations by substituting identity by state (IBS) relations in sib pairs. IBS refers to two alleles with same allelic form. IBS does not consider the ancestor of the alleles. Two alleles identical by descent are also identical by state, but two alleles identical by state may or may not be identical by descent. For highly polymorphic marker system, IBS relations correspond fairly closely to IBD relations. IBS relations are easy to determine and do not require parental information or highly polymorphic marker systems. Lange first proposed the IBS tests to detect linkage using affected sib pairs (1986a, b). He computed the expected proportions of the three IBS categories and proposed to test deviations from the expected proportions with a chi-square goodnees of fit test (1986a). He also constructed a test statistic based on the Central Limit Theorem which could analyze families with arbitrary numbers of affected sibs (1986b). Later, Weeks and Lange generalized the IBS tests to pedigrees and developed an affected-pedigree-member (APM) method (1988). They modified the original Lange test (1986b) by recomputing the theoretical mean and variance for each pedigree and then introduced a weighting factor based on marker allele frequency. Bishop and Williamson (1990) further generalized the method to more distant degrees of relationship and examined the power of a chi-square goodness-of-fit IBS test. They also studied the effect of several factors (the polymorphism of the marker, the distance between loci, the mode of inheritance and the genetic relationship

of the affected individuals) on detecting linkage. The APM method statistic has also been generalized to multiple linked markers (Weeks and Lange, 1992). Later, Weeks et al. (1995) developed a X-linked version of the APM method which can be applied to test for linkage of complex diseases to X-linked markers.

IBS methods are generally less powerful than IBD methods, especially when parental data are available. The power of IBS tests was studied by Bishop and Williamson (1990) and they suggested that power depended on the polymorphism of the marker, the probability of identity by descent at the trait locus, and the recombination fraction between the trait and marker loci. Thomson and Motro (1994) studied the statistical power to detect linkage between a diallelic marker and disease for four IBS tests. They found that two of their tests have undesirable power and the other two tests have more power, especially when there is linkage disequilibrium. Nemesure et al. (1999) constructed a normalized identity by state statistic, which they claimed to give result similar to those obtained by the traditional IBD approach, when most of the pedigrees had at least one homozygous parent or two parents sharing a common allele.

1.3 Model-based linkage analysis

A model-free method (also called a nonparametric method) requires no information about the underlying genetic mechanism of the disease. Most ASP methods are model-free linkage analyses. Actually, one of the main advantage of ASP methods is that there is no need to know the mode of inheritance of the disease. A model-based

method requires the specification of the underlying trait distribution and is based on a likelihood ratio, the logarithm of which is called a lod score. Likelihood ratio tests compute the logarithm of likelihood ratio of the probability of data under a particular value of recombination fraction versus the probability of data under the null hypothesis of no linkage, ie recombination fraction equals 0.5, and reject the null hypothesis if the resulting lod score exceeds the cutoff value. When the mode of inheritance of a disease is well established, lod score linkage analysis will most often be the method of choice in detecting disease susceptibility genes, because likelihood ratio tests will usually yield more power compared to model-free tests. The other advantage of lod score methods is that they can give an estimate of the recombination fraction in most cases. Morton (1955) first proposed a lod score method to detect linkage and numerous modifications have been made since then. Demenais and Amos (1989) showed that a lod score analysis was more powerful than the H-E sib pair linkage analysis (Haseman and Elston, 1972) for a single locus quantitative trait in a nuclear family. Goldin and Weeks (1993) also compared the likelihood method to several nonparametric methods and showed the nonparametric linkage methods had somewhat lower power than does the lod score method. The likelihood ratio method proposed by Risch (1990) can test linkage using IBD information when the marker is not 100% polymorphic. He used a maximum-likelihood method (1989) to estimate the IBD probabilities. A "maximum" lod score (MLS) method is performed to test linkage using affected relative pairs. The main difference between the MLS method and Bishop and Williamson's IBS method is that MLS method take into

account not only how many markers an affected relative pair share but also which ones they share (e.g. they will count a pair of identical homozygotes and a pair of identical heterozygotes separately). Risch's method makes more use of the available marker sharing information and thus should give more power. The "possible triangle" method introduced by Holmans (1993) improved the power of Risch's test, especially for low polymorphism markers, by restricting maximization to a set of IBD probabilities consistent with possible genetic models. Cordell et al. extended the lod score method to be applicable to X-chromosomal data (1995a) and allows the simultaneous detection and modeling of two unlinked disease loci (1995b). Farrall (1997) proposed another likelihood method that can test linkage to a second putative susceptibility gene that happens to map close to an established susceptibility gene. Many other likelihood based linkage methods have been developed recently (Olson, 1997; Knapp 1998). Weeks et al. (1990) and Clerget-Darpoux et al. (1990) studied lod score over alternative segregation models on realistic pedigrees and both found maximizing lod score over many models leads to an increase of evidence for linkage. Knapp et al. (1994b) showed that a lod score analysis for an assumed recessive mode of inheritance, irrespective of the true mode of the disease, is equivalent to the mean test. Their work has been extended to cope with samples with locus heterogeneity (Huang and Vieland, 2001).

Despite the wide use of lod score analysis, there has been much discussion concerning what constitutes a significant lod score. Historically, a base 10 logarithm has been used. Morton (1955) suggested that a lod score of 3 corresponds to a type I

error of 10^{-3}. This suggestion was adopted by many people. However, there is some difficulty in assessing the statistical significance base on lod score of 3. The actual significance of a lod scores depends on how many parameters are estimated and how many tests are performed. Holmans (1993) derived asymptotic distributions for his likelihood-ratio test statistic, enabling test criteria to be found for any required test size and enabling p values to be assigned to results. Maclean et al. (1993) also studied the distribution of lod statistics under an uncertain mode of inheritance in small samples which were mostly composed of affected individuals and found the resulting distribution was a χ^2 distribution.

1.4 Comparison between ASP tests and TDT tests

Linkage may cause association. However, association may have nothing to do with linkage. Many factors, such as population admixture, migration and selection may lead to association between alleles at two well separated loci. Although the case control test is a valid test of association in theory, it could be difficult to apply the test. When there is population stratification, the classic case control study design may detect association between a disease allele and a genetic marker allele, even in the absence of linkage. Population stratification may exist when the cases and controls are not well matched ethnically or the mating is not random. Falk and Rubinstein (1987) first tried to overcome the problem of population stratification by constructing a haplotype relative risk (HRR) test. They used the parental haplotypes present in the affected child as cases and parental haplotypes not present in the affected child as

controls. Now the cases and controls are well matched and the problem of population stratification is eliminated. The haplotype-based haplotype relative risk (HHRR) test proposed by Terwilliger and Ott (1992) studies haplotype-based data rather than genotypic data and gives much higher power than the original HRR test. Another association-based test is the transmission/disequilibrium test (TDT) (Spielman et al., 1993). TDT is also based on the HRR statistic, but TDT studies transmissions from only the heterozygous parents. TDT is not affected by population stratification either.

The affected sib-pair test and transmission/disequilibrium test (TDT) are two of the most widely used tests in linkage and association analysis. ASP methods study the possible genetic linkage between a putative disease susceptibility locus and one or more marker loci by demonstration of nonrandom segregation of parental alleles in affected children within families. The TDT test was originally developed as a test for linkage in the presence of association. Now, the TDT test has been widely used as a test for association in the presence of linkage. The TDT test statistic is based on the HRR statistic, which counts parental marker alleles (or haplotypes) transmitted to an affected child and compares them with those parental alleles not transmitted. Under the null hypothesis of no linkage the heterozygous parents should transmit the two marker alleles to affected offspring with equal probability. If there is linkage and association between marker and disease loci, the associated marker allele should be transmitted with higher probability compared to the other allele. The TDT test procedure compares the number of times that heterozygous parents

transmit the associated marker to an affected offspring with the number of times that they transmit the alternate marker allele.

Compared to ASP tests, the TDT has the advantage that it does not require data either on multiple affected family members or on unaffected sibs. The TDT just needs family trios - two parents and one affected offspring to conduct linkage analysis. It is not affected by the presence of population stratification, which may cause population association and produce false positive in some other linkage analysis. However, the TDT requires both linkage and association between marker and disease loci to detect linkage. Sufficient markers must be typed to ensure that one will be near enough to the disease locus. In contrast, ASP methods do not need linkage disequilibrium to detect linkage. Association analysis has better resolution than linkage analysis, because, ideally, only closely linked loci will show association. Thus, the TDT test is more appropriate for fine scale mapping. In addition, the TDT test can be more powerful than ASP tests even when the same data are used, especially for complex diseases (Risch and Merikangas, 1996). McGinnis (1998) compared the power of TDT and ASP test under different values of recombination fraction, disequilibrium, penetrance and disease allele frequency. He found that the TDT yields higher power than the ASP tests and the superior power of the TDT is greatest when susceptibility loci confer modest disease risk. The power of the TDT is greatly increased if there is strong disequilibrium between marker and disease locus and if the disease allele and positively associated marker allele have similar population frequencies. Although both TDT and ASP methods are used to detect linkage, they can not give estimates

of the recombination fraction between marker and disease locus. Thus, molecular biology methods are necessary to further refine the position of the disease locus.

1.5 Development of linkage methods

One of the earliest ASP methods developed by Haseman and Elston (1972) used logistic regression to compare IBD allele sharing in concordant and discordant sib pairs. ASP methods were initially restricted to sib pairs only. When there were more than two affected siblings in a family, randomly selecting one pair was suggested. However, this will lead to information loss and different results might be obtained when different pairs of sibs are selected. Many different schemes have been proposed that allow utilization of multiple siblings simultaneously. For instance, Ewens and Clarke (1984) proposed a maximum likelihood statistic that can analyze data from both affected and unaffected siblings in a family of any size and allows estimation of parameters associated with *HLA*-linked diseases. Other test statistics which can use multiple sibling information are also available (Suarez and Hodge, 1979; Green et al., 1983; Cockerham and Weir, 1983; Hodge, 1984; Sham, 1997). Recently, Abel et al. (1998) compared four methods that can analyze multiplex sibships and studied their powers and type I error rates. They found that the weighted and likelihood methods yielded high power and consistent type I error rates. They also showed theoretically that the likelihood method is expected to be more powerful than the classical mean test when the sibship has two affected siblings and a common asymptotic type I error is used.

Sometimes, unaffected relatives are available. Typing unaffected relatives can increase power to detect linkage. First, parents are sometimes unavailable and typing unaffected siblings or relatives will allow inference of the parental genotypes. Second, typing unaffected relatives allows detection of genotyping errors in the affected sib pairs. Third, typing unaffected relatives allows estimation of marker allele frequencies. Sribney and Swift (1992) showed that the sample sizes required for sib trios of affected and unaffected siblings are much smaller than the corresponding sample sizes for affected sib pairs, when there is assortative mating and multiple disease loci. Levinson (1993) studied the information gained by typing unaffected relatives under standard lod score linkage methods and found that unaffected siblings can provide information even when both parents are typed. Holmans and Clayton (1995) studied two likelihood based tests and found the power of a test using unaffected siblings to infer parental genotypes was more powerful than another test which was based on IBD sharing states of the entire sibship.

Many diseases follow clear Mendelian, single-locus segregation patterns. In contrast, many complex diseases, such as diabetes and psoriasis, do not exhibit simple Mendelian transmission. More than one locus might be involved in disease susceptibility. Although ASP methods initially aim to identify single disease loci, many new ASP methods have been proposed which can detect two or more disease loci simultaneously. Schork et al. (1993) developed a two-trait-locus, two-marker-locus linkage test and mapped two disease loci simultaneously. They found that two-trait-locus, two-marker-locus linkage analysis can provide substantially more linkage information

compared to the traditional one-trait-locus, one-marker-locus methods. Dizier et al. (1994) extended the marker-association-segregation χ^2 (MASC) method to model effect of two candidate genes simultaneously by using two markers linked to the two candidate genes. Knapp et al. (1994c) evaluated the behavior of different single-marker and two-marker tests under a wide range of two-trait-locus heterogeneity and epistatic model. They got very similar results to those of Schork et al. (1993). Additionally, they found that there are some restrictions in using the two-marker tests, such as the segregation model must be suitable and the second disease locus should have larger effect. Risch's maximum lod score method (1990) has also been extend to detect and model two unlinked disease loci simultaneously (Cordell et al., 1995b). Many other two locus disease linkage analysis are also available (Goldstein et al., 1996; Goldgar and Easton, 1997; Juo et al., 1998; Li and Reich, 2000).

The effect of non-genetic factors on linkage analysis, such as age, environment, and population stratification have been ignored until recently. Li and Hsu (2000) studied effects of incorporation of age at onset information on the power of ASP and TDT tests. They showed the power of both IBD test and TDT test can be greatly affected by the disease onset age of the offspring. Gauderman et al. (1999) proposed a test for linkage in the presence of gene-environment interaction and showed that their test had superior power in detecting linkage compared to the standard mean test. Mandal et al. (1999), Guo (2000), Gauderman and Siegmund (2001) also studied the effect of environment on power of several ASP methods. Gene-gene interaction has also been studied (Niu et al., 1999). Leal and Ott (2000) examined the benefits and costs of

stratifying sib-pair data under three different situations and found that the power of ASP methods to detect linkage will be affected by sib-pair stratification.

1.6 Development of association methods

The simplest way to detect allelic association is to perform a case-control study. However, case-control design is difficult to apply because of the effect of possible population stratification. The family-based association tests, such as HRR test and TDT test, are not affected by population stratification, because in these methods parental alleles not transmitted to the affected child are used as controls. The original TDT test aims to deal with cases where the marker locus has only two alleles. Many modifications have been made to handle multiallelic marker cases. Bickeböller and Clerget-Darpoux (1995) extended the biallelic TDT to compare transmitted and nontransmitted alleles of one parent for a multiallelic marker (TDT_a). They also developed a test statistics, TDT_g in which they studied both parents' allele transmission and compared genotypes formed by the two transmitted alleles and genotypes formed by the two nontransmitted alleles. Spielman and Ewens (1996) also proposed a test statistic (T_{mhet}) that generalized the biallelic TDT test to multiallelic markers. They compared the number of times each marker allele is transmitted to affected offspring to the number of times it is not and summed over all alleles at the marker loci. Their test statistic is relatively easy to use. TDT test was initially constructed to test linkage and association in families with known parental genotypes. When the disease has a late age of onset, it is very difficult or impossible to obtain parental

genotype information. The test statistics proposed by Curtis (1997) circumvent the problem of genotyping parents by using unaffected siblings as controls so that the test is still robust against bias due to population stratification. The S-TDT test proposed by Spielman and Ewens (1998) deals with missing parental data in a similar way. S-TDT compares the observed number of certain alleles in affected children with the number expected with no linkage, conditioned on the observed distribution of marker genotypes in the whole sibship. They also combined TDT and S-TDT into an overall test for data consisting of families with known parental genotypes and families with missing parental data. However, the TDT test has more power than S-TDT. They suggested the use of the TDT if the data can be analyzed by either method. The TDT test has also been extended to more than one marker locus. Wilson (1997) studied the marker transmission of two multiallelic marker loci from parents to affected offspring. The author assumed that the disease locus is located between two marker loci and the parental haplotypes are known. Zhao et al. (2000) proposed a statistical method to analyze multiple tightly linked markers simultaneously. Their simulation results showed that their tests have more power than other existing methods. The TDT test can also be generalized to extended pedigree. Martin et al. (1997) proposed two test statistics that use data from all of the affected children and are more powerful than the normal TDT using a single affected child. Later, Martin (2000) developed a pedigree disequilibrium test (PDT) that can analysis linkage disequilibrium in general pedigrees.

Recently, more and more attention has been attracted to test association based on

haplotype sharing rather than single marker sharing. Conserved ancestral chromosome segments in unrelated individuals usually extend 1-2cM (March, 1999). Haplotype sharing analysis is inherently more powerful than association analysis, because haplotype sharing analysis uses more information than association analysis. Some haplotype sharing test statistics have been developed based on sharing of haplotype identical by descent in affected individuals (Te Meerman et al., 1995; van der Meulen and Te Meerman, 1997). They showed that the affected individuals are not only likely to share alleles at a single locus, but also at the surrounding haplotype. Clayton and Jones (1999) generalized the TDT test to detect association between several adjacent loci. Similar to the TDT test, their approach compared the transmitted haplotypes with untransmitted haplotypes and aimed to detect regions of linkage disequilibrium where the susceptibility gene is located. Recently, Daly et al. (2001) developed a haplotype block approach, in which they grouped single-nucleotide polymorphisms (SNPs) into discrete haplotype groups and performed LD mapping based on haplotype blocks. They found that their approach can clarify LD analysis and reduce false positive results.

1.7 Summary and future direction

Affected sib pair methods and TDT tests have been proved to be useful in identifying disease susceptibility genes (eg., cystic fibrosis, Huntington's Disease). They are easy to use and have many advantages over other linkage analyses: collecting data from sib pairs may be easier than collecting extended pedigree data; the underlying genetic

mechanism of the disease doesn't need to be specified. However, theses methods also have some drawbacks. First, they can't estimate the recombination fraction and disequilibrium coefficients between the marker locus and disease locus. They can only identity a region where disease locus may be located. Molecular biology methods are necessary to further refine the location of disease locus. Second, almost all the linkage and association methods assume that there is no inbreeding and relatedness in the populations. This might not be true all the time. There is slight inbreeding in human populations and inbreeding is significant in some animal and plant populations. Third, linkage analysis alone has limited resolution. Most linkage maps are constructed with a limited number of individuals and generations and thus have limited resolution. Association analysis alone may result in high false positive rates, because many factors other than linkage could lead to association.

Génin and Clerget-Darpoux (1996, 1998) proposed several IBD tests which can be applied to inbreeding populations. However, their IBD tests can not be applied practically because they assumed that the allele IBD state can be determined. In most cases, if the IBD state of the sib pairs can be identified unequivocally, their parents do not share common alleles and can not be related. The whole idea of population inbreeding in ASP tests will be valid only when the markers are not highly polymorphic or the parental information is missing. Allowance for population inbreeding and relatedness will make more sense when identity by state tests are applied since IBS tests relax the IBD relations by substituting sib pair IBS relations. Parental information is no longer needed and the population could be inbred to any degree. In chapter

one, we modified two classical IBS tests proposed by Lange (1986a, b) to allow for population inbreeding. We evaluate the power and false positive rates of the modified tests under three disease models using simulated data. When the population inbreeding coefficient is large, both the false positive rates and power are reduced when the modified test statistics are applied, although power remained high under a recessive disease model. Allowing for inbreeding is therefore appropriate at least for diseases known to be recessive. In the second chapter, we compute the expected proportions of ASP sharing 0, 1 and 2 IBD marker alleles in inbred population and verify our calculations by two consistency checks. In the third chapter, we estimate linkage and linkage disequilibrium coefficients simultaneously in an attempt to increase mapping resolution and decrease false positive rates. We extend our single locus model to deal with two marker loci simultaneously to increase accuracy of estimation. We perform simulation studies to verify our EM algorithm. Our simulation results suggest that our EM algorithm can give consistent estimates of all the parameters considered.

Chapter 2

Affected Sib Pair Tests in Inbred Populations

2.1 Introduction

The affected-sib-pair (ASP) method was first suggested by Penrose (1953) and has become a widely used method in studying the inheritance of human diseases. ASP methods study the possible genetic linkage between a putative disease susceptibility locus and one or more marker loci by demonstration within families of nonrandom segregation of parental alleles in affected children (Penrose, 1953; Day and Simons, 1976; Suarez, Rice and Reich, 1978; deVries et al., 1976; Green and Woodrow, 1977; Fishman et al., 1978). The advantages of sib-pair methods are that they do not need to assume a specific mode of inheritance for the disease being studied and, further, linkage disequilibrium between the marker and disease genes is not necessary for detecting linkage. Many modifications have been made to this method for dealing with different cases, such as when the marker is not highly polymorphic (Risch, 1990; Holmans, 1993) and when there are more than two affected sibs in the family (Hodge, 1984; Green, Low and Woodrow, 1983; Payammi et al., 1985). Most ASP methods

study sharing of alleles identical by descent (IBD), but others study sharing of alleles identical by state (IBS) (Lange, 1986a,b; Weeks and Lange, 1988). There has been little attention paid to the effects of low-level inbreeding and relatedness, and consequent population structure, caused by evolutionary history.

We know that there is slight inbreeding in human population because of limited population sizes during evolutionary history. Génin and Clerget-Darpoux (1996, 1998) first studied the effect of inbreeding on affected sib pair analysis. They modified three ASP tests (Blackwelder and Elston, 1985) which test sharing of IBD alleles. In a non-inbreeding population, and in the absence of linkage, the expected proportions of sib pairs with 0, 1 and 2 marker alleles IBD are 0.25, 0.5, and 0.25, respectively. However, with inbreeding, these IBD proportions are no longer true because all the individuals are related and the two parents could have alleles at a locus that are identical copies of a single ancestral allele. So there must be on average more than a quarter of the sib pairs who share two marker alleles IBD and less than a quarter of the sib pairs who do not share any alleles identical by descent. The null hypothesis proportions of the ASP test need to be modified for inbreeding populations and a high false positive rate may be obtained if traditional ASP tests are used. Génin and Clerget-Darpoux used the notion of IBW state (IBD alleles between and within individuals) and recalculated the expected proportions of sib pairs with 0, 1 and 2 marker alleles IBD when sibs were sampled in a consanguineous population. (There are some errors in their calculation.) In their modified tests, they assumed the marker IBD state of the sib pairs can be identified unambiguously, which made their assumption

of population inbreeding not appropriate. Because in most cases, if the IBD states of the sib pairs can be identified unambiguously, their parents do not share common alleles and thus could not be related; in that case the null values for 0, 1, or 2 marker alleles will still be 0.25, 0.5, and 0.25. The whole idea of population inbreeding in ASP tests will be valid only when the markers are not highly polymorphic or the parental information is missing. Allowance for population inbreeding and relatedness will be more appropriate when identity by state tests are applied since IBS tests relax the IBD relations by substituting sib pair IBS relations. Parental information is no longer needed and the population could be inbred to any degree.

IBS tests are especially useful when parents are not available for typing. In many cases, parental information is very difficult to get, especially for diseases with late onset. Even when parents are available for typing, their mating type may not allow unambiguous inference of the IBD relationships at the marker locus in the sib pairs. IBS tests suggested by Lange (1986a,b) cope with this very well. Lange classified IBS into three categories based on the sib pair's genotypes and calculated the expected number of pairs falling into each of the three categories. For extremely polymorphic loci the IBS relations correspond fairly closely with the usual IBD relations. Deviations from the expected proportions within families imply linkage between the marker locus and disease locus. However, Lange's IBS tests deal only with cases where there is no inbreeding or relatedness in the population. These tests will lead to high false positive rates when performed in inbred populations.

The goals of the present study are:

1. to modify Lange's tests to allow for inbreeding and relatedness, in an attempt to decrease the false positive rate.

2. to use simulation to investigate the power and false positive rates for Lange's tests and our modifications.

2.2 Method

2.2.1 z-test

The z-test proposed by Lange (1986a) is based on the Central Limit Theorem. The IBS relations of the sib pairs can be classified into three categories: concordant if they have the same marker genotypes, discordant if they do not share any marker alleles, and half concordant otherwise. Define Z to be the sum of marker concordances of the affected sibs in each family and $Z = \sum_{i<j} X_{ij}$, where i and j represent different sibs in the family where

$$X_{ij} = \begin{cases} 1 & i \text{ and } j \text{ are concordant} \\ 0.5 & i \text{ and } j \text{ are half concordant} \\ 0 & i \text{ and } j \text{ are discordant} \end{cases}$$

The expected value of Z can be calculated based on the conditional expectations of the marker IBS relations of sib pairs within each mating type and the frequencies of the seven parental mating types.

$$E(Z) = \sum_{i<j} E(X_{ij})$$

$$E(X_{ij}) = \sum_{t} E(X_{ij}|M = t)Pr(M = t)$$

where t represents the mating type. The expected value of Z^2 can be computed from the marker concordances of pairs of sib pairs in each family.

$$\begin{aligned} E(Z^2) &= \sum_{i<j}\sum_{k<l} E(X_{ij}X_{kl}) \\ &= \frac{s(s-1)}{2}E(X_{ij}X_{ij}) + s(s-1)(s-2)E(X_{ij}X_{il}) \\ &\quad + \frac{s(s-1)(s-2)(s-3)}{4}E(X_{ij}X_{kl}) \end{aligned}$$

where i, j, k and l denote different siblings in the family and s is the number of affected sibs in the family. Then, the variance of Z is

$$Var(Z) = E(Z^2) - \{E(Z)\}^2$$

The test statistic for the entire sample is then

$$T = \frac{\Sigma_s \Sigma_r W_{rs}\{Z_{rs} - E(Z_{rs})\}}{\{\Sigma_s \Sigma_r W_{rs}^2 Var(Z_{rs})\}^{1/2}} \tag{2.1}$$

where s still indexes the number of affected sibs in the family, r indexes the sib sets with size s and W_{rs} are the weights. One choice for W_{rs} is

$$W_{rs} = 1/Var(Z_{rs})^{\frac{1}{2}}$$

The probabilities of each of the seven mating types given by Lange are for an infinite population under random mating, i.e. no inbreeding or relatedness. However, when there is inbreeding the genotypic proportions will deviate from the Hardy-Weinberg values. We retained the random mating assumption but allowed for inbreeding and relatedness due to evolutionary forces. Weir (1984) expressed joint

genotypic frequencies in terms of higher-order descent measures (Cockerham, 1971), and then simplified these in Evett and Weir (1998) for the random-mating case. It is well known that there are 15 possible IBD relationships for four alleles of two individuals (Harris, 1964; Gillois, 1965; Cockerham, 1971). Define δ_0 to be the probability that none of the four alleles are IBD; δ_{ij} to be the probability that only two (i, j) of the four alleles are IBD, δ_{ijk} to be the probability that only three (i, j, k) of the four alleles are IBD and δ_{ijkl} to be the probability that all four alleles are IBD. Further, $\delta_{ij.kl}$ is the probability that two pairs of the four alleles (i, j and k, l) are IBD. For a large population with random mating, the IBD status of two alleles is the same whether they are in the same or different individuals and the inbreeding coefficient equals the coancestry coefficient. For a population in evolutionary equilibrium (Evett and Weir, 1998), all these 15 measures can be expressed in terms of the inbreeding coefficient θ (the Dirichlet distribution case). The 15 possible IBD relationships are listed in Table 2.1, together with their probabilities in general, in the general random-mating case when the relationships are symmetric among alleles, in the case when the Dirichlet distribution can be invoked, and in the completely unrelated case.

With these allele IBD relationships, the probabilities of the seven possible parental mating types can be derived (Weir and Cockerham 1984; Evett and Weir, 1998). These mating type frequencies in inbred populations can be represented as functions of θ and allele frequencies p_i for allele A_i. The seven mating type probabilities are shown in Table 2.2, 2.3 and 2.4. The left columns show a typical member of each class and different letters indicate different alleles. The right columns show the probability

Table 2.1: Descent relations among alleles for two individuals: X with alleles a and b and Y with alleles c and d

Alleles IBD	Pr(IBD)	Random	Dirichlet*	Unrelated
none	δ_0	$1-6\theta+8\gamma+3\Delta-6\delta$	$(1-\theta)^3$	1
$a\equiv b$	δ_{ab}	$\theta-2\gamma-\Delta+2\delta$	$\theta(1-\theta)^2$	0
$a\equiv c$	δ_{ac}	$\theta-2\gamma-\Delta+2\delta$	$\theta(1-\theta)^2$	0
$a\equiv d$	δ_{ad}	$\theta-2\gamma-\Delta+2\delta$	$\theta(1-\theta)^2$	0
$b\equiv c$	δ_{bc}	$\theta-2\gamma-\Delta+2\delta$	$\theta(1-\theta)^2$	0
$b\equiv d$	δ_{bd}	$\theta-2\gamma-\Delta+2\delta$	$\theta(1-\theta)^2$	0
$c\equiv d$	δ_{cd}	$\theta-2\gamma-\Delta+2\delta$	$\theta(1-\theta)^2$	0
$a\equiv b\equiv c$	δ_{abc}	$\gamma-\delta$	$2\theta^2(1-\theta)$	0
$a\equiv b\equiv d$	δ_{abd}	$\gamma-\delta$	$2\theta^2(1-\theta)$	0
$a\equiv c\equiv d$	δ_{acd}	$\gamma-\delta$	$2\theta^2(1-\theta)$	0
$b\equiv c\equiv d$	δ_{bcd}	$\gamma-\delta$	$2\theta^2(1-\theta)$	0
$a\equiv b, c\equiv d$	$\delta_{ab.cd}$	$\Delta-\delta$	$\theta^2(1-\theta)$	0
$a\equiv c, b\equiv d$	$\delta_{ac.bd}$	$\Delta-\delta$	$\theta^2(1-\theta)$	0
$a\equiv d, b\equiv c$	$\delta_{ad.bc}$	$\Delta-\delta$	$\theta^2(1-\theta)$	0
$a\equiv b\equiv c\equiv d$	δ_{abcd}	δ	$6\theta^3$	0

* Each term has been multiplied by $(1+\theta)(1+2\theta)$.

of the whole class, and so involves sums over all alleles.

Table 2.2: Mating type probabilities expressed in terms of δ

Genotypes	Probability over all alleles
ii, ii	$\delta_{abcd}\sum_i p_i + (\delta_{abc}+\delta_{abd}+\delta_{acd}+\delta_{bcd})\sum_i p_i^2$ $+(\delta_{ab.cd}+\delta_{ac.bd}+\delta_{ad.bc})\sum_i p_i^2 + (\delta_{ab}+\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd}+\delta_{cd})\sum_i p_i^3$ $+\delta_0\sum_i p_i^4$
ii, jj	$\delta_{ab.cd}\sum_{i\neq j} p_ip_j + \delta_{ab}\sum_{i\neq j} p_ip_j^2 + \delta_{cd}\sum_{i\neq j} p_i^2p_j + \delta_0\sum_{i\neq j} p_i^2p_j^2$
ii, ij	$2(\delta_{abc}+\delta_{abd})\sum_{i\neq j} p_ip_j + 2(2\delta_{ab}+\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd})\sum_{i\neq j} p_i^2p_j$ $+ 4\delta_0\sum_{i\neq j} p_i^3p_j$
ii, jk	$2\delta_{ab}\sum_{i\neq j\neq k} p_ip_jp_k + 2\delta_0\sum_{i\neq j\neq k} p_i^2p_jp_k$
ij, ij	$(\delta_{ac.bd}+\delta_{ad.bc})\sum_{i\neq j} p_ip_j + (\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd})\sum_{i\neq j} p_ip_j(p_i+p_j)/2$ $+2\delta_0\sum_{i\neq j} p_i^2p_j^2$
ij, ik	$(\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd})\sum_{i\neq j\neq k} p_ip_jp_k + 4\delta_0\sum_{i\neq j\neq k} p_i^2p_jp_k$
ij, kl	$\delta_0\sum_{i\neq j\neq k\neq l} p_ip_jp_kp_l$

In evaluating these expressions, use may be made of

$$\begin{aligned}
\sum_{i\neq j} p_ip_j &= 1-\sum_i p_i^2 \\
\sum_{i\neq j} p_i^2p_j &= \sum_i p_i^2 - \sum_i p_i^3 \\
\sum_{i\neq j} p_i^2p_j^2 &= (\sum_i p_i^2)^2 - \sum_i p_i^4 \\
\sum_{i\neq j} p_i^3p_j &= \sum_i p_i^3 - \sum_i p_i^4 \\
\sum_{i\neq j\neq k} p_ip_jp_k &= 1-3\sum_i p_i^2 + 2\sum_i p_i^3 \\
\sum_{i\neq j\neq k} p_i^2p_jp_k &= \sum_i p_i^2 - 2\sum_i p_i^3 - (\sum_i p_i^2)^2 + 2\sum_i p_i^4
\end{aligned}$$

Table 2.3: Mating type probabilities expressed in terms of higher order descent measures

Genotypes	Probability over all alleles
ii, ii	$= \delta \sum_i p_i + (4\gamma + 3\Delta - 7\delta) \sum_i p_i^2 + 6(\theta - 2\gamma - \Delta + 2\delta) \sum_i p_i^3$ $+(1 - 6\theta + 8\gamma + 3\Delta - 6\delta) \sum_i p_i^4$
ii, jj	$= (\Delta - \delta) \sum_{i \neq j} p_i p_j + (\theta - 2\gamma - \Delta + 2\delta) \sum_{i \neq j} p_i p_j (p_i + p_j)$ $+(1 - 6\theta + 8\gamma + 3\Delta - 6\delta) \sum_{i \neq j} p_i^2 p_j^2$
ii, ij	$= 4(\gamma - \delta) \sum_{i \neq j} p_i p_j + 12(\theta - 2\gamma - \Delta + 2\delta) \sum_{i \neq j} p_i^2 p_j$ $+ 4(1 - 6\theta + 8\gamma + 3\Delta - 6\delta) \sum_{i \neq j} p_i^3 p_j$
ii, jk	$= 2(\theta - 2\gamma - \Delta + 2\delta) \sum_{i \neq j \neq k} p_i p_j p_k$ $+2(1 - 6\theta + 8\gamma + 3\Delta - 6\delta) \sum_{i \neq j \neq k} p_i^2 p_j p_k$
ij, ij	$= 2(\Delta - \delta) \sum_{i \neq j} p_i p_j + 2(\theta - 2\gamma - \Delta + 2\delta) \sum_{i \neq j} p_i p_j (p_i + p_j)$ $+ 2(1 - 6\theta + 8\gamma + 3\Delta - 6\delta) \sum_{i \neq j} p_i^2 p_j^2$
ij, ik	$= 4(\theta - 2\gamma - \Delta + 2\delta) \sum_{i \neq j \neq j} p_i p_j p_k$ $+4(1 - 6\theta + 8\gamma + 3\Delta - 6\delta) \sum_{i \neq j \neq k} p_i^2 p_j p_k$
ij, kl	$= (1 - 6\theta + 8\gamma + 3\Delta - 6\delta) \sum_{i \neq j \neq j \neq l} p_i p_j p_k p_l$

Table 2.4: Mating type probabilities expressed in terms of population inbreeding coefficient θ

Genotypes	Probability over all alleles
ii, ii	$= [6\theta^3 + 11\theta^2(1-\theta)\sum_i p_i^2 + 6\theta(1-\theta)^2\sum_i p_i^3$ $+(1-\theta)^3\sum_i p_i^4]/(1+\theta)(1+2\theta)$
ii, jj	$= [\theta^2(1-\theta)\sum_{i\neq j} p_i p_j + \theta(1-\theta)^2\sum_{i\neq j} p_i p_j(p_i+p_j)$ $+(1-\theta)^3\sum_{i\neq j} p_i^2 p_j^2]/(1+\theta)(1+2\theta)$
ii, ij	$= [8\theta^2(1-\theta)\sum_{i\neq j} p_i p_j + 12\theta(1-\theta)^2\sum_{i\neq j} p_i^2 p_j$ $+4(1-\theta)^3\sum_{i\neq j} p_i^3 p_j]/(1+\theta)(1+2\theta)$
ii, jk	$= [2\theta(1-\theta)^2\sum_{i\neq j\neq k} p_i p_j p_k + 2(1-\theta)^3\sum_{i\neq j\neq k} p_i^2 p_j p_k]/(1+\theta)(1+2\theta)$
ij, ij	$= [2\theta^2(1-\theta)\sum_{i\neq j} p_i p_j + 2\theta(1-\theta)^2\sum_{i\neq j} p_i p_j(p_i+p_j)$ $+2(1-\theta)^3\sum_{i\neq j} p_i^2 p_j^2]/(1+\theta)(1+2\theta)$
ij, ik	$= [4\theta(1-\theta)^2\sum_{i\neq j\neq k} p_i p_j p_k + 4(1-\theta)^3\sum_{i\neq j\neq k} p_i^2 p_j p_k]/(1+\theta)(1+2\theta)$
ij, kl	$= [(1-\theta)^3\sum_{i\neq j\neq k\neq l} p_i p_j p_k p_l]/(1+\theta)(1+2\theta)$

$$\sum_{i\neq j\neq k\neq l} p_i p_j p_k p_l \;=\; 1 - 6\sum_i p_i^2 + 8\sum_i p_i^3 + 3(\sum_i p_i^2)^2 - 6\sum_i p_i^4$$

These mating type probabilities are exactly the same as those of Lange when the inbreeding coefficient is zero. The conditional expectations remain the same as given by Lange because the IBS relations are irrelevant to the parental allele IBD relations. The mean and variance of Z change because of changing the mating type frequencies. The expected values for the Xs in inbreeding population can be expressed as follows:

$$\begin{aligned}
E(X_{ij}) &= [24\theta^3 + 11\theta^2(1-\theta)(3+\sum_i p_i^2) + 3\theta(1-\theta)^2(5+3\sum_i p_i^2) \\
&\quad +(1-\theta)^3(\sum_i p_i^4 - 2\sum_i p_i^3 + 3\sum_i p_i^2 + 2)]/4(1+\theta)(1+2\theta) \\
E(X_{ij}, X_{ij}) &= [48\theta^3 + 8\theta^2(1-\theta)(7+4\sum_i p_i^2) + \theta(1-\theta)^2(2\sum_i p_i^3 + 23\sum_i p_i^2 + 23) \\
&\quad +(1-\theta)^3(2(\sum_i p_i^2)^2 - 2\sum_i p_i^3 + 5\sum_i p_i^2 + 3)]/8(1+\theta)(1+2\theta) \\
E(X_{ij}, X_{il}) &= \{192\theta^3 + 2\theta^2(1-\theta)(101+75\sum_i p_i^2) + \theta(1-\theta)^2(14\sum_i p_i^3 \\
&\quad +101\sum_i p_i^2 + 77) + (1-\theta)^3[8\sum_i p_i^4 + 5(\sum_i p_i^2)^2 - 18\sum_i p_i^3 \\
&\quad +29\sum_i p_i^2 + 8]\}/32(1+\theta)(1+2\theta) \\
E(X_{ij}, X_{kl}) &= [384\theta^3 + 2\theta^2(1-\theta)(201+151\sum_i p_i^2) + 3\theta(1-\theta)^2(10\sum_i p_i^3 \\
&\quad +67\sum_i p_i^2 + 51) + (1-\theta)^3(16\sum_i p_i^4 + 9(\sum_i p_i^2)^2 - 34\sum_i p_i^3 \\
&\quad +57\sum_i p_i^2 + 16)]/64(1+\theta)(1+2\theta) \\
E(Z) &= \sum_{i<j} E(X_{ij}) = \frac{s(s-1)}{2}E(X_{ij}) \\
E(Z^2) &= \sum_{i<j}\sum_{k<l} E(X_{ij}X_{kl}) \\
&= \frac{s(s-1)}{2}E(X_{ij}X_{ij}) + s(s-1)(s-2)E(X_{ij}X_{il})
\end{aligned}$$

$$+\frac{s(s-1)(s-2)(s-3)}{4}E(X_{ij}X_{kl})$$

$$Var(Z) = E(Z^2) - \{E(Z)\}^2$$

We can substitute the above expected value of Z and variance of Z into the original test statistics and construct new test statistics for inbred population.

2.2.2 Chi-square test

The chi-square test proposed by Lange (1986b) compared the observed number of sib pairs in each of the three IBS categories in the sample with the expected numbers. Lange calculated the three expected IBS proportions using IBD proportions and mating type frequencies. His chi-square test deals with cases where the population is infinite and the mating is random. In an inbred population, the expected proportions of sib pairs with 0, 1 and 2 marker alleles IBD are no longer 0.25, 0.5, and 0.25 and the parental mating type frequencies also changed. Here, we present another way to calculate the three expected IBS proportions based solely on mating types.

For any pair of individuals chosen to be parents, the probabilities of each type of offspring, and of each class of sib pair is shown in Table 2.5.

As in the previous section, X takes the values 1, 0.5, 0 according to the concordance state of the sib pair. The probabilities of each of the three IBS states can be obtained from the conditional probabilities,

$Pr(X = n) = \sum_t Pr(X = n|t)Pr(t)$

where n=1, 0.5, 0 and t runs over all possible mating types. For unrelated parents, i.e. $\theta = 0$,

Table 2.5: IBS probabilities conditional on mating types

			Probability that		
Parents	Offspring	Sib Pairs	Pr(X=1)	Pr(X=0.5)	Pr(X=0)
ii, ii	ii	ii, ii	1	0	0
ii, jj	ij	ij, ij	1	0	0
ii, ij	$\frac{1}{2}ii + \frac{1}{2}ij$	$\frac{1}{4}ii, ii + \frac{1}{2}ii, ij + \frac{1}{4}ij, ij$	$\frac{1}{2}$	$\frac{1}{2}$	0
ii, jk	$\frac{1}{2}ij + \frac{1}{2}ik$	$\frac{1}{2}ij, ij + \frac{1}{2}ij, ik$	$\frac{1}{2}$	$\frac{1}{2}$	0
ij, ij	$\frac{1}{4}ii + \frac{1}{2}ij$	$\frac{1}{8}ii, ii + \frac{1}{2}ii, ij$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$
	$+\frac{1}{4}jj$	$+\frac{1}{8}ii, jj + \frac{1}{4}ij, ij$			
ij, ik	$\frac{1}{4}ii + \frac{1}{4}ij$	$\frac{1}{16}ii, ii + \frac{1}{4}ii, ij + \frac{1}{8}ii, jk$	$\frac{1}{4}$	$\frac{5}{8}$	$\frac{1}{8}$
	$+\frac{1}{4}ik + \frac{1}{4}jk$	$+\frac{3}{16}ij, ij + \frac{3}{8}ij, ik$			
ij, kl	$\frac{1}{4}ik + \frac{1}{4}il$	$\frac{1}{4}ij, ij + \frac{1}{2}ij, ik + \frac{1}{4}ij, kl$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	$+\frac{1}{4}jk + \frac{1}{4}jl$				

$$
\begin{aligned}
\Pr(X=1) &= \frac{1}{4}+\frac{1}{2}\sum_i p_i^2+\frac{1}{2}(\sum_i p_i^2)^2-\frac{1}{4}\sum_i p_i^4 \\
\Pr(X=0.5) &= \frac{1}{2}+\frac{1}{2}\sum_i p_i^2-\sum_i p_i^3-(\sum_i p_i^2)^2+\sum_i p_i^4 \\
\Pr(X=0) &= \frac{1}{4}-\sum_i p_i^2+\sum_i p_i^3+\frac{1}{2}(\sum_i p_i^2)^2-\frac{3}{4}\sum_i p_i^4
\end{aligned}
$$

as given in Lange's paper after some rearrangement.

When the parents are drawn randomly from an equilibrium population characterized by inbreeding coefficient θ, the three probabilities become:

$$
\begin{aligned}
\Pr(X=1) &= \Big[24\theta^3+\theta^2(1-\theta)(23+21\sum_i p_i^2)+2\theta(1-\theta)^2(4+7\sum_i p_i^2+\sum_i p_i^3) \\
&\quad +(1-\theta)^3\{1+2\sum_i p_i^2+2(\sum_i p_i^2)^2-\sum_i p_i^4\}\Big]/4(1+\theta)(1+2\theta)
\end{aligned}
$$

$$
\begin{aligned}
\Pr(X=0.5) &= (1-\theta)\Big[10\theta^2(1-\sum_i p_i^2)+\theta(1-\theta)(7-5\sum_i p_i^2-2\sum_i p_i^3) \\
&\quad +(1-\theta)^2\{1+\sum_i p_i^2-2\sum_i p_i^3-2(\sum_i p_i^2)^2 \\
&\quad +2\sum_i p_i^4\}\Big]/2(1+\theta)(1+2\theta)
\end{aligned}
$$

$$
\begin{aligned}
\Pr(X=0) &= (1-\theta)\Big[\theta^2(1-\sum_i p_i^2)+2\theta(1-\theta)(1-2\sum_i p_i^2+\sum_i p_i^3) \\
&\quad +(1-\theta)^2\{1-4\sum_i p_i^2+4\sum_i p_i^3+2(\sum_i p_i^2)^2 \\
&\quad -3\sum_i p_i^4\}\Big]/4(1+\theta)(1+2\theta)
\end{aligned}
$$

Then, the usual chi-square goodness of fit test can be applied to test whether or

not there are deviations from the expected numbers of sib pairs in each IBS category. Deviations from expected numbers imply linkage between the marker locus and disease locus.

2.2.3 Simulations

To verify that our new test statistics give lower false positive rates than the original Lange tests and to estimate the powers, we carried out a set of simulations. For both tests, we considered a single locus, two allele disease model. We simulated data under an additive model, a recessive model and a dominant model. We set the two alleles at the disease locus to have equal frequencies and chose the population prevalence of 0.075 for each disease model. The population was allowed to undergo zero to one hundred generations of random mating to get different degrees of inbreeding and relatedness. The powers and false positive rates were evaluated for a two-marker allele model as well as for a four-marker allele model, where the marker alleles were codominant and had equal frequencies. For each sample, the power and false positive rates of our new tests were compared with those of the original Lange tests. Families with two affected siblings were selected. Each sample consisted of 100 families and 1,000 replicated samples were simulated, to give stable estimates of power and false positive rates for these tests. To estimate the powers, we simulated data in which the marker locus was tightly linked to the disease locus (recombination fraction between the marker locus and disease locus is zero) and the initial linkage disequilibrium between the marker alleles and the disease alleles was set to be maximum. To evaluate

the false positive rates, we simulated data under the null hypothesis of no linkage, where the recombination fraction between the marker locus and disease locus was 0.5. Data sets with maximum initial linkage disequilibrium and minimum initial linkage disequilibrium were used when estimating false positive rate.

2.3 Results

2.3.1 False-Positive Rates

The data were simulated for each of three disease models, dominant, additive, and recessive, with prevalence = 0.075, two genetic marker types (the two-allele marker and four-allele marker), two different initial linkage disequilibrium values (minimum and maximum initial linkage disequilibrium) and six different inbreeding coefficient values (θ ranging from 0 to 0.09521) under the null hypothesis of no disease-marker linkage for both the z-test and the chi-square test. The false positive rates of the four methods, obtained by Lange's z-test, our inbred z-test, Lange's chi-square test, our inbred chi-square test were compared. For each statistic, the false positive rates were calculated as the proportion of replicates, out of a total of 1,000 replicates, which showed linkage between the disease locus and marker locus (at the 0.05 significance level). The results are described in Table 2.6 and Table 2.7.

To our surprise, we found that in the initial populations (inbreeding coefficient of the population, θ, equals zero), the empirical false positive rates were very large when the linkage disequilibrium between the marker alleles and the disease alleles was set

Table 2.6: Comparison of the false positive rates between Lange tests and the inbreeding tests for data sets with no initial LD

		Two marker alleles			Four marker alleles		
θ^a	Statistic	Rec[b]	Add	Dom	Rec	Add	Dom
0	Lange z-test	.054	.049	.052	.052	.045	.051
	Inbred z-test	.054	.049	.052	.052	.045	.051
	Lange chi-square	.058	.060	.060	.061	.044	.053
	Inbred chi-square	.058	.060	.060	.061	.044	.053
0.001	Lange z-test	.060	.056	.055	.055	.054	.059
	Inbred z-test	.060	.056	.055	.055	.054	.059
	Lange chi-square	.062	.053	.053	.057	.060	.049
	Inbred chi-square	.062	.053	.054	.057	.061	.049
0.00996	Lange z-test	.070	.065	.054	.062	.059	.064
	Inbred z-test	.070	.065	.054	.062	.059	.064
	Lange chi-square	.064	.038	.038	.062	.060	.048
	Inbred chi-square	.072	.041	.046	.063	.060	.044
0.02957	Lange z-test	.100	.079	.077	.103	.088	.103
	Inbred z-test	.073	.061	.055	.053	.049	.050
	Lange chi-square	.058	.059	.062	.072	.065	.062
	Inbred chi-square	.060	.061	.073	.063	.062	.055
0.04879	Lange z-test	.134	.116	.112	.175	.166	.148
	Inbred z-test	.074	.069	.066	.093	.083	.067
	Lange chi-square	.075	.070	.085	.109	.098	.090
	Inbred chi-square	.093	.074	.089	.079	.071	.069
0.09521	Lange z-test	.204	.233	.212	.309	.288	.314
	Inbred z-test	.121	.131	.117	.131	.093	.108
	Lange chi-square	.123	.130	.133	.179	.166	.184
	Inbred chi-square	.110	.113	.123	.112	.112	.105

[a] θ is inbreeding coefficient of the population

[b] Rec represents recessive disease model; Add represents additive disease model; Dom represents dominant disease model.

Table 2.7: Comparison of the false positive rates between Lange tests and the inbreeding tests for data sets with no initial LD and maximum initial LD

θ^a	Statistic	Four marker alleles					
		Rec[b]		Add		Dom	
		D^c_{min}	D^d_{max}	D_{min}	D_{max}	D_{min}	D_{max}
0	Lange z-test	.052	.381	.045	.094	.051	.060
	Inbred z-test	.052	.381	.045	.094	.051	.060
	Lange chi-square	.061	.209	.044	.075	.053	.060
	Inbred chi-square	.061	.209	.044	.075	.053	.060
0.001	Lange z-test	.055	.110	.054	.059	.059	.052
	Inbred z-test	.055	.110	.054	.059	.059	.052
	Lange chi-square	.057	.060	.060	.054	.049	.051
	Inbred chi-square	.057	.061	.061	.054	.049	.053
0.00996	Lange z-test	.062	.067	.059	.060	.064	.064
	Inbred z-test	.062	.067	.059	.060	.064	.064
	Lange chi-square	.062	.051	.060	.049	.048	.045
	Inbred chi-square	.063	.054	.060	.050	.044	.047
0.02957	Lange z-test	.103	.100	.088	.101	.103	.109
	Inbred z-test	.053	.059	.049	.058	.050	.058
	Lange chi-square	.072	.068	.065	.059	.062	.072
	Inbred chi-square	.063	.063	.062	.057	.055	.059
0.04879	Lange z-test	.175	.156	.166	.150	.148	.150
	Inbred z-test	.093	.073	.083	.070	.067	.074
	Lange chi-square	.109	.104	.098	.089	.090	.085
	Inbred chi-square	.079	.082	.071	.076	.069	.063
0.09521	Lange z-test	.309	.310	.288	.325	.314	.319
	Inbred z-test	.131	.110	.093	.111	.108	.116
	Lange chi-square	.179	.203	.166	.193	.184	.188
	Inbred chi-square	.112	.141	.112	.109	.105	.106

[a] θ is inbreeding coefficient of the population

[b] Rec represents recessive disease model; Add represents additive disease model; Dom represents dominant disease model.

[c] D_{min} denotes no initial linkage disequilibrium.

[d] D_{max} denotes maximum initial linkage disequilibrium.

to its maximum in all four IBS tests, especially in recessive disease model. When there were no initial LD, the empirical false positive rates were close to the nominal value. These suggested that, contrary to our belief, association between marker and disease loci did affect the Lange's IBS tests. Therefore, Lange's IBS tests were tests of both linkage and association. A joint null hypothesis of no linkage disequilibrium and linkage between disease locus and marker locus must be considered. After one generation of random mating (θ equals 0.001), the false positive rates reduced greatly. This was due to the fact that the linkage disequilibrium (LD) in the second generation reduced to half of the initial value because of free recombination events. After more generations of random mating, the false positive rates in data sets with maximum initial LD became similar to those in data sets with minimum initial LD, because LD decayed very fast with free recombination and both data sets had minimum LD after a few generations of random mating.

Therefore, the estimated false positive rates should be the false positive rates for data sets with no LD and no linkage between disease and marker loci as listed in Table 2.6. When the individuals were unrelated (θ equals zero), our inbred z-test and inbred chi-square test performed exactly the same as those of Lange, as expected. This was because that the inbred IBS test statistics and Lange test statistics were exactly the same when θ equaled 0. When θ was small, the empirical false positive rates of the original Lange tests were also similar to those of the inbred tests and all of them were close to the significance level, which indicated that both Lange's z-test and chi square test were not affected by low level of population inbreeding. In general, when

the inbreeding coefficient was large, the false positive rates of Lange tests became very large and the false positive rates of our inbred tests became smaller than those of Lange tests. This is because our inbred tests take into account inbreeding and relatedness of the population. The larger the inbreeding coefficient, the smaller are our false positive rates compared to those of Lange tests. The false positive rates of our inbred z-test became lower than those of Lange's z-test when the inbreeding coefficient reached 0.03. Although the empirical false positive rates of our inbred IBS tests were higher than the nominal value, they reduced to almost a third to a half of those of Lange's z-test when θ is around 0.1. The false positive rates of chi-square test were more insensitive to inbreeding coefficient compared to those of the z test. The false positive rates of our inbred chi-square test also started to become smaller than the Lange test after the inbreeding coefficient became 0.03 when there are four marker alleles in the data set. When there were only two marker alleles, the empirical false positive rates were very interesting. Our chi-square test had a slightly larger false positive rate than those of Lange's chi-square test when θ was between 0.01 and 0.05 which might be caused by random drift of the sample. When θ reached 0.1, our inbred chi-square test had a smaller false positive rate than Lange test.

If we assume that there are many different alleles at the marker locus and each of the alleles has a small frequency, p_i^2, p_i^3, p_i^4 and $(p_i^2)^2$ all become very small. Then the expected value and variance of Z became

$$E(Z) = \frac{s(s-1)(4\theta^3+9\theta^2+9\theta+2)}{8(1+\theta)(1+2\theta)}$$

$$Var(Z) = \frac{s(s-1)(1-\theta)(-64s\theta^5+64s^2\theta^5-106s\theta^4+114s^2\theta^4+76\theta^4+81s^2\theta^3}{-61s\theta^3+158\theta^3-16s\theta^2+136\theta^2+32s^2\theta^2-5s\theta+70\theta+9s^2\theta+16)}{256(1+\theta)^2(1+2\theta)^2}$$

Figure 2.1 showed the theoretical z-statistic values ($E(Z)/\sqrt{Var(Z)}$) under different degrees of population inbreeding (θ ranged from 0 to 0.9). We found that the theoretical z-statistic was a monotone increasing function. It increased slowly when θ was small and increased rapidly when θ became large. This corresponded fairly well with our simulation results. When θ was small, the theoretical z-statistic values were similar to the noninbred theoretical z-statistic value. If we use the noninbred Z values in populations with a low level of inbreeding, the false positive rates would still be close to the nominal value. When θ was large, the theoretical z-statistic values were much bigger than the noninbred theoretical z-statistic value. If we use the noninbred Z values in populations with high level of inbreeding, the false positive rates would be much higher than the nominal value.

2.3.2 Power Estimates

In addition to determining the empirical false positive rates of the four tests, we also estimated the power of the tests to detect linkage. Since we now believe that Lange's IBS tests tested both linkage and association, we set the initial LD between marker and disease alleles to be maximum to estimate power. The same models and parameters were used as above except that the recombination fraction between the disease locus and marker locus was set to zero (complete linkage). Estimates of the power of each test were obtained by the proportion of samples that rejects the null

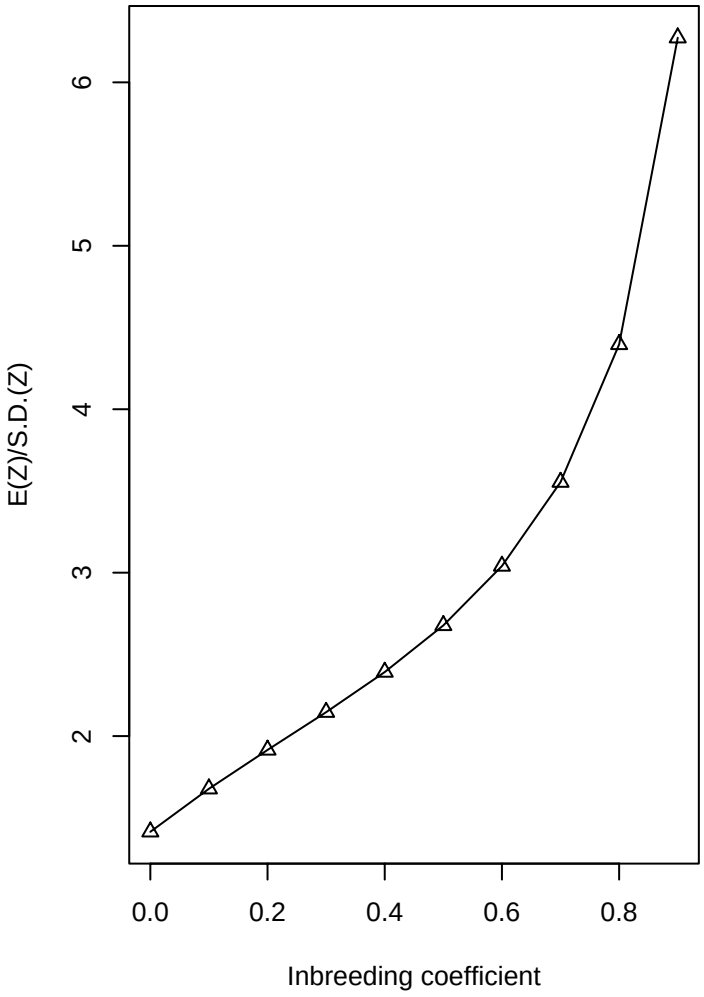

Figure 2.1: The theoretical z-statistic values under different population inbreeding coefficients

hypothesis of no linkage and association. The powers for the four tests are presented in Table 2.8.

Several general trends are obvious from the results. First, the power for the data sets with two marker alleles was greater than the power for the data sets with four marker alleles in additive and dominant disease model. This was interesting because usually for IBS tests, we expect to get higher power when there are more alleles at the marker loci (Bishop and Williamson, 1990). However, this can be explained by the fact that these IBS tests were tests of both linkage and association. In two marker allele data sets, the two marker alleles had equal allele frequencies of 0.5. The two alleles at the disease locus also had equal allele frequencies of 0.5. Thus the maximum LD between the marker alleles and the disease alleles was 0.25. In four marker allele data sets, all four marker alleles had equal allele frequencies of 0.25, which made the maximum possible LD between the marker alleles and the disease alleles equal to 0.125. The LD in four marker allele data sets was only half of the LD in two marker allele data sets. Thus lower powers were obtained in four marker allele data sets. Second, of the three disease models, the power of recessive model was the highest, which was also consistent with other IBS studies (Thomson and Motro, 1994). In the recessive disease model, the power for all the four tests was good; most of them even attained 100 percent power. The power was also high in additive and dominant disease models when two-marker allele data sets were used and z-tests were performed. However, the power was low in other cases. Third, like the IBD tests (Blackwelder and Elston, 1985), the z-tests were more powerful than the chi-square

Table 2.8: Comparison of power between Lange tests and the inbreeding tests

θ^a	Statistic	Two marker alleles			Four marker alleles		
		Rec[b]	Add	Dom	Rec[b]	Add	Dom
0	Lange z-test	1.000	.823	.864	1.000	.557	.478
	Inbred z-test	1.000	.823	.864	1.000	.557	.478
	Lange chi-square	1.000	.535	.596	1.000	.356	.254
	Inbred chi-square	1.000	.535	.596	1.000	.356	.254
0.001	Lange z-test	1.000	.818	.872	1.000	.583	.514
	Inbred z-test	1.000	.818	.872	1.000	.583	.514
	Lange chi-square	1.000	.539	.629	1.000	.378	.276
	Inbred chi-square	1.000	.539	.629	1.000	.378	.276
0.00996	Lange z-test	1.000	.840	.882	1.000	.615	.511
	Inbred z-test	1.000	.840	.882	1.000	.615	.511
	Lange chi-square	1.000	.545	.621	1.000	.369	.299
	Inbred chi-square	1.000	.545	.621	1.000	.336	.266
0.02957	Lange z-test	1.000	.856	.888	1.000	.692	.570
	Inbred z-test	1.000	.810	.847	1.000	.564	.442
	Lange chi-square	1.000	.605	.650	1.000	.479	.382
	Inbred chi-square	1.000	.520	.577	1.000	.364	.271
0.04879	Lange z-test	1.000	.849	.896	1.000	.749	.659
	Inbred z-test	1.000	.763	.809	1.000	.562	.519
	Lange chi-square	1.000	.603	.667	1.000	.582	.440
	Inbred chi-square	1.000	.439	.517	.999	.394	.284
0.09521	Lange z-test	1.000	.888	.890	1.000	.843	.780
	Inbred z-test	1.000	.748	.776	.999	.573	.500
	Lange chi-square	1.000	.691	.703	1.000	.692	.616
	Inbred chi-square	1.000	.489	.493	.996	.393	.328

[a] θ is inbreeding coefficient of the population

[b] Rec represents recessive disease model; Add represents additive disease model; Dom represents dominant disease model.

tests even when inbreeding exists. Fourth, the powers of both inbred tests were lower than the powers of the corresponding Lange tests when the inbreeding coefficient was large and the disease model was dominant and additive. The possible reason for this was that our inbred tests decreased the false positive rates and hence were more insensitive to linkage.

To verify our power and false positive rates estimations, we also performed permutation tests on our simulated data sets. Using the observed marker concordance values (Z_{rs}) of the affected sib pairs minus the expected marker concordance values ($E(Z_{rs})$), we can get a set of numbers from our original data set for the permutation test. By randomly assigning plus and minus signs to these numbers, we will get permuted data sets. Substituting these permuted numbers into the z-test statistics we described above, we can get a set of T values. The p-value of the permutation test can be computed as the ratio of the number of tests whose T values are greater or equal to the original T value divided by the total number of permutations. If the resulted p-value is less than 0.05, we reject the null hypothesis of no linkage. We performed permutation tests on both two marker alleles data sets and four marker alleles data sets and used the same set of parameters as we used in the z-tests. The resulting false positive rate and power estimates are very close to the values we get using the z-tests. This suggests that the estimations we made are correct.

In general, the power of our two inbred tests is good in the recessive disease model even when the association between disease and marker loci is not strong (data not shown here) and we suggest the use of the two inbred IBS tests only in recessive disease

model. Although the empirical false positive rates of our two inbred tests exceed the nominal value, they are improved compared to those of Lange tests, especially for large inbreeding coefficients. The power of the inbred z test is superior to that of inbred chi-square test.

2.4 Discussion

A crucial step in finding gene loci that contribute to a genetic disease is to demonstrate linkage between a candidate gene locus and a marker locus. When the mode of inheritance for the disease being studied is unknown, affected sib pair analysis will often be the method of choice. The IBD tests are powerful ASP tests when marker allele IBD relations of the sibs can be determined unambiguously. Tremendous efforts have been made to improve the power of IBD tests (Blackwelder and Elston, 1985; Knapp, Seuchter and Baur, 1994) and to modify the test statistic to deal with different cases. However, little has be done to take care of population inbreeding and relatedness. Inbreeding is an important factor that affects population structure, especially in animal and plant populations where inbreeding can be significant and genealogical relationships are easy to establish. The normal ASP tests will lead to high false positive rates when applied to inbreeding population because of relatedness of parents. Génin and Clerget-Darpoux (1996, 1998) first studied the effect of inbreeding and relatedness on ASP tests and extended the ASP methods to the situation of sib pairs sampled from a consanguineous population. However, their calculations of the extended kinship coefficients are not correct because they assumed independence

of the genes. Furthermore, their IBD tests have few practical applications. In most cases, if the IBD state of four alleles from the siblings can be identified unambiguously, the parents must not share common alleles and thus can not be descended from a common progenitor.

In IBS tests parental information is not known, thus the parents could be inbred. Our aim in this study was to modify Lange's IBS tests to allow for inbreeding. In the chi-square test, we present another way to compute the expected proportions of the three different IBS categories which does not require knowledge of expected IBD proportions. As we can see from simulation results, the false positive rates of Lange tests greatly exceed the empirical values in highly inbred populations. Thus, when the population inbreeding coefficient is large, false conclusions of linkage may result if inbreeding is ignored. The empirical false positive rates of our inbred tests are superior to those of Lange's tests for populations with a high inbreeding coefficient. Thus our inbred IBS tests should be used when the population inbreeding coefficient is large. Our results also showed that, for only slightly inbred populations, the false positive rates of Lange's IBS tests are not affected by inbreeding, which suggests that it is safe to use Lange's IBS tests when the population inbreeding coefficient is known to be small. Chi-square tests have lower power than z-tests in general. Although the power of both of inbred IBS tests are not good in dominant and additive disease models when the two alleles at the disease locus have equal initial frequencies, the powers are good in the recessive disease model. When the initial disease allele is rare, the power of all the IBS tests in dominant and additive disease model increase (data

not shown here). However, in highly inbred populations with affected individuals, it is unlikely that the disease allele has a low initial frequency because such a disease allele will become lost with random genetic drift.

Another important conclusion we obtained is that the IBS tests we studied are tests of both linkage and association. Association mapping has a better resolution than linkage mapping. Ideally, only two closely linked loci will show association. Therefore, if we find one marker locus that rejects the null hypothesis, it is highly likely that it is both in linkage and association with the disease locus, which suggests that it is much closer to the disease locus than the marker locus linked but in linkage equilibrium with the disease locus.

Although ASP tests are thought to be robust for all the genetic models, our simulation results show that the powers of the IBS tests are good only in recessive disease model and they are not good in additive or dominant disease models. It appears that the inbred IBS tests could serve as a good screen only when the disease model are known to be recessive. Thus, we suggest to use the inbred IBS tests if we know the population is highly inbred and the disease model is recessive.

In this paper, we have demonstrated one way to use population inbreeding information to refine ASP tests. Other more powerful ASP tests will become more accurate in mapping a disease locus after some modification which accounts for inbreeding.

Chapter 3

The effect of higher-order descent measures on expectation of identical by descent proportions in inbred populations

3.1 Introduction

Génin and Clerget-Darpoux (1996) first studied the effect of inbreeding on traditional affected sib pair (ASP) tests and extended ASP methods to the situation of sib pairs sampled from a consanguineous population. In their paper, they computed the probabilities of nine condensed identity states of two individuals as a function of the inbreeding coefficient and calculated the expected identical by descent (IBD) proportions in inbreeding population. They also derived the identity state probabilities for sib pairs from first-cousin mating sampled from an inbreeding population and constructed a "N_a test". Their overall idea was novel and good, but there are

errors in their calculation. Weeks and Sinsheimer (1998) pointed out that Génin and Clerget-Darpoux derived "IBW-state probabilities" (IBW refers to IBD alleles between and within individuals) that could not satisfy two consistency checks (Jacqard, 1974; Karigl, 1981) and their derived extended kinship coefficients for two sibs from a first-cousin marriage have been criticized. Génin and Clerget-Darpoux recalculated the IBW-state probabilities and kinship coefficients in their reply (1998) to Weeks and Sinsheimer's letter. Génin and Clerget-Darpoux computed eight extended-kinship coefficients of parents randomly selected from an inbreeding population using formulas presented by Karigl (1981) and derived IBW-state probabilities by use of the inverse of matrix D (as given in Table 3.7) multiplied by the kinship coefficient vector. However, their calculation of extended kinship coefficients leads to questionable derived IBW-state probabilities.

The coancestry coefficient is the probability that two alleles taken at random from two individuals are IBD. The inbreeding coefficient, on the other hand, is the probability that two alleles from the same individual are IBD. For a dioecious mating system, by definition, the inbreeding coefficient in one generation is the same as the coancestry coefficient in the previous generation. However, in large populations with random mating, the difference between the coancestry coefficient and the inbreeding coefficient is small and the coancestry coefficient can be approximated by the inbreeding coefficient. Thus the IBD status of pairs of alleles is the same whether they came from the same or different individuals. In a random mating population, the probability that any two alleles at a locus are IBD is the inbreeding coefficient θ.

However, the probabilities that any three or four alleles at a locus are IBD are not simply products of θ since the alleles are not independent of each other. Instead, the IBD relations for three and four alleles should be expressed in terms of higher-order descent measures (Cockerham, 1971). Define γ, δ, Δ to be the probability that three, four and two pairs respectively of alleles selected at random from a population are IBD. These quantities can be expressed as functions of θ when the population is in an evolutionary equilibrium (Li, 1996). However, they are quite different from θ^2, θ^3 and θ^2, which were used by Génin and Clerget-Darpoux throughout their paper. Both Weeks and Sinsheimer (1998) and Cannings (1998) pointed out errors in the IBW-state probabilities of Génin and Clerget-Darpoux and recomputed the allele-identity states probabilities. We believe their calculations are not correct either because they also assumed independence of some of the alleles.

In this paper, we derive the probabilities for the fifteen detailed identity states for two sibs randomly selected from an inbreeding population and the probabilities for the nine condensed identity states for sib pairs from a first-cousin marriage sampled from an inbreeding population using higher-order decent measures and we simplify them as functions of θ.

3.2 Identity state probabilities for two sibs randomly sampled in an inbreeding population

For the four alleles at one locus of two individuals, it is well known that there are fifteen possible identity states (Harris, 1964; Gillois, 1965; Cockerham, 1971). We consider the fifteen identity states are more informative than nine condensed identity states and thus we compute the probabilities of all fifteen. Define δ_0 to be the probability that none of the four alleles are IBD; δ_{ij} to be the probability that only two (i, j) of the four alleles are IBD, δ_{ijk} to be the probability that only three (i, j, k) of the four alleles are IBD and δ_{ijkl} to be the probability that all four alleles are IBD. Further, $\delta_{ij.kl}$ is the probability that two pairs of the four alleles (i, j and k, l) are IBD. We name the four alleles to be (a_{11}, a_{12}, a_{21}, a_{22}) according to the method of Thompson (1974), where the first subscript indicates the first or second individual and the second subscript indicates the paternal or maternal allele. We give the IBD alleles the same label. Different (a_{11}, a_{12}, a_{21}, a_{22}) corresponds to different IBW state. For example, IBW state (1122) stands for $a_{11} = 1$, $a_{12} = 1$, $a_{21} = 2$, and $a_{22} = 2$, which means that the paternal and maternal alleles of the first individual are IBD, and the paternal and maternal alleles of the second individual are IBD, but the alleles of the first individual are not IBD with the alleles of the second individual. These fifteen IBW probabilities can be expressed as functions of θ. Table 3.1 lists the fifteen possible identity relationships of two parents and their probabilities in general, and in a completely random-mating case, and in an evolutionary equilibrium case.

Table 3.1: Descent relations among alleles for two parents: parent X with alleles a and b and parent Y with alleles c and d

IBW states $(a_{11}, a_{12}, a_{21}, a_{22})$	Pr(IBD)	Random	Equilibrium*	IBD
$S_1(1111)$	δ_{abcd}	δ	$6\theta^3$	2
$S_2(1122)$	$\delta_{ab.cd}$	$\Delta - \delta$	$\theta^2(1-\theta)$	0
$S_3(1212)$	$\delta_{ac.bd}$	$\Delta - \delta$	$\theta^2(1-\theta)$	2
$S_4(1221)$	$\delta_{ad.bc}$	$\Delta - \delta$	$\theta^2(1-\theta)$	2
$S_5(1112)$	δ_{abc}	$\gamma - \delta$	$2\theta^2(1-\theta)$	1
$S_6(1121)$	δ_{abd}	$\gamma - \delta$	$2\theta^2(1-\theta)$	1
$S_7(1211)$	δ_{acd}	$\gamma - \delta$	$2\theta^2(1-\theta)$	1
$S_8(2111)$	δ_{bcd}	$\gamma - \delta$	$2\theta^2(1-\theta)$	1
$S_9(1123)$	δ_{ab}	$\theta - 2\gamma - \Delta + 2\delta$	$\theta(1-\theta)^2$	0
$S_{10}(2311)$	δ_{cd}	$\theta - 2\gamma - \Delta + 2\delta$	$\theta(1-\theta)^2$	0
$S_{11}(1213)$	δ_{ac}	$\theta - 2\gamma - \Delta + 2\delta$	$\theta(1-\theta)^2$	1
$S_{12}(1231)$	δ_{ad}	$\theta - 2\gamma - \Delta + 2\delta$	$\theta(1-\theta)^2$	1
$S_{13}(2113)$	δ_{bc}	$\theta - 2\gamma - \Delta + 2\delta$	$\theta(1-\theta)^2$	1
$S_{14}(2131)$	δ_{bd}	$\theta - 2\gamma - \Delta + 2\delta$	$\theta(1-\theta)^2$	1
$S_{15}(1234)$	δ_0	$1 - 6\theta + 8\gamma + 3\Delta - 6\delta$	$(1-\theta)^3$	0

* Each term has been multiplied by $(1+\theta)(1+2\theta)$.

The sib pair identity state probabilities can be easily derived conditional on the 15 possible parental mating types, as shown in Table 3.2.

Table 3.2: Sib pair identity state probabilities conditional on parental mating types

sib IBW	parental IBW state														
state	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
S_1	1	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_2	0	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0	0	0	0	0	0	0	0	0	0
S_3	0	1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{4}$
S_4	0	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0	0	0	0	0	0	0	0	0	0
S_5	0	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_6	0	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0	$\frac{1}{4}$	$\frac{1}{4}$	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_7	0	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_8	0	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0	$\frac{1}{4}$	$\frac{1}{4}$	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_9	0	0	0	0	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_{10}	0	0	0	0	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_{11}	0	0	0	0	0	0	0	0	$\frac{1}{2}$	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
S_{12}	0	0	0	0	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_{13}	0	0	0	0	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	0
S_{14}	0	0	0	0	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
S_{15}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{4}$

The identity-state probabilities for the sib pair are thus the product of sib pair conditional identity state probabilities and parental mating type probabilities. Table 3.3 shows the sib pair identity state probabilities.

Karigl (1981) presented a matrix K that related the fifteen detailed identity coefficients δ_1, ..., δ_{15} to the parental kinship coefficients if neither of the two individuals under consideration is an ancestor of the other one. Let P_k denote the probability for identity state S_k (k=1-15). Let I_{ij} represent the vector of P_k vector for individual i and individual j. Let V_{ij} stand for the vector of kinship coefficients for the parents of individual i and individual j. Then matrix K multiplied by I_{ij} equals V_{ij}. Matrix K, vectors I_{ij} and V_{ij} are described in the Table 3.4.

Let individual 1 and 2 be the parents of sib pair 3 and 4. Let individual f, m denote parents of individual 1 and let $\bar{f}$, $\bar{m}$ denote parents of individual 2. The fifteen extended-kinship coefficients in vector V_{12} can be calculated as follow:

- $\Phi_{fm\bar{f}\bar{m}} = \delta$
- $\Phi_{fm\bar{f}} = \Phi_{fm\bar{m}} = \Phi_{f\bar{f}\bar{m}} = \Phi_{m\bar{f}\bar{m}} = \gamma$
- $\Phi_{fm,\bar{f}\bar{m}} = \Phi_{f\bar{f},m\bar{m}} = \Phi_{f\bar{m},m\bar{f}} = \Delta$
- $\Phi_{fm} = \Phi_{\bar{f}\bar{m}} = \Phi_{f\bar{f}} = \Phi_{f\bar{m}} = \Phi_{m\bar{f}} = \Phi_{m\bar{m}} = \theta$

where γ, δ, Δ can be expressed as functions of θ (Li, 1996). The fifteen parental identity state probabilities as listed in Table 3.1 satisfy Karigl's linear constraints as expected.

Likewise, the fifteen extended-kinship coefficients of parent 1 and 2 can be computed using the recursive methods as shown in Table 3.5.

Table 3.3: Sib pair identity state probabilities

IBW state	Probability*
$S_1(1,1,1,1)$	$\theta(16\theta^2+7\theta+1)$
$S_2(1,1,2,2)$	$\theta^2(1-\theta)$
$S_3(1,2,1,2)$	$(1-\theta)(7\theta^2+5\theta+1)$
$S_4(1,2,2,1)$	$\theta^2(1-\theta)$
$S_5(1,1,1,2)$	$\theta(1-\theta)(1+4\theta)$
$S_6(1,1,2,1)$	$\theta(1-\theta)(1+4\theta)$
$S_7(1,2,1,1)$	$\theta(1-\theta)(1+4\theta)$
$S_8(2,1,1,1)$	$\theta(1-\theta)(1+4\theta)$
$S_9(1,1,2,3)$	$\theta(1-\theta)^2$
$S_{10}(2,3,1,1)$	$\theta(1-\theta)^2$
$S_{11}(1,2,1,3)$	$(1-\theta)^2(1+3\theta)$
$S_{12}(1,2,3,1)$	$\theta(1-\theta)^2$
$S_{13}(2,1,1,3)$	$\theta(1-\theta)^2$
$S_{14}(2,1,3,1)$	$(1-\theta)^2(1+3\theta)$
$S_{15}(1,2,3,4)$	$(1-\theta)^3$

* Each term has been multiplied by $4(1+\theta)(1+2\theta)$.

Table 3.4: Matrix K and vector I and V

K															I_{ij}	V_{ij}^*
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	P_1	$\Phi_{fm\bar{f}\bar{m}}$
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	P_5	$\Phi_{fm\bar{f}}$
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	P_6	$\Phi_{fm\bar{m}}$
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	P_7	$\Phi_{f\bar{f}\bar{m}}$
1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	P_8	$\Phi_{m\bar{f}\bar{m}}$
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	P_2	$\Phi_{fm,\bar{f}\bar{m}}$
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	P_9	$\Phi_{f\bar{f},m\bar{m}}$
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	P_{10}	$\Phi_{f\bar{m},m\bar{f}}$
1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	P_3	Φ_{fm}
1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	P_{14}	$\Phi_{\bar{f}\bar{m}}$
1	1	0	1	0	0	0	0	1	1	0	0	0	0	0	P_{11}	$\Phi_{f\bar{f}}$
1	0	1	1	0	0	0	0	0	0	0	1	1	0	0	P_4	$\Phi_{f\bar{m}}$
1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	P_{12}	$\Phi_{m\bar{f}}$
1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	P_{13}	$\Phi_{m\bar{m}}$
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	P_{15}	1

*f, m and $\bar{f}$, $\bar{m}$ denote the parents of individuals i and j.

Table 3.5: Kinship coefficients for parent 1 and parent 2

kinship coefficients	probability in terms of θ
Φ_{11}, Φ_{22}	$= \frac{1}{2}(1 + \Phi_{\bar{f}\bar{m}})$
	$= \frac{1}{2}(1 + \theta)$
	$= (4\theta^3 + 10\theta^2 + 8\theta + 2)^*$
Φ_{12}, Φ_{21}	$= \theta$
	$= (8\theta^3 + 12\theta^2 + 4\theta)^*$
$\Phi_{121}, \Phi_{122}, \Phi_{112}, \Phi_{212}$	$= \frac{1}{2}(\Phi_{12} + \Phi_{fm2})$
	$= \frac{1}{2}(\theta + \gamma)$
	$= (12\theta^3 + 10\theta^2 + 2\theta)^*$
Φ_{1212}	$= \frac{1}{2}(\Phi_{122} + \Phi_{fm22})$
	$= \frac{1}{2}\Phi_{122} + \frac{1}{4}(\Phi_{2fm} + \Phi_{\bar{f}\bar{m}fm})$
	$= \frac{1}{4}(\theta + 2\gamma + \delta)$
	$= (16\theta^3 + 7\theta^2 + \theta)^*$
$\Phi_{11,22}$	$= \frac{1}{2}(\Phi_{22} + \Phi_{fm,22})$
	$= \frac{1}{2}\Phi_{22} + \frac{1}{4}(\Phi_{fm} + \Phi_{\bar{f}\bar{m},fm})$
	$= \frac{1}{4}(1 + \theta + \theta + \Delta)$
	$= (9\theta^3 + 9\theta^2 + 5\theta + 1)^*$
$\Phi_{12,21}, \Phi_{12,12}$	$= \frac{1}{4}(2\Phi_{122} + 2\Phi_{f2,m2})$
	$= \frac{1}{2}\Phi_{122} + \frac{1}{8}(2\Phi_{2fm} + \Phi_{\bar{f}f,\bar{m}m} + \Phi_{\bar{m}f,\bar{f}m})$
	$= \frac{1}{4}(\theta + \gamma) + \frac{1}{8}(2\gamma + 2\Delta)$
	$= (15\theta^3 + 8\theta^2 + \theta)^*$

*Each term has been multiplied by $4(1 + \theta)(1 + 2\theta)$

Using matrix K multiply the fifteen identity-state probabilities of the sib pairs as listed in Table 3.3, we can get the set of extended kinship coefficients for parent 1 and parent 2, which are exactly the same as we computed using the recursive methods. Thus, our fifteen identity state probabilities verify one consistency check discussed by Weeks and Sinsheimer (1998).

Weeks and Sinsheimer (1998) did another consistency check: they used a formula presented by Jackquard (1974) to compute the kinship coefficient between siblings, Φ_{34} and claimed that it should equal $(1+3\theta)/4$, which was derived using a classical recursion method. Let Δ_i represent the probabilities of the nine condensed sib pair identity states. All the Δ_i can be expressed in terms of P_k. Thus we can also calculate Φ_{34} using Jacquard's formula:

$$\begin{aligned}
\Phi_{34} &= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8 \\
&= P_1 + \frac{1}{2}(P_5 + P_6 + P_7 + P_8 + P_3 + P_4) + \frac{1}{4}(P_{11} + P_{12} + P_{13} + P_{14}) \\
&= \frac{1}{4} + \frac{3}{4}\theta
\end{aligned}$$

Thus, our fifteen sib pair identity probabilities verify the two consistency checks performed by Weeks and Sinsheimer.

Once the identity state probabilities of the sib pairs are known, the expected IBD proportions of the sib pairs under the null hypothesis of no linkage can be easily obtained. If the two sibs are in identity states S_1, S_3, S_4, they share two IBD alleles. If the sibs are in identity states $S_5, S_6, S_7, S_8, S_{11}, S_{12}, S_{13}, S_{14}$, they share one IBD allele. Otherwise, if the sibs are in identity states S_2, S_9, S_{10}, S_{15}, they share zero IBD allele. Table 3.6 describes the probabilities for the three IBD category.

Table 3.6: Expected IBD proportions in inbreeding population

IBD Category	Probabilities($\theta \geq 0$)*	Probabilities($\theta = 0$)
IBD=2	$(1+2\theta)(4\theta^2+3\theta+1)$	$\frac{1}{4}$
IBD=1	$2(1-\theta^2)(1+4\theta)$	$\frac{1}{2}$
IBD=0	$1-\theta$	$\frac{1}{4}$

* Each term has been multiplied by $4(1+\theta)(1+2\theta)$.

3.3 Identity state probabilities for two sibs from first cousin marriage

In Génin and Clerget-Darpoux (1996, 1998) paper, they computed extended kinship coefficients for first cousins. We will not utilize the result of their calculation since they assumed independence of the alleles. Here we recalculate these kinship coefficient in terms of higher-order descent measures (Cockerham, 1971). The pedigree in which kinship coefficients have been computed is shown in Figure 3.1.

The kinship coefficient Φ_{77}, Φ_{88}, Φ_{78}, Φ_{87} in Génin and Clerget-Darpoux paper, involve only two alleles and thus are correct:

$\Phi_{77} = \Phi_{88} = \frac{1}{2}(1+\theta)$

$\Phi_{78} = \Phi_{87} = \frac{1}{16}(1+15\theta)$

To compute other kinship coefficients of first cousin, we must calculate some intermediate kinship coefficients first:

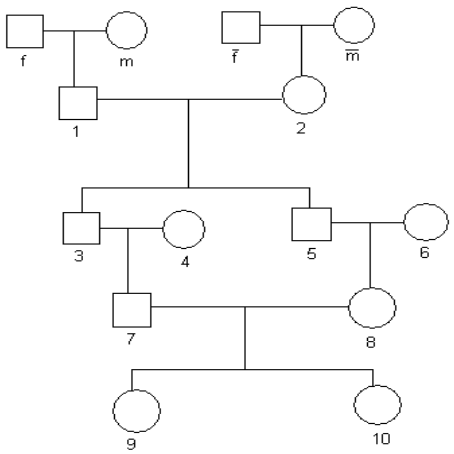

Figure 3.1: Pedigree, individual 7 and 8 are first cousins, individual 9 and 10 are sib pair from first cousin marriage.

$$\begin{aligned}
\Phi_{345} &= \frac{1}{2}(\Phi_{145} + \Phi_{245}) \\
&= \frac{1}{4}(\Phi_{141} + \Phi_{142}) + \frac{1}{4}(\Phi_{241} + \Phi_{242}) \\
&= \frac{1}{8}(\Phi_{14} + \Phi_{fm4}) + \frac{1}{2}\Phi_{142} + \frac{1}{8}(\Phi_{24} + \Phi_{\bar{f}\bar{m}4}) \\
&= \frac{1}{4}\theta + \frac{3}{4}\gamma
\end{aligned}$$

$$\begin{aligned}
\Phi_{3456} &= \frac{1}{2}(\Phi_{1456} + \Phi_{2456}) \\
&= \frac{1}{4}(\Phi_{1416} + \Phi_{1426}) + \frac{1}{4}(\Phi_{2416} + \Phi_{2426}) \\
&= \frac{1}{8}(\Phi_{146} + \Phi_{fm46}) + \frac{1}{2}\Phi_{1426} + \frac{1}{8}(\Phi_{246} + \Phi_{\bar{f}\bar{m}46}) \\
&= \frac{1}{8}(\gamma + \delta) + \frac{1}{2}\delta + \frac{1}{8}(\gamma + \delta) \\
&= \frac{1}{4}\gamma + \frac{3}{4}\delta
\end{aligned}$$

$$\Phi_{34,56} = \Phi_{36,45}$$

$$
\begin{aligned}
&= \frac{1}{4}(\Phi_{14,56} + \Phi_{24,56} + \Phi_{34,16} + \Phi_{34,26}) \\
&= \frac{1}{8}(\Phi_{14,16} + \Phi_{14,26}) + \frac{1}{8}(\Phi_{24,16} + \Phi_{24,26}) + \frac{1}{8}(\Phi_{14,16} + \Phi_{24,16}) \\
&\quad + \frac{1}{8}(\Phi_{14,26} + \Phi_{24,26}) \\
&= \frac{1}{16}(2\Phi_{146} + \Phi_{f4,m6} + \Phi_{m4,f6}) + \frac{1}{16}(2\Phi_{246} + \Phi_{\bar{f}4,\bar{m}6} + \Phi_{\bar{m}4,\bar{f}6}) + \frac{1}{4}\Delta + \frac{1}{4}\Delta \\
&= \frac{1}{4}\gamma + \frac{3}{4}\Delta
\end{aligned}
$$

$$
\begin{aligned}
\Phi_{35,46} &= \frac{1}{2}(\Phi_{15,46} + \Phi_{25,46}) \\
&= \frac{1}{4}(\Phi_{11,46} + \Phi_{12,46}) + \frac{1}{4}(\Phi_{21,46} + \Phi_{22,46}) \\
&= \frac{1}{8}(\Phi_{46} + \Phi_{fm,46}) + \frac{1}{8}(\Phi_{46} + \Phi_{\bar{f}\bar{m},46}) + \frac{1}{2}\Delta \\
&= \frac{1}{4}\theta + \frac{3}{4}\Delta
\end{aligned}
$$

Then it is easy to derive the first cousin kinship coefficient:

$$
\begin{aligned}
\Phi_{778} &= \Phi_{887} \\
&= \frac{1}{2}(\Phi_{78} + \Phi_{348}) \\
&= \frac{1}{2}\Phi_{78} + \frac{1}{4}(\Phi_{345} + \Phi_{346}) \\
&= \frac{1}{32}(1 + 15\theta) + \frac{1}{4}(\frac{1}{4}\theta + \frac{3}{4}\gamma) + \frac{1}{4}\gamma \\
&= \frac{1}{32} + \frac{17}{32}\theta + \frac{7}{16}\gamma
\end{aligned}
$$

$$
\begin{aligned}
\Phi_{7788} &= \frac{1}{2}(\Phi_{788} + \Phi_{3488}) \\
&= \frac{1}{2}\Phi_{788} + \frac{1}{4}(\Phi_{348} + \Phi_{3456})
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\Phi_{788} + \frac{1}{8}(\Phi_{345} + \Phi_{346}) + \frac{1}{4}\Phi_{3456} \\
&= \frac{1}{64} + \frac{19}{64}\theta + \frac{1}{2}\gamma + \frac{3}{16}\delta
\end{aligned}
$$

$$
\begin{aligned}
\Phi_{77,88} &= \frac{1}{2}(\Phi_{88} + \Phi_{34,88}) \\
&= \frac{1}{2}\Phi_{88} + \frac{1}{4}(\Phi_{34} + \Phi_{34,56}) \\
&= \frac{1}{4} + \frac{1}{2}\theta + \frac{1}{16}\gamma + \frac{3}{16}\Delta
\end{aligned}
$$

$$
\begin{aligned}
\Phi_{78,78} &= \frac{1}{4}(2\Phi_{788} + 2\Phi_{38,48}) \\
&= \frac{1}{2}\Phi_{788} + \frac{1}{8}(2\Phi_{348} + \Phi_{35,46} + \Phi_{36,45}) \\
&= \frac{1}{2}\Phi_{788} + \frac{1}{8}(\Phi_{345} + \Phi_{346}) + \frac{1}{8}\Phi_{35,46} + \frac{1}{8}\Phi_{36,45} \\
&= \frac{1}{64} + \frac{21}{64}\theta + \frac{15}{32}\gamma + \frac{3}{16}\Delta
\end{aligned}
$$

Once these extended kinship coefficients of first cousin are known, the identity probabilities for individuals 7 and 8 can be easily deduced. Here we will compute only the nine condensed identity state probabilities as described by Génin and Clerget-Darpoux for simplicity. Karigl (1981) also gave a matrix D that relates the nine condensed identity state probability vector E_{ij} to the extended kinship coefficient vector Q_{ij}. Matrix D and the vectors E_{ij} and Q_{ij} are presented in table 3.7.

The nine identity state probabilities for first cousin can be obtained by multiplying the inverse of matrix D and vector Q_{78} as follow:

$P(S_1) = \Phi_{78} - 2\Phi_{778} - 2\Phi_{788} + 4\Phi_{7788} = \frac{1}{4}\gamma + \frac{3}{4}\delta$

$P(S_2) = 1 - 2\Phi_{77} - 2\Phi_{88} - \Phi_{78} + 2\Phi_{778} + 2\Phi_{788} - 4\Phi_{7788} + 4\Phi_{77,88} = \frac{3}{4}\Delta - \frac{3}{4}\delta$

Table 3.7: Matrix D and vector E and Q

D									E_{ij}	Q^*_{ij}
1	1	1	1	1	1	1	1	1	$P(S_1)$	1
2	2	2	2	1	1	1	1	1	$P(S_2)$	$2\Phi_{ii}$
2	2	1	1	2	2	1	1	1	$P(S_3)$	$2\Phi_{jj}$
4	0	2	0	2	0	2	1	0	$P(S_4)$	$4\Phi_{ij}$
8	0	4	0	2	0	2	1	0	$P(S_5)$	$8\Phi_{iij}$
8	0	2	0	4	0	2	1	0	$P(S_6)$	$8\Phi_{ijj}$
16	0	4	0	4	0	2	1	0	$P(S_7)$	$16\Phi_{iijj}$
4	4	2	2	2	2	1	1	1	$P(S_8)$	$4\Phi_{ii,jj}$
16	0	4	0	4	0	4	1	0	$P(S_9)$	$16\Phi_{ij,ij}$

$$P(S_3) = -4\Phi_{78} + 8\Phi_{778} + 4\Phi_{788} - 8\Phi_{7788} = \tfrac{1}{4}\theta + \tfrac{5}{4}\gamma - \tfrac{3}{2}\delta$$

$$\begin{aligned} P(S_4) &= -2 + 4\Phi_{77} + 2\Phi_{88} + 4\Phi_{78} - 8\Phi_{778} - 4\Phi_{788} + 8\Phi_{7788} - 4\Phi_{77,88} \\ &= \frac{3}{4}\theta - \frac{3}{2}\gamma + \frac{3}{2}\delta - \frac{3}{4}\Delta \end{aligned}$$

$$P(S_5) = -4\Phi_{78} + 4\Phi_{778} + 8\Phi_{788} - 8\Phi_{7788} = \tfrac{1}{4}\theta + \tfrac{5}{4}\gamma - \tfrac{3}{2}\delta$$

$$\begin{aligned} P(S_6) &= -2 + 2\Phi_{77} + 4\Phi_{88} + 4\Phi_{78} - 4\Phi_{778} - 8\Phi_{788} + 8\Phi_{7788} - 4\Phi_{77,88} \\ &= \frac{3}{4}\theta - \frac{3}{2}\gamma + \frac{3}{2}\delta - \frac{3}{4}\Delta \end{aligned}$$

$$P(S_7) = -8\Phi_{7788} + 8\Phi_{78,78} = \tfrac{1}{4}\theta - \tfrac{1}{4}\gamma - \tfrac{3}{2}\delta + \tfrac{3}{2}\Delta$$

$$\begin{aligned} P(S_8) &= 16\Phi_{78} - 16\Phi_{778} - 16\Phi_{788} + 32\Phi_{7788} - 16\Phi_{78,78} \\ &= \frac{1}{4} + \frac{9}{4}\theta - \frac{11}{2}\gamma + 6\delta - 3\Delta \end{aligned}$$

Table 3.8: Sib pair identity state conditional on parental mating types (Génin and Clerget-Darpoux (1996))

	Parents identity state								
Sibs identity state	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
S_1	1	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{8}$	$\frac{1}{16}$	0
S_2	0	0	0	0	0	0	$\frac{1}{8}$	0	0
S_3	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{8}$	0
S_4	0	0	0	0	0	0	0	$\frac{1}{16}$	0
S_5	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{8}$	0
S_6	0	0	0	0	0	0	0	$\frac{1}{16}$	0
S_7	0	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{4}$
S_8	0	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{3}{8}$	$\frac{1}{2}$
S_9	0	0	0	0	0	0	0	0	$\frac{1}{4}$

$$\begin{aligned} P(S_9) &= 4 - 4\Phi_{77} - 4\Phi_{88} - 16\Phi_{78} + 16\Phi_{778} + 16\Phi_{788} - 24\Phi_{7788} + 4\Phi_{77,88} + 8\Phi_{78,78} \\ &= \frac{3}{4} - \frac{9}{2}\theta + 6\gamma - \frac{9}{2}\delta + \frac{9}{4}\Delta \end{aligned}$$

Génin and Clerget-Darpoux (1996) showed sib-pair identity probabilities conditional on parental mating types. It is easy to derive the identity state probabilities for sib pairs from first cousin marriage using conditional identity probabilities matrix M_{ps}, as described in Table 3.8 and the nine identity state probabilities of first cousin. Table 3.9 shows identity state probabilities for sib pairs from first cousin marriage.

Again, we can check these sib pair identity probabilities from first cousin mating using Karigl's matrix (1981). The kinship coefficient of individuals 9 and 10 can be

Table 3.9: Identity state probabilities of sib pairs from first cousin marriage

Identity state	Prob in higher order terms	Equilibrium prob*
$S_1(1111)$	$\frac{1}{64} + \frac{19}{64}\theta + \frac{1}{2}\gamma + \frac{3}{16}\delta$	$238\theta^3 + 123\theta^2 + 22\theta + 1$
$S_2(1122)$	$\frac{1}{32}\theta - \frac{1}{32}\gamma - \frac{3}{16}\delta + \frac{3}{16}\Delta$	$2\theta(1-\theta)(1+8\theta)$
$S_3(1112)$	$\frac{1}{32} + \frac{15}{32}\theta - \frac{1}{8}\gamma - \frac{3}{8}\delta$	$2(1-\theta)(58\theta^2 + 19\theta + 1)$
$S_4(1123)$	$\frac{1}{64} + \frac{9}{64}\theta - \frac{11}{32}\gamma + \frac{3}{8}\delta - \frac{3}{16}\Delta$	$(1+14\theta)(1-\theta)^2$
$S_5(1222)$	$\frac{1}{32} + \frac{15}{32}\theta - \frac{1}{8}\gamma - \frac{3}{8}\delta$	$2(1-\theta)(58\theta^2 + 19\theta + 1)$
$S_6(1233)$	$\frac{1}{64} + \frac{9}{64}\theta - \frac{11}{32}\gamma + \frac{3}{8}\delta - \frac{3}{16}\Delta$	$(1+14\theta)(1-\theta)^2$
$S_7(1212)$	$\frac{15}{64} + \frac{15}{64}\theta - \frac{15}{32}\gamma - \frac{3}{8}\delta + \frac{3}{8}\Delta$	$3(1-\theta)(38\theta^2 + 25\theta + 5)$
$S_8(1213)$	$\frac{15}{32} - \frac{21}{32}\theta - \frac{9}{16}\gamma + \frac{3}{2}\delta - \frac{3}{4}\Delta$	$6(5+18\theta)(1-\theta)^2$
$S_9(1234)$	$\frac{3}{16} - \frac{9}{8}\theta + \frac{3}{2}\gamma - \frac{9}{8}\delta + \frac{9}{16}\Delta$	$12(1-\theta)^3$

*Each term has been multiplied by $64(1+\theta)(1+2\theta)$

easily obtained:

$$\begin{aligned}\Phi_{99} &= \Phi_{1010} \\ &= \frac{1}{2}(1+\Phi_{78}) \\ &= \frac{17}{32}+\frac{15}{32}\theta\end{aligned}$$

$$\begin{aligned}\Phi_{910} &= \frac{1}{4}(\Phi_{77}+\Phi_{78}+\Phi_{87}+\Phi_{88}) \\ &= \frac{9}{32}+\frac{23}{32}\theta\end{aligned}$$

$$\begin{aligned}\Phi_{9910} &= \Phi_{91010} \\ &= \frac{1}{2}(\Phi_{910}+\Phi_{7810}) \\ &= \frac{1}{2}\Phi_{910}+\frac{1}{4}(\Phi_{787}+\Phi_{788}) \\ &= \frac{5}{32}+\frac{5}{8}\theta+\frac{7}{32}\gamma\end{aligned}$$

$$\begin{aligned}\Phi_{991010} &= \frac{1}{2}(\Phi_{91010}+\Phi_{781010}) \\ &= \frac{1}{2}\Phi_{91010}+\frac{1}{4}(\Phi_{7810}+\Phi_{7878}) \\ &= \frac{23}{256}+\frac{133}{256}\theta+\frac{11}{32}\gamma+\frac{3}{64}\delta\end{aligned}$$

$$\begin{aligned}\Phi_{99,1010} &= \frac{1}{2}(\Phi_{1010}+\Phi_{78,1010}) \\ &= \frac{1}{2}\Phi_{1010}+\frac{1}{4}(\Phi_{78}+\Phi_{78,78}) \\ &= \frac{73}{256}+\frac{141}{256}\theta+\frac{15}{128}\gamma+\frac{3}{64}\Delta\end{aligned}$$

$$\begin{aligned}\Phi_{910,910} &= \frac{1}{4}(2\Phi_{91010} + 2\Phi_{710,810}) \\ &= \frac{1}{2}\Phi_{91010} + \frac{1}{8}(2\Phi_{7810} + \Phi_{77,88} + \Phi_{78,87}) \\ &= \frac{61}{512} + \frac{281}{512}\theta + \frac{73}{256}\gamma + \frac{3}{64}\Delta\end{aligned}$$

We can get exactly the same set of extended kinship coefficients by using the matrix D times our identity probabilities of sib pairs from first cousin mating.

When computing Φ_{910} using Jacquard's formula, we can get:

$\Phi_{910} = P(S_1) + \frac{1}{2}(P(S_3) + P(S_5) + P(S_7)) + \frac{1}{4}P(S_8) = \frac{9}{32} + \frac{23}{32}\theta$

which equal the kinship coefficient of individuals 9 and 10 which we obtained using the classical method. Thus, our identity coefficients of sib pairs from first cousin marriages also verify two consistency checks performed by Weeks and Sinsheimer.

3.4 Discussion

Identifying disease susceptibility genes through affected sib-pair analysis has been one of the most striking success of the genome project to date. Traditional ASP methods assume no inbreeding and relatedness in the population. However, there will be a low level of inbreeding in finite populations because of evolutionary history. Traditional ASP methods will lead to a high false positive rate if applied to inbreeding populations. Génin and Clerget-Darpoux first studied the effect of inbreeding on traditional ASP tests and extended the ASP methods to the situation of sib pairs sampled from a consanguineous population (1996, 1998). However, there are errors in their calculations because they assumed independence of some of the alleles.

In this paper, we first computed the fifteen detailed sib pair identity states probabilities using higher-order decent measures when sibs are sampled from an inbreeding population, and then derived the expected proportions of sib pairs sharing zero, one, or two marker alleles. Our fifteen detailed sib pair identity states probabilities verify the two consistency checks performed by Weeks and Sinsheimer (1998). Our results are quite different from that of Génin and Clerget-Darpoux's. We believe our results are correct since we take into account allele dependency. In the second part of the paper, we compute the nine condensed identity state probabilities of sib-pairs from first cousin marriages. These results also satisfy Weeks and Sinsheimer's two consistency checks and are different from the results obtained by Génin and Clerget-Darpoux. From the results, it is easy to see that we got almost the same set of coefficients as Génin and Clerget-Darpoux. As expected, we just need to replace α^3 with δ and replace α^2 with γ or Δ. Cannings also pointed out that the identity state probabilities computed by Génin and Clerget-Darpoux are incorrect. He recomputed the nine identity state probabilities in two different ways. Although he argued that α^3 was not appropriate for indicating all four alleles being IBD, he claimed in his Table 1 that the correct probability should be $\alpha^2\beta$ which still assumed independence of some of the alleles and therefore is incorrect. In his second way of computing the identity state probabilities, he used a set of parameters α_2, α_3 and α_4 to represent the probabilities of two, three and four alleles being IBD. His α_2 is the same as our θ, α_3 is the same as our γ and α_4 is the same as our δ. However, he used α^2 to indicate the probability of two pairs of alleles being IBD. This is incorrect because again, he

assumed independence of some of the alleles.

Génin and Clerget-Darpoux (1996) modified three ASP tests (Blackwelder, 1985) after they recalculated the three expected sib pair IBD proportions. They also did a "N_a test" for sib-pairs from first cousin marriage. However, their tests have few practical applications. In most cases, if the IBD state of four alleles from the siblings can be identified unambiguously, the parents must not share common alleles and thus can not be inbred. The whole idea of population inbreeding on IBD tests will be valid only when the marker alleles are not highly polymorphic or the parents are missing. Population inbreeding will be more relevant when identity by state tests are applied since IBS tests relax the IBD relations by substituting sib pair IBS relations. Parental information is no longer needed and the population could be inbred to any degree.

Chapter 4

Linkage and linkage disequilibrium analysis for a single QTL using family data

4.1 Introduction

Identification of disease susceptibility genes is one of the primary aims of contemporary genetic research. Advances in recent decades in molecular genetics enable more and more DNA polymorphisms to be used as markers for linkage and association analysis. The basis for association mapping is that a chromosome with a disease allele should show a distinctive haplotype in the vicinity of the disease locus similar to the haplotype of the ancestral chromosome. The basis for linkage analysis is that affected individuals in a pedigree should show a different marker segregation pattern than expected at marker loci linked to a disease locus. Linkage analysis has been successfully used to map major genes causing simple Mendelian disorders, such as Duchenne muscular dystrophy (Murray et al., 1982; Monaco and Kunkel, 1988)

and cystic fibrosis (Kerem et al., 1989; Dean et al., 1990). However, most human diseases are complex diseases caused by genes at multiple loci. Association analysis has had success in mapping genes or regions affecting complex diseases such as the corticotropin-releasing hormone genomic region in rheumatoid arthritis (Fife et al., 2002).

Linkage analysis seldom provides resolution better than 1 centiMorgan (CM) (Hastbacka et al., 1992). Two well separated marker loci may not be distinguished if there is no cross over between them. Most linkage maps are constructed with a limited number of pedigrees and thus have limited resolution. Association analysis has better resolution than linkage analysis, because ideally only closely linked loci show association. Associations decay over time. The decay of association can be modeled in terms of the recombination fraction. The higher the recombination fraction between two loci, the faster the decay of the association. Association reflects recombination events that have occurred over the entire population history. However, association analysis alone may not be adequate to distinguish between an association caused by linkage and an association caused by a biologically irrelevant effect such as population admixture, migration or selection. So association analysis alone may result in high false positive rates.

Another problem with most linkage and association methods is that they study each marker locus separately (single marker analysis). This will cause a waste of information when multilocus marker information is available. Multipoint analysis studies marker loci simultaneously and has been shown to be able to increase the

power of mapping in both linkage and association analysis (Kruglyak et al., 1996; Zeng, 1994). However, little has been done to study linkage and linkage disequilibrium and multiple marker loci simultaneously.

Quantitative trait loci (QTL) refers to loci affecting a continuous trait, such as loci causing alcoholism and hypertension. Many methods have been developed to locate QTL, such as the interval mapping method (Lander and Botstein, 1989) and composite interval mapping (Jansen 1993; Zeng 1993, 1994). Interval mapping is based on linear regression to map single QTL. It tests for the presence of a QTL at many positions between two mapped marker loci. After fitting the model, the goodness of fit of the model is tested using a likelihood ratio test. Composite interval mapping uses a combination of interval mapping with multiple regression to map multiple QTL one by one. Composite interval mapping gives more efficiency and precision then interval mapping because the variance from other QTL would be accounted for by including partial regression coefficients from markers in other regions of the genome. Both interval mapping and composite interval mapping study samples from backcross designs. In some cases, such as in human populations, it is impossible to control the mating. QTL mapping also can use unrelated individuals. Wang and Zeng (2001) developed an EM algorithm to estimate the disequilibrium coefficients between the marker alleles and the disease alleles, the QTL allele frequencies, and QTL model parameters based on phenotypes and genotypes of unrelated individuals. They developed models to study single QTL as well as multiple QTL simultaneously.

In this chapter, we study linkage and linkage disequilibrium simultaneously for

mapping single QTL using unrelated family data in an attempt to increase mapping resolution and reduce false positive results. We develop an EM algorithm that extends Wang and Zeng's EM algorithm (2001) to family members and study allele transmission from parent to offspring so that the recombination fractions can be estimated. In order to speed up the algorithm, we derive closed form solutions to update the QTL parameters at the M-step in the EM algorithm. We perform single marker analysis as well as two marker analysis to increase accuracy of the estimates.

4.2 Single marker analysis using family data

For simplicity, we assume that the quantitative trait is affected by a single QTL. We also assume that our sample consists of unrelated families and that each family has two parents and a given number of offspring. In QTL analysis, we usually assume that phenotypic trait values and marker genotypes of the individuals can be observed. In this section, we will study marker loci one by one. We will estimate QTL allele frequencies, linkage disequilibrium coefficients between the marker alleles and the disease alleles, recombination fraction between the marker locus and the disease locus and QTL model parameters based on the observed phenotypic and genotypic data.

4.2.1 Construct the log likelihood of the data

Consider a natural random mating population which is in Hardy-Weinberg equilibrium. Suppose we have family data: two parents and a given number of offspring in

each family. Assume all the families in the sample are unrelated. Suppose we can observe: unphased marker genotypes of the parents (x_{i1}, x_{i2}) (i=1,2,$\cdots$,n where n represents the total number of families in the sample) and phenotypes of the parents (y_{i1}, y_{i2}). "Unphased genotype" means that we can observe two alleles at each locus, but we do not know which allele is located on which chromosome. "Phased genotype" means that we have knowledge about not only which alleles exist at the locus, but also which chromosomes the alleles are located on at that locus. For each family i, let s_i represent the number of siblings in the family. Assume we can also observe offspring unphased marker genotypes $(x_{oi1}, \cdots, x_{ois_i})$, and offspring trait phenotypes $(y_{oi1}, \cdots, y_{ois_i})$. Let Θ represent all the parameters including QTL allele frequencies, linkage disequilibrium between the marker alleles and the QTL alleles, recombination fraction between the marker locus and the QTL locus and QTL effect parameters. Let z_1, z_2 represent the phased QTL genotypes of the parents and let z_1, z_2 vary over all N_Q possible phased QTL genotypes. Let m_1, m_2 represent the phased marker genotypes of the parents and m_1, m_2 vary over all N_m possible phased marker genotypes. Likewise, let z_o represent the phased offspring QTL genotypes and m_o represent the phased offspring marker genotypes.

Then the likelihood function of the observed data

$Y_{obs} = (x_{i1}, x_{i2}, y_{i1}, y_{i2}, x_{oi1}, \cdots, x_{ois_i}, y_{oi1}, \cdots, y_{ois_i})$, $i = 1, 2, \cdots, n$

can be specified as

$$L(Y_{obs}, \Theta) = \Pi_i^n P(x_{i1}, x_{i2}, y_{i1}, y_{i2}, x_{oi1}, \cdots, x_{ois_i}, y_{oi1}, \cdots, y_{ois_i})$$

$$
\begin{aligned}
&= \Pi_i^n \sum_{z_1}^{N_Q}\sum_{m_1}^{N_m}\sum_{z_2}^{N_Q}\sum_{m_2}^{N_m}\sum_{z_o}^{N_Q}\sum_{m_o}^{N_m} P(z_1, m_1, x_{i1}, y_{i1}, z_2, m_2, x_{i2}, y_{i2}, z_o, m_o, \\
&\quad x_{oi1}, \cdots, x_{ois_i}, y_{oi1}, \cdots, y_{ois_i}) \\
&= \Pi_i^n \sum_{z_1}^{N_Q}\sum_{m_1}^{N_m}\sum_{z_2}^{N_Q}\sum_{m_2}^{N_m}\sum_{z_o}^{N_Q}\sum_{m_o}^{N_m} P(z_1, m_1, x_{i1}, y_{i1}) P(z_2, m_2, x_{i2}, y_{i2}) \\
&\quad \times P(y_{oi1}, \cdots, y_{ois_i} | z_o) \\
&\quad \times P(z_o, m_o, x_{oi1}, \cdots, x_{ois_i} | z_1, m_1, x_{i1}, y_{i1}, z_2, m_2, x_{i2}, y_{i2}) \\
&= \Pi_i^n \sum_{z_1}^{N_Q}\sum_{m_1}^{N_m} P(y_{i1}|z_1) P(z_1, m_1, x_{i1}) \sum_{z_2}^{N_Q}\sum_{m_2}^{N_m} P(y_{i2}|z_2) P(z_2, m_2, x_{i2}) \\
&\quad \times \prod_{j}^{si}\sum_{m_o}^{N_m}\sum_{z_o}^{N_Q} P(y_{oij}|z_o) P(z_o, m_o, x_{oij} | z_1, m_1, x_{i1}, z_2, m_2, x_{i2})
\end{aligned}
$$

where

$$
P(z_1, m_1, x_{i1}) = \begin{cases} P(z_1, m_1) & \text{if } x_{i1} \text{ and } m_1 \text{ are compatible} \\ 0 & \text{otherwise} \end{cases}
$$

$$
P(z_2, m_2, x_{i2}) = \begin{cases} P(z_2, m_2) & \text{if } x_{i2} \text{ and } m_2 \text{ are compatible} \\ 0 & \text{otherwise} \end{cases}
$$

$P(z_o, m_o, x_{oij} | z_1, m_1, x_{i1}, z_2, m_2, x_{i2})$

$$= \begin{cases} P(z_o, m_o | z_1, m_1, z_2, m_2) & \text{if } x_{i1} \text{ and } m_1 \text{ are compatible} \\ & x_{i2} \text{ and } m_2 \text{ are compatible} \\ & x_{oij} \text{ and } m_o \text{ are compatible} \\ 0 & \text{otherwise} \end{cases}$$

If phased and unphased genotypes have same alleles, we call them compatible. For instance, if phased marker genotype is M_1/M_0 and unphased marker genotype is M_0M_1, they are compatible. If phased marker genotype is M_1/M_0 and unphased marker genotype is M_0M_0, they are not compatible.

For simplicity, we assume there are only two alleles (Q_0, Q_1)at the QTL and only two marker alleles (M_0, M_1) at the marker locus. If we assume that z represents the phased QTL genotype $Q_{l_1}/Q_{l'_1}$ and m represents the phased marker genotype $M_{h_1}/M_{h'_1}$. Then P(z,m) becomes $P(Q_{l_1}M_{h_1}/Q_{l'_1}M_{h'_1})$. Under Hardy-Weinberg equilibrium, genotype frequency equals the product of two haplotypes. Suppose the disequilibrium coefficient between allele Q_0 and allele M_0 is D, then the probability of joint phased QTL and marker genotype can be specified as

$$\begin{aligned} P(z,m) &= P(Q_{l_1}M_{h_1})P(Q_{l'_1}M_{h'_1}) \\ &= (p_{l_1}q_{h_1} + (-1)^{l_1+h_1}D)(p_{l'_1}q_{h'_1} + (-1)^{l'_1+h'_1}D) \end{aligned}$$

$P(z_o, m_o|z_1, m_1, z_2, m_2)$ is a function of recombination fraction r between marker locus and disease locus. Table 4.1 shows the probabilities of offspring phased marker and QTL genotype conditional on parental genotypes.

Table 4.1: Offspring genotype frequencies conditional on parental mating types

	offspring genotype							
Parental mating	$\frac{a_1b_1}{a_2b_2}$	$\frac{a'_1b_1}{a_2b_2}$	$\frac{a_1b'_1}{a_2b_2}$	$\frac{a'_1b'_1}{a_2b_2}$	$\frac{a_1b_1}{a'_2b_2}$	$\frac{a_1b_1}{a_2b'_2}$	$\frac{a_1b_1}{a'_2b'_2}$	$\frac{a'_1b_1}{a'_2b_2}$
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a_2b_2}$	1	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a_2b'_2}$	$\frac{1}{2}$	0	0	0	0	$\frac{1}{2}$	0	0
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a'_2b_2}$	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0	0	0
$\frac{a_1b_1}{a'_1b_1} \times \frac{a_2b_2}{a_2b_2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b'_1} \times \frac{a_2b_2}{a_2b_2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a'_2b'_2}$	$\frac{1-r}{2}$	0	0	0	$\frac{r}{2}$	$\frac{r}{2}$	$\frac{1-r}{2}$	0
$\frac{a_1b_1}{a'_1b'_1} \times \frac{a_2b_2}{a_2b_2}$	$\frac{1-r}{2}$	$\frac{r}{2}$	$\frac{r}{2}$	$\frac{1-r}{2}$	0	0	0	0
$\frac{a_1b_1}{a_1b'_1} \times \frac{a_2b_2}{a_2b'_2}$	$\frac{1}{4}$	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$	0	0
$\frac{a_1b_1}{a_1b'_1} \times \frac{a_2b_2}{a'_2b_2}$	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	0	0
$\frac{a_1b_1}{a'_1b_1} \times \frac{a_2b_2}{a_2b'_2}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$	0	0
$\frac{a_1b_1}{a'_1b_1} \times \frac{a_2b_2}{a'_2b_2}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$
$\frac{a_1b_1}{a_1b'_1} \times \frac{a_2b_2}{a'_2b'_2}$	$\frac{1-r}{4}$	0	$\frac{1-r}{4}$	0	$\frac{r}{4}$	$\frac{r}{4}$	$\frac{1-r}{4}$	0
$\frac{a_1b_1}{a'_1b_1} \times \frac{a_2b_2}{a'_2b'_2}$	$\frac{1-r}{4}$	$\frac{1-r}{4}$	0	0	$\frac{r}{4}$	$\frac{r}{4}$	$\frac{1-r}{4}$	$\frac{r}{4}$
$\frac{a_1b_1}{a'_1b'_1} \times \frac{a_2b_2}{a'_2b_2}$	$\frac{1-r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$\frac{1-r}{4}$	$\frac{1-r}{4}$	0	0	$\frac{r}{4}$
$\frac{a_1b_1}{a'_1b'_1} \times \frac{a_2b_2}{a_2b'_2}$	$\frac{1-r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$\frac{1-r}{4}$	0	$\frac{1-r}{4}$	0	0
$\frac{a_1b_1}{a'_1b'_1} \times \frac{a_2b_2}{a'_2b'_2}$	$\frac{(1-r)^2}{4}$	$\frac{r(1-r)}{4}$	$\frac{r(1-r)}{4}$	$\frac{(1-r)^2}{4}$	$\frac{r(1-r)}{4}$	$\frac{r(1-r)}{4}$	$\frac{(1-r)^2}{4}$	$\frac{r^2}{4}$

Table 4.1 (continued)

Parental mating	$\frac{a_1'b_1}{a_2b_2'}$	$\frac{a_1'b_1}{a_2'b_2'}$	$\frac{a_1b_1'}{a_2'b_2}$	$\frac{a_1b_1'}{a_2b_2'}$	$\frac{a_1b_1'}{a_2'b_2'}$	$\frac{a_1'b_1'}{a_2'b_2}$	$\frac{a_1'b_1'}{a_2b_2'}$	$\frac{a_1'b_1'}{a_2'b_2'}$
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a_2b_2}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a_2b_2'}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a_2'b_2}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1'b_1} \times \frac{a_2b_2}{a_2b_2}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1'} \times \frac{a_2b_2}{a_2b_2}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1} \times \frac{a_2b_2}{a_2'b_2'}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1'b_1'} \times \frac{a_2b_2}{a_2b_2}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1'} \times \frac{a_2b_2}{a_2b_2'}$	0	0	0	$\frac{1}{4}$	0	0	0	0
$\frac{a_1b_1}{a_1b_1'} \times \frac{a_2b_2}{a_2'b_2}$	0	0	$\frac{1}{4}$	0	0	0	0	0
$\frac{a_1b_1}{a_1'b_1} \times \frac{a_2b_2}{a_2b_2'}$	$\frac{1}{4}$	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1'b_1} \times \frac{a_2b_2}{a_2'b_2}$	0	0	0	0	0	0	0	0
$\frac{a_1b_1}{a_1b_1'} \times \frac{a_2b_2}{a_2'b_2'}$	0	0	$\frac{r}{4}$	$\frac{r}{4}$	$\frac{1-r}{4}$	0	0	0
$\frac{a_1b_1}{a_1'b_1} \times \frac{a_2b_2}{a_2'b_2'}$	$\frac{r}{4}$	$\frac{1-r}{4}$	0	0	0	0	0	0
$\frac{a_1b_1}{a_1'b_1'} \times \frac{a_2b_2}{a_2'b_2}$	0	0	$\frac{r}{4}$	0	0	$\frac{1-r}{4}$	0	0
$\frac{a_1b_1}{a_1'b_1'} \times \frac{a_2b_2}{a_2b_2'}$	$\frac{r}{4}$	0	0	$\frac{r}{4}$	0	0	$\frac{1-r}{4}$	0
$\frac{a_1b_1}{a_1'b_1'} \times \frac{a_2b_2}{a_2'b_2'}$	$\frac{r^2}{4}$	$\frac{r(1-r)}{4}$	$\frac{r^2}{4}$	$\frac{r^2}{4}$	$\frac{r(1-r)}{4}$	$\frac{r(1-r)}{4}$	$\frac{r(1-r)}{4}$	$\frac{(1-r)^2}{4}$

The phenotypic trait value (P) can usually be expressed in terms of the genotypic value (G), the environmental deviation (E), and the genotype by environment interaction ($G \times E$).

$$P = G + E + G \times E$$

Then, the phenotypic variance can be expressed as

$$V_p = V_G + V_E + V_{G\times E} + 2Cov(G, E) + 2Cov(G, G \times E) + 2Cov(E, G \times E)$$

If we assume that there is no genotype and environment interaction and no covariance between genotype and environment, then the phenotypic trait value can usually be expressed as

$$y_i = \mu + X_{jj'}E + e_i$$

where μ represents the overall population mean of the trait. E is a vector of QTL effect parameters and $E = (a, d)^T$, where a refers to additive effect and d refers to dominance effect. $X_{jj'}$ is a vector of coefficients corresponds to different QTL genotypes. For example, both X_{10} and X_{01} corresponds to QTL genotype Q_1Q_0. We define $X_{00} = (-1, -1/2)$, $X_{01} = X_{10} = (0, 1/2)$, $X_{11} = (1, -1/2)$. e_i stands for the residual error of the model. Usually, we assume that e_i follows a normal distribution and $e_i \sim N(0, \sigma^2)$. Then $P(y_i|z_{jj'})$ follows a normal distribution with mean of $\mu + X_{jj'}E$ and variance of σ^2:

$$P(y_i|z_{jj'}) = \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{(y_i - \mu - X_{jj'}E)^2}{2\sigma^2}\}$$

In other words,

$$P(y_i|z_{jj'}) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{(y_i-\mu-a+d/2)^2}{2\sigma^2}\} & \text{if } z_{jj'} = \frac{Q_1}{Q_1} \\ \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{(y_i-\mu-d/2)^2}{2\sigma^2}\} & \text{if } z_{jj'} = \frac{Q_1}{Q_0} \text{ or } \frac{Q_0}{Q_1} \\ \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{(y_i-\mu+a+d/2)^2}{2\sigma^2}\} & \text{if } z_{jj'} = \frac{Q_0}{Q_0} \end{cases}$$

Define $f_{i1} = P(y_{i1}|z_1)$, $P_{i1} = P(z_1, m_1, x_{i1})$, $f_{i2} = P(y_{i2}|z_2)$, $P_{i2} = P(z_2, m_2, x_{i2})$, $P_{oij} = P(z_o, m_o, x_{oij}|z_1, m_1, x_{i1}, z_2, m_2, x_{i2})$ and $f_{oij} = P(y_{oij}|z_o)$. Then the log-likelihood function of the observed data is given by

$$\begin{aligned} \log L(Y_{obs}, \Theta) &= \sum_i^n \log \sum_{z_1} \sum_{m_1} P(y_{i1}|z_1) P(z_1, m_1, x_{i1}) \sum_{z_2} \sum_{m_2} P(y_{i2}|z_2) P(z_2, m_2, x_{i2}) \\ &\quad \times \prod_j \sum_{m_o} \sum_{z_o} P(y_{oij}|z_o) P(z_o, m_o, x_{oij}|z_1, m_1, x_{i1}, z_2, m_2, x_{i2}) \\ &= \sum_i^n \log \sum_{z_1} \sum_{m_1} f_{i1} P_{i1} \sum_{z_2} \sum_{m_2} f_{i2} P_{i2} \prod_j \sum_{m_o} \sum_{z_o} P_{oij} f_{oij} \end{aligned}$$

4.2.2 EM algorithm

After constructing the log likelihood of the observed data, we can develop a searching method based on the expectation- maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). By using the EM algorithm, we can search for optimal parameter values that give the maximum value of the log likelihood function. The EM algorithm is simple and stable and has been widely used to fit models with incomplete data. To get maximum likelihood estimates of the data, we need to take derivative of the log likelihood function with respect to Θ, where Θ refers to all parameters including QTL allele frequencies, disequilibrium coefficients between marker alleles and disease

alleles, recombination fraction between marker locus and QTL locus and QTL effect parameters. We can show that the derivative is a function of f, P and π_{i1}, π_{i2}, and π_{i3} (see detail in appendix A).

$$\begin{aligned}\frac{\partial L(\Theta|\Theta^{(t)})}{\partial\Theta} &= \sum_{i}^{n}\sum_{l_1l_1'}\sum_{h_1h_1'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)[\frac{\partial \log f_{i1}P_{i1}}{\partial\Theta}+\sum_{l_2l_2'}\sum_{h_2h_2'}\pi^{l_2h_2}_{l_2'h_2',i2}(t)\{\frac{\partial \log f_{i2}P_{i2}}{\partial\Theta}\\ &\quad +\sum_{s}\sum_{jj'}\sum_{kk'}\pi^{jk}_{j'k',is3}(t)\frac{\partial \log P_{ois}f_{ois}}{\partial\Theta}\}]\end{aligned}$$

where

$$\pi^{l_1h_1}_{l_1'h_1',i1}(t)=\frac{f_{i1}P_{i1}\sum_{l_2l_2'}\sum_{h_2h_2'}f_{i2}P_{i2}\prod_s\sum_{jj'}\sum_{kk'}P_{ois}f_{ois}|\Theta=\Theta^{(t)}}{\sum_{l_1l_1'}\sum_{h_1h_1'}f_{i1}P_{i1}\sum_{l_2l_2'}\sum_{h_2h_2'}f_{i2}P_{i2}\prod_s\sum_{jj'}\sum_{kk'}P_{ois}f_{ois}|\Theta=\Theta^{(t)}}$$

$$\pi^{l_2h_2}_{l_2'h_2',i2}(t)=\frac{f_{i2}P_{i2}\prod_s\sum_{jj'}\sum_{kk'}P_{ois}f_{ois}|\Theta=\Theta^{(t)}}{\sum_{l_2l_2'}\sum_{h_2h_2'}f_{i2}P_{i2}\prod_s\sum_{jj'}\sum_{kk'}P_{ois}f_{ois}|\Theta=\Theta^{(t)}}$$

$$\pi^{jk}_{j'k',is3}(t)=\frac{P_{ois}f_{ois}|\Theta=\Theta^{(t)}}{\sum_{jj'}\sum_{kk'}P_{ois}f_{ois}|\Theta=\Theta^{(t)}}$$

Setting the derivative to zero and solving for Θ, we can get maximum likelihood estimates of Θ. On the E-step at the t-th iteration, we need to compute posterior probabilities $\pi^{l_1h_1}_{l_1'h_1',i1}(t)$, $\pi^{l_2h_2}_{l_2'h_2',i2}(t)$ and $\pi^{jk}_{j'k',is3}(t)$ for all $l_1, l_1', h_1, h_1', l_2, l_2', h_2, h_2', j, j', k, k' = 0, 1$ and $i = 1, 2, \cdots, n$ given $\Theta = \Theta^{(t)}$. On M-step at the t-th iteration, we need to maximize the L-function and update Θ as $\widehat{\Theta}^{(t+1)}$. Step E and Step M need to be repeated until convergence is reached. One thing need to be noticed is that the EM procedure could lead to a local maximum. We should assign the parameters with different sets of initial values to search for the global maximum. Different sets of estimators may be obtained with different sets of initial values and the estimates

corresponding to the maximum likelihood value should be chosen. In the following sections, we will derive closed form solutions for $\Theta^{(t+1)}$.

Estimation of QTL allele frequency and LD

Let p_1 represent the QTL allele frequency of Q_1 and D denote the disequilibrium coefficient between allele Q_0 and allele M_0. Taking the derivative of the log-likelihood with respect to p_1, we get (detail in Appendix A)

$$\begin{aligned}\frac{\partial L(\Theta|\Theta^{(t)})}{\partial p_1} &= \sum_{l_1h_1}[\sum_i \sum_{l_1'h_1'}\{\pi^{l_1h_1}_{l_1'h_1',i1}(t)+\pi^{l_1'h_1'}_{l_1h_1,i1}(t)\}]\frac{(-1)^{1+l_1}q_{h_1}}{P_{l_1h_1}} \\ &\quad + \sum_{l_2h_2}[\sum_i \sum_{l_1l_1'}\sum_{h_1h_1'}\sum_{l_2'h_2'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)\{\pi^{l_2h_2}_{l_2'h_2',i2}(t)+\pi^{l_2'h_2'}_{l_2h_2,i2}(t)\}]\frac{(-1)^{1+l_2}q_{h_2}}{P_{l_2h_2}} \\ &= 0\end{aligned}$$

Define

$a^{(t)}_{l_1h_1} = \sum_i \sum_{l_1'h_1'}\{\pi^{l_1h_1}_{l_1'h_1',i1}(t)+\pi^{l_1'h_1'}_{l_1h_1,i1}(t)\}$

$b^{(t)}_{l_2h_2} = \sum_i \sum_{l_1l_1'}\sum_{h_1h_1'}\sum_{l_2'h_2'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)\{\pi^{l_2h_2}_{l_2'h_2',i2}(t)+\pi^{l_2'h_2'}_{l_2h_2,i2}(t)\}$

Then, we can get

$$\sum_{l_1h_1} a^{(t)}_{l_1h_1}\frac{(-1)^{1+l_1}q_{h_1}}{P_{l_1h_1}} + \sum_{l_2h_2} b^{(t)}_{l_2h_2}\frac{(-1)^{1+l_2}q_{h_2}}{P_{l_2h_2}} = 0$$

By index substitution, we can simplify the above expression as

$$\sum_{l_1h_1}(a^{(t)}_{l_1h_1}+b^{(t)}_{l_1h_1})\frac{(-1)^{1+l_1}q_{h_1}}{P_{l_1h_1}} = 0 \qquad (4.1)$$

Likewise, taking derivative of the log-likelihood with respect to D, we have (detail in Appendix A)

$$\frac{\partial L(\Theta|\Theta^{(t)})}{\partial D} = \sum_{l_1 h_1} (a_{l_1 h_1}^{(t)} + b_{l_1 h_1}^{(t)}) \frac{(-1)^{l_1+h_1}}{P_{l_1 h_1}} = 0 \tag{4.2}$$

By solving equation (4.1) and (4.2) together, we can get

$$\begin{cases} \sum_{h_1} [\sum_{l_1} (a_{l_1 h_1}^{(t)} + b_{l_1 h_1}^{(t)}) \frac{(-1)^{l_1}}{P_{l_1 h_1}}] q_{h_1} = 0 \\ \sum_{h_1} [\sum_{l_1} [(a_{l_1 h_1}^{(t)} + b_{l_1 h_1}^{(t)}) \frac{(-1)^{l_1}}{P_{l_1 h_1}}] (-1)^{h_1} = 0 \end{cases}$$

If we define $c_{h_1}(t) = \sum_{l_1} (a_{l_1 h_1}^{(t)} + b_{l_1 h_1}^{(t)}) \frac{(-1)^{l_1}}{P_{l_1 h_1}}$, we will obtain

$$\begin{cases} c_0(t) q_0 + c_1(t) q_1 = 0 \\ c_0(t) - c_1(t) = 0 \end{cases}$$

which leads to

$$c_0(t) = c_1(t) = 0$$

In other words,

$$\sum_{l_1} (a_{l_1 h_1}^{(t)} + b_{l_1 h_1}^{(t)}) \frac{(-1)^{l_1}}{P_{l_1 h_1}} = 0$$

$$\Longrightarrow \frac{a_{0 h_1}^{(t)} + b_{0 h_1}^{(t)}}{P_{0 h_1}} = \frac{a_{1 h_1}^{(t)} + b_{1 h_1}^{(t)}}{P_{1 h_1}}$$

$$\Longrightarrow P_{1h_1} = \frac{a_{1h_1}^{(t)} + b_{1h_1}^{(t)}}{a_{0h_1}^{(t)} + b_{0h_1}^{(t)}} P_{0h_1}$$

$$\Longrightarrow \begin{cases} P_{10} = \frac{a_{10}(t)+b_{10}(t)}{a_{00}(t)+b_{00}(t)} P_{00} \\ \\ P_{11} = \frac{a_{11}(t)+b_{11}(t)}{a_{01}(t)+b_{01}(t)} P_{01} \end{cases}$$

Let q_0 denote the marker allele frequency of M_0 and q_1 denote the marker allele frequency of M_1. q_0 and q_1 can be estimated based on marker information alone, so we treat them as constants. Then,

$$\begin{cases} q_0 = P_{00} + P_{10} = (1 + \frac{a_{10}(t)+b_{10}(t)}{a_{00}(t)+b_{00}(t)}) P_{00} \\ \\ q_1 = P_{01} + P_{11} = (1 + \frac{a_{11}(t)+b_{11}(t)}{a_{01}(t)+b_{01}(t)}) P_{01} \end{cases}$$

$$\begin{aligned} p_1^{(t+1)} &= P_{10}^{(t)} + P_{11}^{(t)} \\ &= \frac{a_{10}^{(t)} + b_{10}^{(t)}}{a_{00}^{(t)} + b_{00}^{(t)}} P_{00} + \frac{a_{11}^{(t)} + b_{11}^{(t)}}{a_{01}^{(t)} + b_{01}^{(t)}} P_{01} \\ &= \frac{a_{10}^{(t)} + b_{10}^{(t)}}{a_{00}^{(t)} + b_{00}^{(t)}} \times \frac{q_0(a_{00}^{(t)} + b_{00}^{(t)})}{(a_{00}^{(t)} + b_{00}^{(t)}) + (a_{10}^{(t)} + b_{10}^{(t)})} \\ &\quad + \frac{a_{11}^{(t)} + b_{11}^{(t)}}{a_{01}^{(t)} + b_{01}^{(t)}} \times \frac{q_1(a_{01}^{(t)} + b_{01}^{(t)})}{(a_{01}^{(t)} + b_{01}^{(t)}) + (a_{11}^{(t)} + b_{11}^{(t)})} \\ &= \frac{q_0(a_{10}^{(t)} + b_{10}^{(t)})}{(a_{00}^{(t)} + b_{00}^{(t)}) + (a_{10}^{(t)} + b_{10}^{(t)})} + \frac{q_1(a_{11}^{(t)} + b_{11}^{(t)})}{(a_{01}^{(t)} + b_{01}^{(t)}) + (a_{11}^{(t)} + b_{11}^{(t)})} \end{aligned}$$

Similarly, we update D as

$$
\begin{aligned}
D^{(t+1)} &= P_{11}^{(t)} - p_1^{(t)} q_1 \\
&= \frac{q_1(a_{11}^{(t)} + b_{11}^{(t)})}{(a_{01}^{(t)} + b_{01}^{(t)}) + (a_{11}^{(t)} + b_{11}^{(t)})} - [\frac{q_0(a_{10}^{(t)} + b_{10}^{(t)})}{(a_{00}^{(t)} + b_{00}^{(t)}) + (a_{10}^{(t)} + b_{10}^{(t)})} \\
&\quad + \frac{q_1(a_{11}^{(t)} + b_{11}^{(t)})}{(a_{01}^{(t)} + b_{01}^{(t)}) + (a_{11}^{(t)} + b_{11}^{(t)})}] q_1 \\
&= [\frac{(a_{11}^{(t)} + b_{11}^{(t)})}{(a_{01}^{(t)} + b_{01}^{(t)}) + (a_{11}^{(t)} + b_{11}^{(t)})} - \frac{(a_{10}^{(t)} + b_{10}^{(t)})}{(a_{00}^{(t)} + b_{00}^{(t)}) + (a_{10}^{(t)} + b_{10}^{(t)})}] q_0 q_1
\end{aligned}
$$

Estimation of recombination fraction

In the log-likelihood, only P_{ois} is a function of recombination fraction, r, between the marker locus and the QTL locus. Taking the derivative with respect to r, we can get

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^t)}{\partial r} &= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi_{l_1' h_1', i1}^{l_1 h_1}(t) [0 + \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi_{h_2 h_2', i2}^{l_2 l_2'}(t) \{0 \\
&\quad + \sum_s \sum_{jj'} \sum_{kk'} \pi_{j'k', is3}^{jk}(t) \frac{\partial \log P_{ois}}{\partial r}\}] \\
&= \sum_i \sum_s \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \sum_{jj'} \sum_{kk'} \pi_{l_1' h_1', i1}^{l_1 h_1}(t) \pi_{l_2' h_2', i2}^{l_2 h_2}(t) \pi_{j'k', is3}^{jk}(t) \\
&\quad \times [-\frac{1}{(1-r)} I(A) + \frac{1}{r} I(B) - \frac{2}{(1-r)} I(C) \\
&\quad + (\frac{1}{r} - \frac{1}{(1-r)}) I(D) + \frac{2}{r} I(E)] \\
&= 0
\end{aligned}
$$

where the I's are a set of indicator variables defined as

$$I(A) = \begin{cases} 1 & \text{if } P(\frac{Q_j M_k}{Q_{j'} M_{k'}} | \frac{Q_{l_1} M_{h_1}}{Q_{l'_1} M_{h'_1}}, \frac{Q_{l_2} M_{h_2}}{Q_{l'_2} M_{h'_2}}) = \frac{1-r}{2} \text{ or } \frac{1-r}{4} \\ 0 & \text{otherwise} \end{cases}$$

$$I(B) = \begin{cases} 1 & \text{if } P(\frac{Q_j M_k}{Q_{j'} M_{k'}} | \frac{Q_{l_1} M_{h_1}}{Q_{l'_1} M_{h'_1}}, \frac{Q_{l_2} M_{h_2}}{Q_{l'_2} M_{h'_2}}) = \frac{r}{2} \text{ or } \frac{r}{4} \\ 0 & \text{otherwise} \end{cases}$$

$$I(C) = \begin{cases} 1 & \text{if } P(\frac{Q_j M_k}{Q_{j'} M_{k'}} | \frac{Q_{l_1} M_{h_1}}{Q_{l'_1} M_{h'_1}}, \frac{Q_{l_2} M_{h_2}}{Q_{l'_2} M_{h'_2}}) = \frac{(1-r)^2}{4} \\ 0 & \text{otherwise} \end{cases}$$

$$I(D) = \begin{cases} 1 & \text{if } P(\frac{Q_j M_k}{Q_{j'} M_{k'}} | \frac{Q_{l_1} M_{h_1}}{Q_{l'_1} M_{h'_1}}, \frac{Q_{l_2} M_{h_2}}{Q_{l'_2} M_{h'_2}}) = \frac{r(1-r)}{4} \\ 0 & \text{otherwise} \end{cases}$$

$$I(E) = \begin{cases} 1 & \text{if } P(\frac{Q_j M_k}{Q_{j'} M_{k'}} | \frac{Q_{l_1} M_{h_1}}{Q_{l'_1} M_{h'_1}}, \frac{Q_{l_2} M_{h_2}}{Q_{l'_2} M_{h'_2}}) = \frac{r^2}{4} \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$\sum_i \sum_s [-rA_{is} + (1-r)B_{is} - 2rC_{is} + (1-2r)D_{is} + 2(1-r)E_{is}] = 0$$

Where $A_{is} = \sum_i \sum_s \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \sum_{jj'} \sum_{kk'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \pi^{jk}_{j'k', is3}(t)$ and the index of $\pi^{l_1 h_1}_{l_1' h_1', i1}(t)$, $\pi^{l_2 h_2}_{l_2' h_2', i2}(t)$, and $\pi^{jk}_{j'k', is3}(t)$ should be in set A. Likewise B_{is} represents the sum of π's whose indexes are in set B, etc. Then, $r^{(t+1)}$ can be expressed in terms of A_{is}, B_{is}, C_{is}, D_{is}, E_{is} as

$$r^{(t+1)} = \frac{\sum_i \sum_s (B_{is} + D_{is} + 2E_{is})}{\sum_i \sum_s (A_{is} + B_{is} + 2C_{is} + 2D_{is} + 2E_{is})}$$

Estimation of QTL effects

In the L-function, only f_{i1},f_{i2}, f_{ois} are functions of QTL effect parameters including the overall population mean of the trait, the additive and dominance effects of the QTL and the variance of the QTL. If we take derivatives of P_{i1}, P_{i2} and P_{ois} in the L-function with respect to the QTL effect parameters, they will equal zero. To get maximum likelihood estimates of all the QTL effect parameters, we need to take derivatives of the log likelihood function with respect to each QTL effect parameter and set the derivatives equal to zero, then solve for the parameters. For simplicity, we assume that the i-family has s_i offspring. The following are closed form solutions for the QTL effect parameters (see detail in Appendix A)

$$\begin{aligned}
\mu^{(t+1)} &= \frac{1}{\sum_i (2 + s_i)} [\sum_i (y_{i1} + y_{i2} + \sum_s y_{ois}) \\
&\quad - \sum_i \sum_{jj'} (c_{jj',i1}(t) + c_{jj',i2}(t)) X_{jj'} E^{(t)} - \sum_i \sum_s \sum_{jj'} c_{jj',is3}(t) X_{jj'} E^{(t)}] \\
a^{(t+1)} &= [\sum_i \{(c_{11,i1}(t) - c_{00,i1}(t))(y_{i1} - \mu^{(t+1)} + \frac{d^{(t)}}{2}) + (c_{11,i2}(t) - c_{00,i2}(t))
\end{aligned}$$

$$
\begin{aligned}
& \times(y_{i2}-\mu^{(t+1)}+\frac{d^{(t)}}{2})+\sum_s(c_{11,i3}(t)-c_{00,i3}(t))(y_{ois}-\mu^{(t+1)}+\frac{d^{(t)}}{2})\}] \\
& /[\sum_i\{c_{11,i1}(t)+c_{00,i1}(t)+c_{11,i2}(t)+c_{00,i2}(t)+\sum_s(c_{11,is3}(t)+c_{00,is3}(t))\}] \\
d^{(t+1)} &= \frac{2}{\sum_i(2+s_i)}\sum_i[(c_{01,i1}(t)+c_{10,i1}(t)-c_{11,i1}(t)-c_{00,i1}(t))(y_{i1}-\mu^{(t+1)}) \\
& +(c_{01,i2}(t)+c_{10,i2}(t)-c_{11,i2}(t)-c_{00,i2}(t))(y_{i2}-\mu^{(t+1)}) \\
& +\sum_s\{(c_{01,is3}(t)+c_{10,is3}(t)-c_{11,is3}(t)-c_{00,is3}(t))(y_{ois}-\mu^{(t+1)})\} \\
& +\{c_{11,i1}(t)-c_{00,i1}(t)+c_{11,i2}(t)-c_{00,i2}(t) \\
& +\sum_s(c_{11,is3}(t)-c_{00,is3}(t))\}a^{(t+1)}] \\
(\sigma^2)^{(t+1)} &= \frac{1}{\sum_i(2+s_i)}[\sum_i\sum_{jj'}\{c_{jj',i1}(t)(y_{i1}-\mu^{(t+1)}-X_{jj'}E^{(t+1)})^2+c_{jj',i2}(t) \\
& \times(y_{i2}-\mu^{(t+1)}-X_{jj'}E^{(t+1)})^2 \\
& +\sum_s c_{jj',is3}(t)(y_{ois}-\mu^{(t+1)}-X_{jj'}E^{(t+1)})^2\}]
\end{aligned}
$$

where

$c_{l_1l_1',i1}(t)=\sum_{h_1h_1'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)$

$c_{l_2l_2',i2}(t)=\sum_{l_1l_1'}\sum_{h_1h_1'}\sum_{h_2h_2'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)\pi^{l_2h_2}_{l_2'h_2',i2}(t)$

$c_{jj',is3}(t)=\sum_{l_1l_1'}\sum_{h_1h_1'}\sum_{l_2l_2'}\sum_{h_2h_2'}\sum_{kk'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)\pi^{l_2h_2}_{l_2'h_2',i2}(t)\pi^{jk}_{j'k',is3}(t).$

and

$\sum_{l_1l_1'}c_{l_1l_1',i1}(t)=\sum_{l_2l_2'}c_{l_2l_2',i2}(t)=1$

$\sum_s\sum_{jj'}c_{jj',is3}(t)=s_i.$

4.3 Two marker analysis using family data

4.3.1 Construct the log likelihood of the data

After performing single marker analysis, we extend our model to study two marker loci simultaneously in an attempt to increase the accuracy of the estimations. As before, suppose there is only one QTL affecting the trait. Assume we still can observe the parental and offspring trait values and marker genotypes. Now the marker genotypes involve alleles at two different loci. Similarly to the single marker model, let N_Q stand for the total number of possible phased QTL genotypes and N_m stand for the total number of possible phased marker genotypes. The log likelihood function for the two marker loci and single QTL model is similar to the single marker and single QTL model. The log likelihood function still can be expressed as

$$\begin{aligned}
\log L(Y_{obs}, \Theta) &= \sum_{i}^{n} \log \sum_{z_1}^{N_Q} \sum_{m_1}^{N_m} P(y_{i1}|z_1) P(z_1, m_1, x_{i1}) \sum_{z_2}^{N_Q} \sum_{m_2}^{N_m} P(y_{i2}|z_2) P(z_2, m_2, x_{i2}) \\
&\quad \times \prod_{j}^{si} \sum_{m_o}^{N_m} \sum_{z_o}^{N_Q} P(y_{oij}|z_o) P(z_o, m_o, x_{oij}|z_1, m_1, x_{i1}, z_2, m_2, x_{i2}) \\
&= \sum_{i}^{n} \log \sum_{z_1}^{N_Q} \sum_{m_1}^{N_m} f_{i1} P_{i1} \sum_{z_2}^{N_Q} \sum_{m_2}^{N_m} f_{i2} P_{i2} \prod_{j}^{si} \sum_{m_o}^{N_m} \sum_{z_o}^{N_Q} P_{oij} f_{oij}
\end{aligned}$$

Here, y_{i1}, y_{i2}, z_1 and z_2 still represent the trait values and phased QTL genotypes of the two parents; y_{oij} and z_o denote trait values and phased QTL genotypes of the j-th child in the i-th family; x_{i1}, x_{i2}, m_1 and m_2 stands for observed unphased two marker genotypes and phased two marker genotypes of the two parents; x_{oij}

and m_o stand for observed unphased two marker genotypes and phased two marker genotypes of the j-th child in the i-th family. If we ignore the genotype by environment interaction and assume there is no covariance between genotype and environment as in the single marker model, the probability of trait values conditional on the phased QTL genotypes still follow a normal distribution with mean of $\mu + X_{jj'}E$ and variance of σ^2. The probability of an unphased genotype still equals the probability of the phased genotype if the observed unphased marker genotype and the phased marker genotype are compatible and it equals 0 otherwise. The probability of joint phased marker and QTL genotypes still equals the product of two haplotype frequencies. Now the haplotypes involve alleles at three different loci. Assume there are only two alleles (Q_0, Q_1) at the QTL locus, two alleles (M_0, M_1) at the first marker locus M, and two alleles (N_0, N_1) at the second marker locus N. The haplotype frequencies can be specified as functions of allele frequencies and three linkage disequilibrium coefficients including the disequilibrium coefficient between allele M_0 and allele Q_0, D_{01}; the disequilibrium coefficient between allele N_0 and allele Q_0, D_{02}, and the disequilibrium coefficient between allele M_0, allele Q_0 and allele N_0, D_{012}. Assume $P(Q_0) = p_0$, $P(Q_1) = p_1$, $P(M_0) = q_0$, $P(M_1) = q_1$, and $P(N_0) = w_0$, $P(N_1) = w_1$. Then, the genotype frequencies become

$$\begin{aligned} P(\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l_1'}Q_{j_1'}N_{h_1'}}) &= P(M_{l_1}Q_{j_1}N_{h_1})P(M_{l_1'}Q_{j_1'}N_{h_1'}) \\ &= [(-1)^{j_1+l_1+h_1}D_{012} + (-1)^{j_1+l_1}w_{h_1}D_{01} \\ &\quad +(-1)^{j_1+h_1}q_{l_1}D_{02} + p_{j_1}P_{l_1h_1}] \end{aligned}$$

$$\times[(-1)^{j_1'+l_1'+h_1'}D_{012}+(-1)^{j_1'+l_1'}w_{h_1'}D_{01}$$

$$+(-1)^{j_1'+h_1'}q_{l_1'}D_{02}+p_{j_1'}P_{l_1'h_1'}]$$

where $l_1, l_1', j_1, j_1', h_1, h_1' = 0, 1$.

The offspring joint phased QTL and marker genotypes conditional on parental phased genotypes can be expressed in terms of the recombination fractions between QTL locus and marker loci. Let r_1 denote the recombination fraction between marker locus M and the QTL locus, r_2 represent the recombination fraction between QTL locus and marker locus N and r_3 stands for the recombination fraction between marker loci M and N. Since r_3 can be estimated based on marker information alone, we treat r_3 as a known value. Assume there is no double crossover event and $r_1 + r_2 = r_3$. Define $\alpha = r_1/r_3$, then $1-\alpha = r_2/r_3$. Thus r_1, r_2 can be expressed in terms of r_3 and α: $r_1 = \alpha r_3$ and $r_2 = (1-\alpha)r_3$. Then probabilities the offspring joint QTL and marker genotypes conditional on parental genotypes can be expressed in terms of α and r_3, where only α is an unknown parameter. Suppose the offspring phased genotype is $M_{l_o}Q_{j_o}N_{h_o}/M_{l_o'}Q_{j_o'}N_{h_o'}$ and the first chromosome is the maternal chromosome, the second chromosome is the paternal chromosome. Assume the mother phased genotype is $M_{l_1}Q_{j_1}N_{h_1}/M_{l_1'}Q_{j_1'}N_{h_1'}$ and the father phased genotype is $M_{l_2}Q_{j_2}N_{h_2}/M_{l_2'}Q_{j_2'}N_{h_2'}$, then

$$P(\frac{M_{l_o}Q_{j_o}N_{h_o}}{M_{l_o'}Q_{j_o'}N_{h_o'}}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l_1'}Q_{j_1'}N_{h_1'}},\frac{M_{l_2}Q_{j_2}N_{h_2}}{M_{l_2'}Q_{j_2'}N_{h_2'}}) \quad = \quad P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l_1'}Q_{j_1'}N_{h_1'}})$$

$$\times P(M_{l'_o}Q_{j'_o}N_{h'_o}|\frac{M_{l_2}Q_{j_2}N_{h_2}}{M_{l'_2}Q_{j'_2}N_{h'_2}})$$

When $l_1 = l'_1$, $j_1 = j'_1$ and $h_1 = h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l'_1}Q_{j'_1}N_{h'_1}}) = \begin{cases} 1 & \text{if } l_o j_o h_o = l_1 j_1 h_1 \\ 0 & \text{otherwise} \end{cases}$$

When $l_1 = l'_1$, $j_1 = j'_1$ and $h_1 \neq h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l'_1}Q_{j'_1}N_{h'_1}}) = \begin{cases} \frac{1}{2} & \text{if } l_o j_o h_o = l_1 j_1 h_1 \text{ or } l_1 j_1 h'_1 \\ 0 & \text{otherwise} \end{cases}$$

When $l_1 = l'_1$, $j_1 \neq j'_1$ and $h_1 = h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l'_1}Q_{j'_1}N_{h'_1}}) = \begin{cases} \frac{1}{2} & \text{if } l_o j_o h_o = l_1 j_1 h_1 \text{ or } l_1 j'_1 h_1 \\ 0 & \text{otherwise} \end{cases}$$

When $l_1 \neq l'_1$, $j_1 = j'_1$ and $h_1 = h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l'_1}Q_{j'_1}N_{h'_1}}) = \begin{cases} \frac{1}{2} & \text{if } l_o j_o h_o = l_1 j_1 h_1 \text{ or } l'_1 j_1 h_1 \\ 0 & \text{otherwise} \end{cases}$$

When $l_1 = l'_1$, $j_1 \neq j'_1$ and $h_1 \neq h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l'_1}Q_{j'_1}N_{h'_1}}) = \begin{cases} \frac{1-(1-\alpha)r_3}{2} & \text{if } l_oj_oh_o = l_1j_1h_1 \text{ or } l_1j'_1h'_1 \\ \frac{(1-\alpha)r_3}{2} & \text{if } l_oj_oh_o = l_1j'_1h_1 \text{ or } l_1j_1h'_1 \\ 0 & \text{otherwise} \end{cases}$$

When $l_1 \neq l'_1$, $j_1 = j'_1$ and $h_1 \neq h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l'_1}Q_{j'_1}N_{h'_1}}) = \begin{cases} \frac{1-r_3}{2} & \text{if } l_oj_oh_o = l_1j_1h_1 \text{ or } l'_1j_1h'_1 \\ \frac{r_3}{2} & \text{if } l_oj_oh_o = l'_1j_1h_1 \text{ or } l_1j_1h'_1 \\ 0 & \text{otherwise} \end{cases}$$

When $l_1 \neq l'_1$, $j_1 \neq j'_1$ and $h_1 = h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l'_1}Q_{j'_1}N_{h'_1}}) = \begin{cases} \frac{1-\alpha r_3}{2} & \text{if } l_oj_oh_o = l_1j_1h_1 \text{ or } l'_1j'_1h_1 \\ \frac{\alpha r_3}{2} & \text{if } l_oj_oh_o = l'_1j_1h_1 \text{ or } l_1j'_1h_1 \\ 0 & \text{otherwise} \end{cases}$$

When $l_1 \neq l'_1$, $j_1 \neq j'_1$ and $h_1 \neq h'_1$

$$P(M_{l_o}Q_{j_o}N_{h_o}|\frac{M_{l_1}Q_{j_1}N_{h_1}}{M_{l_1'}Q_{j_1'}N_{h_1'}}) = \begin{cases} \frac{(1-\alpha r_3)(1-(1-\alpha)r_3)}{2} & \text{if } l_o j_o h_o = l_1 j_1 h_1 \text{ or } l_1' j_1' h_1' \\ \frac{\alpha r_3(1-(1-\alpha)r_3)}{2} & \text{if } l_o j_o h_o = l_1' j_1 h_1 \text{ or } l_1 j_1' h_1' \\ \frac{\alpha(1-\alpha)r_3^2}{2} & \text{if } l_o j_o h_o = l_1 j_1' h_1 \text{ or } l_1' j_1 h_1' \\ \frac{(1-\alpha)(1-\alpha r_3)r_3}{2} & \text{if } l_o j_o h_o = l_1 j_1 h_1' \text{ or } l_1' j_1' h_1 \\ 0 & \text{otherwise} \end{cases}$$

Similarly, $P(M_{l_o'}Q_{j_o'}N_{h_o'}|M_{l_2}Q_{j_2}N_{h_2}/M_{l_2'}Q_{j_2'}N_{h_2'})$ is also a function of r_3 and α. Thus the probabilities of offspring phased genotype conditional on parental phased genotype are functions of r_3 and α.

4.3.2 EM algorithm

After constructing the log likelihood of the data, we can estimate the parameters using an EM algorithm as before. We want to estimate QTL allele frequencies, the three disequilibrium coefficients and the QTL model parameters. So Θ includes p_1, p_2, D_{01}, D_{02}, D_{012}, μ, a, d and σ^2. Taking the derivative of the log-likelihood with respect to Θ, we can get a similar expression to that in the single marker model as follows

$$\frac{\partial L(\Theta|\Theta^{(t)})}{\partial \Theta} = \sum_{i}^{n}\sum_{j_1 j_1'}\sum_{l_1 l_1'}\sum_{h_1 h_1'} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t)[\frac{\partial \log f_{i1} P_{i1}}{\partial \Theta} + \sum_{j_2 j_2'}\sum_{l_2 l_2'}\sum_{h_2 h_2'} \pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t)$$

$$\times\{\frac{\partial \log f_{i2}P_{i2}}{\partial\Theta}+\sum_s\sum_{j_oj'_o}\sum_{l_ol'_o}\sum_{h_oh'_o}\pi^{j_ol_oh_o}_{j'_ol'_oh'_o,is3}(t)\frac{\partial \log P_{ois}f_{ois}}{\partial\Theta}\}]$$

where

$$\pi^{j_1l_1h_1}_{j'_1l'_1h'_1,i1}(t)$$
$$=\frac{f_{i1}P_{i1}\sum_{j_2j'_2}\sum_{l_2l'_2}\sum_{h_2h'_2}f_{i2}P_{i2}\prod_s\sum_{j_oj'_o}\sum_{l_ol'_o}\sum_{h_oh'_o}P_{ois}f_{ois}|_{\Theta=\Theta^{(t)}}}{\sum_{j_1j'_1}\sum_{l_1l'_1}\sum_{h_1h'_1}f_{i1}P_{i1}\sum_{j_2j'_2}\sum_{l_2l'_2}\sum_{h_2h'_2}f_{i2}P_{i2}\prod_s\sum_{j_oj'_o}\sum_{l_ol'_o}\sum_{h_oh'_o}P_{ois}f_{ois}|_{\Theta=\Theta^{(t)}}}$$

$$\pi^{j_2l_2h_2}_{j'_2l'_2h'_2,i2}(t)=\frac{f_{i2}P_{i2}\prod_s\sum_{j_oj'_o}\sum_{l_ol'_o}\sum_{h_oh'_o}P_{ois}f_{ois}|_{\Theta=\Theta^{(t)}}}{\sum_{j_2j'_2}\sum_{l_2l'_2}\sum_{h_2h'_2}f_{i2}P_{i2}\prod_s\sum_{j_oj'_o}\sum_{l_ol'_o}\sum_{h_oh'_o}P_{ois}f_{ois}|_{\Theta=\Theta^{(t)}}}$$

$$\pi^{j_ol_oh_o}_{j'_ol'_oh'_o,is3}(t)=\frac{P_{ois}f_{ois}|_{\Theta=\Theta^{(t)}}}{\sum_{j_oj'_o}\sum_{l_ol'_o}\sum_{h_oh'_o}P_{ois}f_{ois}|_{\Theta=\Theta^{(t)}}}$$

Here, i indexes the family number. s is the offspring number in the i-th family. j_1 and j'_1 represent the phased QTL genotype of the first parent. l_1 and l'_1 represent the phased marker genotype at the first marker locus of the first parent. h_1 and h'_1 are the phased marker genotype at the second marker locus of the first parent. For example, $l_1 = 1$, $l'_1 = 0$, $j_1 = 0$, $j'_1 = 1$, $h_1 = 0$ and $h'_1 = 1$ stand for $M_1Q_0N_0/M_0Q_1N_1$. Similarly, j_2 and j'_2 stand for phased QTL genotypes of the second parent. l_2, l'_2 and h_2 and h'_2 represent the phased marker genotypes of the second parent. j_o, j'_o, l_o, l'_o, h_o, h'_o stand for the phased QTL and marker genotypes of the children.

In the log-likelihood function, only P_{i1} and P_{i2} are functions of QTL allele frequency, p_1 and the three disequilibrium coefficients D_{01}, D_{02} and D_{012}. Taking derivatives with respect to p_1, D_{01}, D_{02} and D_{012} and solving for the parameters, we can get closed form solutions for the four parameters as follows (details in appendix B)

$$
\begin{aligned}
p_j^{(t+1)} &= \sum_{l_1 h_1} P^{(t)}_{Q_j M_{l_1} N_{h_1}} \quad \text{for j=0, 1} \\
D_{01}^{(t+1)} &= P^{(t)}_{Q_0 M_0} - p_0^{(t+1)} q_0 \\
D_{02}^{(t+1)} &= P^{(t)}_{Q_0 N_0} - p_0^{(t+1)} w_0 \\
D_{012}^{(t+1)} &= P^{(t)}_{Q_0 M_0 N_0} - w_0 D_{01}^{(t+1)} - q_0 D_{02}^{(t+1)} - p_0^{(t+1)} P_{M_0 N_0}
\end{aligned}
$$

where

$$
\begin{aligned}
P^{(t)}_{Q_1 M_{l_1} N_{h_1}} &= \frac{a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)}{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t) + a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)} P_{l_1 h_1} \\
P^{(t)}_{Q_0 M_{l_1} N_{h_1}} &= \frac{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t)}{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t) + a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)} P_{l_1 h_1} \\
P^{(t)}_{M_{l_1} Q_0} &= P^{(t)}_{Q_0 M_{l_1} N_0} + P^{(t)}_{Q_0 M_{l_1} N_1} \\
P^{(t)}_{Q_0 N_{h_1}} &= P^{(t)}_{Q_0 M_0 N_{h_1}} + P^{(t)}_{Q_0 M_1 N_{h_1}}
\end{aligned}
$$

and

$$
\begin{aligned}
a_{j_1 l_1 h_1}(t) &= \sum_i \sum_{j_1' l_1' h_1'} (\pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) + \pi^{j_1' l_1' h_1'}_{j_1 l_1 h_1, i1}(t)) \\
b_{j_2 l_2 h_2}(t) &= \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{j_2' l_2' h_2'} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) (\pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) + \pi^{j_2' l_2' h_2'}_{j_2 l_2 h_2, i2}(t)) \\
c_{l_1 h_1}(t) &= \sum_{j_1} \frac{(-1)^{j_1}}{P_{j_1 l_1 h_1}} (a_{j_1 l_1 h_1}(t) + b_{j_1 l_1 h_1}(t))
\end{aligned}
$$

We need to estimate the haplotype frequency $P(M_{l_1} N_{h_1})$ before we can implement the EM algorithm. The marker haplotype frequency can be estimated based on

marker information alone as described in Wang and Zeng (2001). We could estimate the marker haplotype frequencies using EM algorithm. We could use parental genotype information to estimate marker haplotype frequencies. The log-likelihood can be derived as follows

$$L_M(\Theta_m|\Theta_m^{(t)}) = \sum_i \sum_j \sum_{l_1 l_1'} \sum_{h_1 h_1'} (logP_{l_1 h_1} + logP_{l_1' h_1'}) \pi_{l_1' h_1', ij}^{l_1 h_1}(t)$$

where i sums form 1 to n and j sums form 1 to 2 because there are two parents in each family. $\pi_{l_1' h_1', ij}^{l_1 h_1}(t)$ is the posterior probability of the phased marker genotype $M_{l_1} N_{h_1} / M_{l_1'} N_{h_1'}$ conditional on the unphased marker genotype $x_i = M_{l_1} M_{l_1'} N_{h_1} N_{h_1'}$ of the j-th parent in the i-th family.

$$P_{l_1 h_1}^{(t+1)} = \frac{\Pi_{l_1 h_1}^{(t)}}{\sum_{l_1 h_1} \Pi_{l_1 h_1}^{(t)}} \text{ for } l_1,h_1=0,1$$

where $\Pi_{l_1 h_1}^{(t)} = \sum_i \sum_j \sum_{l_1' h_1'} [\pi_{l_1' h_1', ij}^{l_1 h_1}(t) + \pi_{l_1 h_1, ij}^{l_1' h_1'}(t)]$.

To get maximum likelihood estimates of the QTL model parameters, we need to take derivatives of the log-likelihood with respect to each QTL model parameter. Similar to the one marker case, only f_{i1}, f_{i2} and f_{oij}, $i = 1, 2, ..., n, j = 1, \cdots, s_i$, terms involve genetic model parameters. The estimation of the QTL effect parameters can be done in the same way as the single marker case. For simplicity, we assume that there are s_i children in the ith-family.

$$
\begin{aligned}
\mu^{(t+1)} &= \frac{1}{\sum_i (2+s_i)}[\sum_i (y_{i1}+y_{i2}+\sum_s y_{ois}) \\
&\quad -\sum_i \sum_{jj'} (c_{jj',i1}(t)+c_{jj',i2}(t))X_{jj'}E^{(t)} - \sum_i \sum_s \sum_{jj'} c_{jj',is3}(t)X_{jj'}E^{(t)}] \\
a^{(t+1)} &= [\sum_i \{(c_{11,i1}(t)-c_{00,i1}(t))(y_{i1}-\mu^{(t+1)}+\frac{d^{(t)}}{2}) + (c_{11,i2}(t)-c_{00,i2}(t)) \\
&\quad \times (y_{i2}-\mu^{(t+1)}+\frac{d^{(t)}}{2}) + \sum_s (c_{11,is3}(t)-c_{00,is3}(t))(y_{ois}-\mu^{(t+1)}+\frac{d^{(t)}}{2})\}] \\
&\quad /[\sum_i \{c_{11,i1}(t)+c_{00,i1}(t)+c_{11,i2}(t)+c_{00,i2}(t) \\
&\quad + \sum_s (c_{11,is3}(t)+c_{00,is3}(t))\}] \\
d^{(t+1)} &= \frac{2}{\sum_i (2+s_i)} \sum_i [(c_{01,i1}(t)+c_{10,i1}(t)-c_{11,i1}(t)-c_{00,i1}(t))(y_{i1}-\mu^{(t+1)}) \\
&\quad + (c_{01,i2}(t)+c_{10,i2}(t)-c_{11,i2}(t)-c_{00,i2}(t))(y_{i2}-\mu^{(t+1)}) \\
&\quad + \sum_s \{(c_{01,is3}(t)+c_{10,is3}(t)-c_{11,is3}(t)-c_{00,is3}(t))(y_{ois}-\mu^{(t+1)})\} \\
&\quad + \{c_{11,i1}(t)-c_{00,i1}(t)+c_{11,i2}(t)-c_{00,i2}(t) \\
&\quad + \sum_s (c_{11,is3}(t)-c_{00,is3}(t))\}a^{(t+1)}] \\
(\sigma^2)^{(t+1)} &= \frac{1}{\sum_i (2+s_i)}[\sum_i \sum_{jj'} \{c_{jj',i1}(t)(y_{i1}-\mu^{(t+1)}-X_{jj'}E^{(t+1)})^2 \\
&\quad + c_{jj',i2}(t)(y_{i2}-\mu^{(t+1)}-X_{jj'}E^{(t+1)})^2 \\
&\quad + \sum_s c_{jj',is3}(t)(y_{ois}-\mu^{(t+1)}-X_{jj'}E^{(t+1)})^2\}]
\end{aligned}
$$

where

$$
\begin{aligned}
c_{j_1 j_1',i1}(t) &= \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1',i1}(t) \\
c_{j_2 j_2',i2}(t) &= \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1',i1}(t) \pi^{j_2 l_2 h_2}_{j_2' l_2' h_2',i2}(t)
\end{aligned}
$$

$$c_{j_o j'_o, is3}(t) = \sum_{j_1 j'_1} \sum_{l_1 l'_1} \sum_{h_1 h'_1} \sum_{j_2 j'_2} \sum_{l_2 l'_2} \sum_{h_2 h'_2} \sum_{l_o l'_o} \sum_{h_o h'_o} \pi^{j_1 l_1 h_1}_{j'_1 l'_1 h'_1, i1}(t) \pi^{j_2 l_2 h_2}_{j'_2 l'_2 h'_2, i2}(t) \pi^{j_o l_o h_o}_{j'_o l'_o h'_o, is3}(t)$$

and

$\sum_{j_1 j'_1} c_{j_1 j'_1, i1}(t) = \sum_{j_2 j'_2} c_{l_2 l'_2, i2}(t) = 1$

$\sum_s \sum_{j_o j'_o} c_{j_o j'_o, is3}(t) = s_i.$

4.3.3 Estimation of recombination fraction

Unlike the single marker model, we estimate the recombination fractions between marker loci and QTL locus based on an assumed QTL locus. As we mentioned before, f_{ois} can be expressed in terms of α and r_3, where r_3 can be estimated from marker data. We can assume that the QTL locates at a known point between locus M and locus N. In other words, we can assign a value to α (α should be between 0 and 1). Given the value of α, we can perform the above EM algorithm at that known point and estimate QTL allele frequencies, the three disequilibrium coefficients and QTL effect parameters. After estimating the QTL parameters, we can substitute the parameters to the log likelihood function and compute the log likelihood of the data. Repeating the above process at each point between loci M and N (assign different values between 0 to 1 to α), a set of different log likelihood values will be obtained and each value corresponds to one point between locus M and locus N. The α value for the point with maximum log likelihood value is the maximum likelihood estimate of α. That point is the estimated QTL locus.

4.4 Results

To verify our algorithm, we performed a set of simulations for single marker analysis. We considered a trait affected by one QTL with two alleles. We considered a single two-allele marker locus in linkage and linkage disequilibrium with the QTL locus. We simulated the parental and offspring genotypes and phenotypes according to a set of known parameters. For QTL and marker genotypes, we simulated haplotypes first and put two haplotypes together to form genotypes. The haplotypes were simulated according to the marker allele frequencies, QTL allele frequencies and linkage disequilibrium between marker and QTL loci. The phenotypes were simulated according to the QTL effect parameters, μ, a, d and σ^2. a and d are the additive and dominance effects of the QTL. σ^2 represented the phenotypic variance caused by environment. Trait values were generated as follows:

$$y_i = \begin{cases} \mu + a - \frac{d}{2} + e_i & \text{for QTL genotype } Q_1Q_1 \\ \mu + \frac{d}{2} + e_i & \text{for QTL genotype } Q_1Q_0 \\ \mu - a - \frac{d}{2} + e_i & \text{for QTL genotype } Q_0Q_0 \end{cases}$$

where $e_i \sim N(0, \sigma^2)$. Then we used the simulated data sets to estimate the QTL allele frequencies, linkage disequilibrium and recombination fraction between marker and disease loci, and the genetic model parameters according to the algorithm we developed above. Each sample consisted of 500 families with 1 or 2 offspring in each family. 50 replicate samples were simulated for each set of parameter estimations.

Table 4.2 and Table 4.3 list the means and standard deviations of the parameter estimations for each selected parameter set. h^2 is the broad sense heritability: $h^2 = (V_A + V_D)/V_P$. V_A is the additive genetic variance; V_D is the dominant variance and V_P is the overall phenotypic variance.

$$\begin{aligned} V_A &= 2P_{Q_1}P_{Q_0}[a' + d'(P_{Q_0} - P_{Q_1})]^2 \\ V_D &= (2P_{Q_1}P_{Q_0}d')^2 \\ V_P &= V_A + V_D + \sigma^2 \end{aligned}$$

where $a' = (Q_1Q_1 - Q_0Q_0)/2$ and $d' = Q_1Q_0 - (Q_0Q_0 + Q_1Q_1)/2$. For our case, $a' = a$ and $d' = d$.

Table 4.2: Mean and standard error of the MLE of the genetic parameters from 50 replicate samples. Each sample is consisted of 500 families with 1 offspring in each family

Case		$P(Q_1)$	$P(m_0)$	D^d	r^e	μ	a	d	σ^2	h^2
	parameter[a]	0.5	0.5	0.2	0.1	0.5	0.3	0.1	0.0064	0.8813
1	mean[b]	0.5013	0.4986	0.2006	0.1026	0.5002	0.2998	0.1007	0.0064	0.8807
	S.E.[c]	0.0106	0.0100	0.0040	0.0217	0.0025	0.0033	0.0049	0.0003	0.0473
	parameter	0.5	0.5	0.1	0.15	0.5	0.3	0.1	0.02	0.7037
2	mean	0.4983	0.5003	0.1016	0.1548	0.5024	0.2961	0.1053	0.0205	0.6959
	S.E.	0.0145	0.0112	0.0084	0.0592	0.0062	0.0078	0.0107	0.0012	0.0394
	parameter	0.5	0.5	0.1	0.15	0.5	0.3	0.1	0.08	0.3725
3	mean	0.4179	0.5017	0.1085	0.1463	0.5438	0.2624	0.1456	0.0835	0.3514
	S.E.	0.0336	0.0116	0.0135	0.0740	0.0206	0.0241	0.0301	0.0052	0.0388
	parameter	0.9	0.5	0.05	0.02	10.0	1.2	0.5	0.02	0.8604
4	mean	0.9015	0.4982	0.04937	0.0238	9.9856	1.2242	0.5197	0.0200	0.8718
	S.E.	0.0054	0.0140	0.0027	0.0186	0.0120	0.0225	0.0241	0.0007	0.1429
	parameter	0.5	0.3	0.1	0.08	10.0	1.2	0.5	0.5	0.6101
5	mean	0.5208	0.3021	0.0991	0.0756	9.9414	1.2135	0.4492	0.5029	0.6006
	S.E.	0.0523	0.0120	0.0074	0.0612	0.1342	0.0408	0.1717	0.0364	0.0405
	parameter	0.7	0.5	0.1	0.05	10.0	1.2	0.5	0.5	0.4814
6	mean	0.7240	0.5002	0.0947	0.0390	9.927	1.241	0.4614	0.4981	0.4791
	S.E.	0.0265	0.0105	0.0110	0.0434	0.0779	0.0522	0.0834	0.0336	0.0448

a parameter value.
b mean estimate value.
c standard deviation of the parameter estimations.
d Linkage disequilibrium between marker and disease loci.
e Recombination fraction between marker and disease loci.

Our simulation results showed that our algorithm can give consistent estimates of all the parameters considered. Better estimates were obtained for data sets with higher heritability than for data sets with lower heritability as expected. This is because higher heritability implied that higher proportion of the trait value was contributed by the genetic effects and lower proportion of the trait value was contributed by the random environmental effects. When heritability is high, the estimates for families with one offspring is almost as good as the estimates for families with two offspring. When heritability is low, data sets with two offspring families generated slightly better estimates than data sets with one offspring families.

Table 4.3: Mean and standard error of the MLE of the genetic parameters from 50 replicate samples. Each sample is consisted of 500 families with 2 offspring in each family

Case		$P(Q_1)$	$P(m_0)$	D^d	r^e	μ	a	d	σ^2	h^2
	parameter[a]	0.5	0.5	0.2	0.1	0.5	0.3	0.1	0.0064	0.8813
1	mean[b]	0.4997	0.5005	0.1996	0.1007	0.5001	0.2995	0.1007	0.0064	0.8844
	S.E.[c]	0.0120	0.0108	0.0043	0.0137	0.0019	0.0025	0.0043	0.0003	0.0211
	parameter	0.5	0.5	0.1	0.15	0.5	0.3	0.1	0.02	0.7037
2	mean	0.4991	0.5006	0.0988	0.1368	0.5020	0.2973	0.1053	0.0204	0.6950
	S.E.	0.0121	0.0122	0.0078	0.0328	0.0052	0.0066	0.0083	0.0011	0.0278
	parameter	0.5	0.5	0.1	0.15	0.5	0.3	0.1	0.08	0.3725
3	mean	0.4437	0.5001	0.1007	0.1501	0.5320	0.2775	0.1269	0.0806	0.3612
	S.E.	0.0339	0.0109	0.0111	0.0619	0.0161	0.0199	0.0288	0.0044	0.0271
	parameter	0.9	0.5	0.05	0.02	10.0	1.2	0.5	0.02	0.8604
4	mean	0.9002	0.4997	0.0499	0.0189	9.9872	1.2236	0.5216	0.0201	0.8730
	S.E.	0.0069	0.0117	0.0032	0.0107	0.0089	0.0152	0.0145	0.0006	0.0924
	parameter	0.5	0.3	0.1	0.08	10.0	1.2	0.5	0.5	0.6101
5	mean	0.5181	0.3001	0.0988	0.0769	9.9586	1.2033	0.4547	0.5044	0.6029
	S.E.	0.0526	0.0108	0.0076	0.0388	0.1273	0.0361	0.1608	0.0334	0.0387
	parameter	0.7	0.5	0.1	0.05	10.0	1.2	0.5	0.5	0.4814
6	mean	0.7171	0.4983	0.0977	0.0405	9.9579	1.2206	0.4964	0.5026	0.4659
	S.E.	0.0189	0.0116	0.0077	0.0374	0.0524	0.0448	0.0537	0.0190	0.0334

a parameter value.
b mean estimate value.
c standard deviation of the parameter estimations.
d Linkage disequilibrium between marker and disease loci.
e Recombination fraction between marker and disease loci.

4.5 Discussion

In this chapter, we develop EM algorithms to estimate a set of QTL parameters including QTL allele frequencies, recombination fractions between the QTL locus and the marker loci, linkage disequilibrium coefficients between the marker alleles and the QTL alleles and QTL model parameters. We extend Wang and Zeng (2001) by considering family data instead of unrelated individual data. By doing so, we can study the allele transmission from one generation to the next generation and estimate the recombination fraction between the disease locus and the marker loci. The families we studied consist of two parents and a given number of children. These two generation family data are relatively easier to obtain than extended pedigree

data. To verify our EM algorithm, we carry out simulation studies for single marker model. Our simulation results show that our algorithm can give consistent estimates for all the parameters considered.

As we mentioned before, linkage analysis has limited resolution and association analysis is more appropriate for fine scale mapping. By studying linkage and linkage disequilibrium simultaneously, we can increase mapping resolution. If we perform the above algorithm, we might find multiple marker loci closely linked to the QTL locus, but only the markers that are strongly associated with the QTL physically locate close to the QTL locus. This is because that the association between two far away loci will be broken down by generations of recombination events. Studying linkage and association at the same time also enables us to reduce false positive rates. Association may be caused by non genetic factors, such as population stratification, migration and selection. If a marker allele is in strong linkage disequilibrium with the QTL allele, but there is no evidence of linkage between the marker locus and the disease locus, then it suggests that the association between the marker allele and the disease allele is caused by things other than linkage. This is probably a false positive result.

After performing single marker analysis, we extend our model to study two marker loci simultaneously. When multiple marker information exists, studying marker loci one by one may cause loss of information. Previous studies have shown that studying multiple markers simultaneously for the fine mapping of QTL can increase the mapping power and improve the accuracy of the gene location (Meuwissen and Goddard,

2000; Zeng, 1994). We believe that the two marker model will give better estimation than a single marker model.

List of References

[1] Abel L, Alcais A and Mallet A. Comparison of four sib-pair linkage methods for analyzing sibships with more than two affecteds: interest of the binomial maximum likelihood approach. *Genet Epidemiol* 15:371-390, 1998.

[2] Bickeböller H, Clerget-Darpoux F. Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol* 12:865-870, 1995.

[3] Bishop DT, Williamson JA. The power of identity-by-state methods for linkage analysis. *Am J Hum Genet* 46:254-265, 1990.

[4] Blackwelder WC and Elston RC. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* 2:85–98, 1985.

[5] Cannings C. On the probabilities of identity states in permutable populations. *Am J Hum Genet* 62:698-702, 1998.

[6] Clayton D, Jones H. Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 65:1161-1169, 1999.

[7] Clerget-Darpoux F, Babron MC and Bonaiti-Pellie C. Assessing the effect of multiple linkage tests in complex disease. *Genet Epidemiol* 7:245-253, 1990.

[8] Cockerham C.C. Higher order probability functions of identity of allelles by descent. *Genetics* **69,** 235-246,1971.

[9] Cockerham CC, Weir BS. Linkage between a marker locus and a quantitative trait of sibs. *Am J Hum Genet* 35:263-273, 1983.

[10] Cordell HJ, Kawaguchi Y, Todd JA, Farrall M. -a. An extension of the maximum lod score method to X-linked loci. *Ann Hum Genet* 59: 435-449, 1995

Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M. -b. 2-locus maximum lod score analysis of a multificatorial trait - joint consideration of IDDM2 and IDDM4 with IDDM1 in type-I diabetes. *Am J Hum Genet* 57:920-934, 1995.

[11] Cudworth, AG and Woodrow. Evidence for HLA-linked genes in "juvenile" diabetes mellitus. *Br. Med. J.* 3:133-135, 1975.

[12] Curtis D. Use of siblings as controls in case-control associqation studies. *Ann Hum Genet* 61:319-333, 1997.

[13] Daly MJ, Rioux JD, Schaffner SF, et al. High-resolution haplotype structure in the human genome. *Nat Genet* 29:229-232, 2001.

[14] Davies JL, Kawaguchi Y, Bennett ST et al. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* 371:130-136, 1994.

[15] Davis S and Weeks DE. Comparison of Nonparametric Statistics for Detection of Linkage in Nuclear Families: Single-Marker Evaluation. *Am. J. Hum. Genet.* 61:1431-1444, 1997.

[16] Day NE and Simons MJ. Disease susceptibility genes-Their identification by multiple case family studies. *Tissue Antigens* 8:109-119, 1976.

[17] Dean M, Drumm ML, Stewart C, Gerrard B, Perry A, Hidaka N, Cole JL, Collins, F.S., and Iannuzzi MC. Approaches to localizing disease genes as applied to cystic fibrosis. *Nucleic Acids Res* 18:345-350, 1990.

[18] Demenais FM, Amos CI. Power of the sib pair and lod score method for linkage analysis of quantitative traits. *Prog Clin Biol Res* 329:201-206, 1989.

[19] Dempster, A.P., Laird,N.M. and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc* 39:1-38, 1977.

[20] de Vries R.R.P., Fat R.F.M., Lai A., Nijenhuis L.E., Van Rood J.J. HLA-linked genetic cotrol of host response of Mycobacterium leprae. *Lancet* **2,** 1328-1330, 1976.

[21] Dizier MH, Babron MC and Clerget-Darpoux F. Interactive effect of two candidate genes in a disease: extension of marker-association-segregation χ^2 method. *Am J Hum Genet* 55:1042-1049, 1994.

[22] Evett, I.W. and Weir B.S. *Interpreting DNA Evidence*. Sinauer, Sunderland, MA. 1998.

[23] Ewens WJ and Clarke CP. Maximum likelihood estimation of genetic parameters of HLA-linked disease using data from families of various sizes. *Am J Hum Genet* 36:858-872, 1984.

[24] Falk CT, Rubinstein P. Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 51:227-233, 1987.

[25] Farrall M. Affected sibpair linkage tests for multiple linked susceptibility genes. *Genet Epidemiol* 14: 103-115, 1997.

[26] Field LL, Tobias R, Thompson G, Plon S. Susceptibility to insulin-dependent diabetes mellitus maps to a locus (IDDM11) on human chromosome 14q24.3-q31. *Genomics* 33:1-8, 1996.

[27] Fife M, Steer S, Fisher S, et al. Association of familial and sporadic rheumatoid arthritis with a single corticotropin-releasing hormone genomic region (8q12.3) haplotype. *Arthritis Rheum* 46:75-82, 2002.

[28] Fishman P.M., Suarez B., Hodge S.E., Reich T. A robust method for the detection of linkage in familial disease. *American Journal of Human Genetics* **30,** 308-321, 1978.

[29] Gauderman WJ, Morrison JL, Siegmund KD, Thomas DC. A joint test of linkage and gene x environment interaction, with affected sib pairs. *Genet Epidemiol* 17 Suppl 1:S563-8, 1999.

[30] Gauderman WJ, Siegmund KD. Gene-environment interaction and affected sib pair linkage analysis. *Hum Hered* 52:34-46, 2001.

[31] Génin E and Clerget-Darpoux F. Consanguinity and sib-pair method: an approach using identity by descent between and within individuals. *Am J Hum Genet* 59:1149-1162, 1996.

[32] Génin E and Clerget-Darpoux F. Letters to the editor: Reply to Weeks and Sinsheimer *Am J Hum Genet* 62:731-736, 1998.

[33] Génin E and Clerget-Darpoux F. On the probability of identity states in permutable populations: reply to Cannings *Am J Hum Genet* 62:726-727, 1998.

[34] Gillois M. Relation didentite en genetique. I postulats et axiomes mendeliens. II. Correlation genetique dans le cas de dominance. *Annales De L Institut Henri Poincare Section B-Calcul Des Probabilites Et Statistique* **2,** 1-94, 1965.

[35] Goldin LR and Weeks DE. Two locus models of disease: comparison of likelihood and nonparametric linkage methods. *Am J Hum Genet* 53:908-915, 1993.

[36] Goldstein AM, Goldin LR, Dracopoli NC, Clark WH, Tucker MA. Two-locus linkage analysis of cutaneous malignant melanoma/dysplastic nevi *Am J Hum Genet* 58: 1050-1056, 1996.

[37] Goldgar DE and Easton DF. Optimal strategies for mapping complex diseases in the presence of multiple loci. *Am J Hum Genet* 60: 1222-1232, 1997.

[38] Green JR, Low HC, Woodrow JC. Influence on inheritance of disease using repetitions of HLA haplotypes in affected siblings. *Ann Hum Genet* 47:73-82, 1983.

[39] Green J.R., Woodrow J.C. Sibling method for detecting HLA-linked genes in disease. *tissue Antigens* **9,** 31-35, 1977.

[40] Guo SW. Gene-environment interactions and the affected-sib-pair designs. *Hum Hered* 50:271-285, 2000.

[41] Harris D.L. Genotypic covariances between inbred relatives. *Genetics* **50,** 1319-1348, 1964.

[42] Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3-19, 1972.

[43] Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* 2:204-211, 1992.

[44] Hodge SE. The information contained in multiple sibling pairs. *Genet Epidemiol* 1:109-122, 1984.

[45] Holmans P. Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet* 52:362–374, 1993.

[46] Holmans P and Clayton D. Efficiency of typing unaffected relatives in an affected-sib-pair linkage study with single-locus and multiple tightly linked markers. *Am J Hum Genet* 57:1221-1232, 1995.

[47] Huang J and Vieland VJ. Comparison of 'model-free' and 'model-based' linkage statistics in the presence of locus heterogeneity: Single data set and multiple data set applications *Hum Hered* 51: 217-225, 2001.

[48] Jacquard A. *The genetic structure of populations. Springer-Verlag, New York*, 1974.

[49] Jansen RC. Interval mapping of multiple quantitative trait loci. *Genetics* 135:205-211, 1993.

[50] Juo SH, Beaty TH, Xu J, Prenger VL, Coresh J, Kwiterovich PO . Segregation analysis of two-locus models regulating apolipoprotein-A1 levels. *Genet Epidemiol* 15: 73-86, 1998.

[51] Kaplan NL, Hill WG, Weir BS. Likelihood methods for locating disease genes in nonequilibrium populations. *Am J Hum Genet* 56(1):18-32, 1995.

[52] Karigl G. A recursive algorithm for the calculation of identity coefficients. *Ann Hum Genet* 45:299-305, 1981.

[53] Kerem B, Rommens JM, Buchana JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui L-C. Identification of the cystic fibrosis gene: genetic analysis. *Science* 245:1073-1080, 1989.

[54] Knapp M, Seuchter SA and Baur MP.

-a. Linkage analysis in nuclear families. I. Optimality criteria for affected sib-pair tests. *Hum Hered* 44:37-43, 1994.

-b. Linkage analysis in nuclear families. II. Relationship between affected sib-pair tests and lod score analysis. *Hum Hered* 44:44-51, 1994.

-c. 2-locus disease-models with 2-marker loci - the power of affected-sib-pair tests. *Am J Hum Genet* 55:1030-1041, 1994.

[55] Knapp M. Evaluation of a restricted likelihood ratio test for mapping quantitative trait loci with extreme discordant sib pairs *Ann Hum Genet* 62: 75-87, 1998.

[56] Khoury M, Flanders W, Lipton R et al. The affected sib-pair method in the context of an epidemiologic study design. *Genet. Epidemiol.* 8:277-282, 1991.

[57] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 58:1347-1363, 1996.

[58] Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185-199, 1989.

[59] Lange K.

-a. The affected sib-pair method using identity by descent relations. *Am J Hum Genet* 39:148-150, 1986.

-b. A test statistic for the affected-sib-set method. *Ann Hum Genet* 50:283–290, 1986.

[60] Levinson DF. Linkage information in small family structures: comparison of pedigrees with three to five affected memebers. *Psychiatr Genet* 3:45-56, 1993.

[61] Leal SM, Ott J. Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. *Am J Hum Genet* 66:567-575, 2000.

[62] Li H, Hsu L. Effects of age at onset on the power of the affected sib pair and transmission/disequilibrium tests. *Ann Hum Genet* 64:239-54, 2000.

[63] Li WT and Reich JA. Complete enumeration and classification of two-locus disease models. *Hum Hered* 50: 334-349, 2000.

[64] Li YJ. Characterizing the structure of genetic population. *Ph.D. Thesis, North Carolina State University* 1996.

[65] Luria, SE and Delbruck, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491-511, 1943.

[66] MacLean CJ, Bishop DT, Sherman SL, Diehl SR. Distribution of lod scores under uncertain mode of inheritance. *Am J Hum Genet* 52:354-361, 1993.

[67] Mandal DM, Sorant AJ, Pugh EW, Marcus SE, Klein AP, Mathias RA, O'Neill J, Temiyakarn LF, Wilson AF, Bailey-Wilson JE. Environmental covariates: effects on the power of sib-pair linkage methods. *Genet Epidemiol* 17 Suppl 1:S643-S648, 1999.

[68] March RE. Gene mapping by linkage and association analysis. *Mol Biotechnol* 13:113-122, 1999.

[69] Martin ER, Kaplan NL, Weir BS. Tests for linkage and association in nuclear families. em Am J Hum Genet 61(2):439-448, 1997.

[70] Martin ER, Lai EH, Gilbert JR, et. al. SNPing Away at Complex Diseases: Analysis of Single-Nucleotide Polymorphisms around APOE in Alzheimer Disease. *Am J Hum Genet* 67:383394, 2000.

[71] Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 67:146-154, 2000.

[72] Mcginnis RE. Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT). *Ann Hum Genet* 62, 159-179, 1998.

[73] Meuwissen TH, Goddard ME. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155:421-430, 2000.

[74] Monaco AP, Kunkel LM. Cloning of the Duchenne/Becker muscular dystrophy locus. *Adv Hum Genet* 17:61-98, 1988.

[75] Morton NE. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7:277-318, 1955.

[76] Murray JM, Davies KE, Harper PS, Meredith L, Mueller CR, Williamson R. Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature* 300:69-71, 1982.

[77] Nemesure BB, He Q, Mendell NR. A Normalized identity-by-state statistic for linkage analysis of sib pairs. *Genet Epidemiol* 17(Suppl 1):S673-S677, 1999.

[78] Niu T, Xu X, Cordell HJ, Rogus J, Zhou Y, Fang Z, Lindpaintner K. Linkage analysis of candidate genes and gene-gene interactions in chinese hypertensive sib pairs. *Hypertension* 33:1332-1337, 1999.

[79] Olson JM. Likelihood-based models for genetic linkage analysis using affected sib pairs *Hum Hered* 47: 110-120, 1997.

[80] Payami H, Thomson G, Motro U, Louis EJ, Hudes E. The affected sib method. IV. sib trios. *Ann Hum Genet* 49:303-314, 1985.

[81] Penrose LS. The general-purpose sib-pair linkage test. *Ann Eugenics* 18:120-124, 1953.

[82] Risch N. Genetics of IDDM: evidence for complex inheritance with HLA. *Genet Epidemiol* 6:143-148, 1989.

[83] Risch N. Linkage strategies for genetically complex traits. iii. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* 46:242–253, 1990.

[84] Risch N and Merikangas, K. The future of genetic studies of complex human diseases. *Science* 273:1516-1517, 1996.

[85] Schork NJ, Boehnke M, Terwilliger JD, Ott J. 2-trait-locus linkage analysis -

a powerful strategy for mapping complex genetic-traits. *Am J Hum Genet* 53: 1127-1136, 1993.

[86] Schwab S, Albus M, Hallmayer J, Honig S, Borrmann M, Lichtermann D, Ebstein R, et al. Evaluation of susceptibility gene for schizophrenia on chromosome 6p by multipoint affected sib-pair linkage analysis. *Nat Genet* 11:325-327, 1995.

[87] Sham PC, Zhao JH, Curtis D. Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann Hum Genet* 61: 61-69, 1997.

[88] Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506-516, 1993.

[89] Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Gen* 59:983-989, 1996.

[90] Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Gen* 62:450-458, 1998.

[91] Sribney WM, Swift M. Power of sib-pair and sib-trio linkage analysis with assortative mating and multiple disease loci. *Am J Hum Genet* 51:773-784, 1992.

[92] Suarez BK, Hodge SE. Simple method to detect linkage for rare recessive disease - application to juvenile diabetes. *Clin Genet* 15:126-136, 1979.

[93] Suarez BK, Rice J, Reich T. The generalized sib pair IBD distribution: Its use in the detection of linkage. *Ann Hum Genet* 42:87-94, 1978.

[94] Suarez BK. The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. *Tissue Antigens* 12:87-93, 1978.

[95] Terwilliger JD. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet* 56:777-787, 1995.

[96] Terwilliger JD and Ott J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* 42:337-346, 1992.

[97] Terwilliger JD and Ott J. Handbook of human genetic linkage. *Johns Hopkins University Press, Baltimore and London* 1994.

[98] Thompson EA. Gene identities and multiple relationships. *Biometrics* 30:667-680, 1974.

[99] Thomson G and Motro U. Affected sib pair identity by state analyses. *Genet Epidemiol* 11:353-364, 1994.

[100] Van der Meulen MA, te Meerman GJ. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genet Epidemiol* 14:915-920, 1997.

[101] Wang T and Zeng ZB. Ph.D. dissertation. *North Carolina State University* 2001.

[102] Weeks D, Lathrop G. Polygenic disease: methods for mapping complex disease traits. *Trends Genet* 11:513-519, 1995.

[103] Weeks DE, Lange K. The affected-pedigree member method of linkage analysis. *Am J Hum Genet* 42:315-326, 1988.

[104] Weeks DE, Lange K. A multilocus extension of the affected-pedigree-member method of linkage analysis. *Am J Hum Genet* 50: 859-868, 1992.

[105] Weeks DE, Lehner T, Squires-Wheeler E, Kaufmann C, Ott J. Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. *Genet Epidemiol* 7:237-243, 1990.

[106] Weeks DE and Sinsheimer JS. Consanguinity and relative-pair methods for linkage analysis. *Am J Hum Genet* 62:728-731, 1998.

[107] Weeds DE, Valappil TI, Schroeder M, et al. An X-linked version of the affected-pedigree-member method of linkage analysis. *Hum Hered* 45: 25-33, 1995.

[108] Weir BS and Cockerham CC. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* 40:157-164, 1984.

[109] Weitkamp LR, Stancer HC, et al. Depressive disorders and HLA: a gene on chromosome 6 that can affect behaviour. *N Engl J Med* 305:1301-1306, 1981.

[110] Wilson SR. On extending the transmission/disequilibrium test (TDT). *Ann Hum Genet* 61:151-61, 1997.

[111] Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 67:936-946, 2000.

[112] Zeng ZB. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci U S A* 90:10972-10976, 1993.

[113] Zeng ZB. Precision mapping of quantitative trait loci. *Genetics* 136(4):1457-1468, 1994.

Appendix A

Derivation of closed form solutions for QTL parameters in the single marker model

In this appendix, we derive closed form solutions for all the parameters considered in section 4.2.2.

To get maximum likelihood estimates of all the parameters, we need to take derivatives of the log-likelihood function of the data. Then we should set the derivatives equal zero and solve for the parameters. Let Θ stand for the parameter set which includes QTL allele frequencies, disequilibrium coefficients between marker alleles and disease alleles, recombination fraction between marker locus and QTL locus and QTL effect parameters such as additive effect of QTL and dominance effect of QTL. We will get the following expression

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial \Theta} &= \sum_i^n \frac{1}{\sum_{z_1}\sum_{m_1} f_{i1}P_{i1}\sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}} \\
&\quad \times \sum_{z_1}\sum_{m_1} \frac{\partial f_{i1}P_{i1}\sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}}{\partial \Theta} \\
&= \sum_i^n \frac{1}{\sum_{z_1}\sum_{m_1} f_{i1}P_{i1}\sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}} \\
&\quad \times \sum_{z_1}\sum_{m_1}[f_{i1}P_{i1}\frac{\partial \log f_{i1}P_{i1}}{\partial \Theta}\sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij} \\
&\quad + f_{i1}P_{i1}\sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij} \\
&\quad \times \frac{\partial \log \sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}}{\partial \Theta}] \\
&= \sum_i^n \sum_{z_1}\sum_{m_1}[\pi_{i1}(t)\frac{\partial \log f_{i1}P_{i1}}{\partial \Theta} + \pi_{i1}(t)\frac{1}{\sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}} \\
&\quad \times \sum_{z_2}\sum_{m_2}\frac{\partial f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}}{\partial \Theta}] \\
&= \sum_i^n \sum_{z_1}\sum_{m_1}[\pi_{i1}(t)\frac{\partial \log f_{i1}P_{i1}}{\partial \Theta} + \pi_{i1}(t)\frac{1}{\sum_{z_2}\sum_{m_2} f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}} \\
&\quad \times \sum_{z_2}\sum_{m_2}[f_{i2}P_{i2}\frac{\partial \log f_{i2}P_{i2}}{\partial \Theta}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij} \\
&\quad + f_{i2}P_{i2}\prod_j \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}\frac{\partial \sum_j \log \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}}{\partial \Theta}] \\
&= \sum_i^n \sum_{z_1}\sum_{m_1}\pi_{i1}(t)[\frac{\partial \log f_{i1}P_{i1}}{\partial \Theta} + \sum_{z_2}\sum_{m_2}(\pi_{i2}(t)\frac{\partial \log f_{i2}P_{i2}}{\partial \Theta} \\
&\quad + \pi_{i2}(t)\sum_j \frac{\partial \log \sum_{m_o}\sum_{z_o} P_{oij}f_{oij}}{\partial \Theta})] \\
&= \sum_i^n \sum_{z_1}\sum_{m_1}\pi_{i1}(t)[\frac{\partial \log f_{i1}P_{i1}}{\partial \Theta} + \sum_{z_2}\sum_{m_2}\pi_{i2}(t)\{\frac{\partial \log f_{i2}P_{i2}}{\partial \Theta} \\
&\quad + \sum_j \frac{1}{\sum_{m_o}\sum_{z_o} P_{oij}f_{oij}}\sum_{m_o}\sum_{z_o}\frac{\partial P_{oij}f_{oij}}{\partial \Theta}\}] \\
&= \sum_i^n \sum_{z_1}\sum_{m_1}\pi_{i1}(t)[\frac{\partial \log f_{i1}P_{i1}}{\partial \Theta} + \sum_{z_2}\sum_{m_2}\pi_{i2}(t)\{\frac{\partial \log f_{i2}P_{i2}}{\partial \Theta}
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_j \frac{1}{\sum_{m_o}\sum_{z_o} P_{oij} f_{oij}} \sum_{m_o}\sum_{z_o} (P_{oij} f_{oij} \frac{\partial \log P_{oij} f_{oij}}{\partial \Theta})\}] \\
= &\sum_i^n \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) [\frac{\partial \log f_{i1} P_{i1}}{\partial \Theta} + \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \{\frac{\partial \log f_{i2} P_{i2}}{\partial \Theta} \\
&+\sum_s \sum_{jj'} \sum_{kk'} \pi^{jk}_{j'k', is3}(t) \frac{\partial \log P_{ois} f_{ois}}{\partial \Theta}\}]
\end{aligned}
$$

where

$$
\pi^{l_1 h_1}_{l_1' h_1', i1}(t) = \frac{f_{i1} P_{i1} \sum_{l_2 l_2'} \sum_{h_2 h_2'} f_{i2} P_{i2} \prod_s \sum_{jj'} \sum_{kk'} P_{ois} f_{ois} | \Theta = \Theta^{(t)}}{\sum_{l_1 l_1'} \sum_{h_1 h_1'} f_{i1} P_{i1} \sum_{l_2 l_2'} \sum_{h_2 h_2'} f_{i2} P_{i2} \prod_s \sum_{jj'} \sum_{kk'} P_{ois} f_{ois} | \Theta = \Theta^{(t)}}
$$

$$
\pi^{l_2 h_2}_{l_2' h_2', i2}(t) = \frac{f_{i2} P_{i2} \prod_s \sum_{jj'} \sum_{kk'} P_{ois} f_{ois} | \Theta = \Theta^{(t)}}{\sum_{l_2 l_2'} \sum_{h_2 h_2'} f_{i2} P_{i2} \prod_s \sum_{jj'} \sum_{kk'} P_{ois} f_{ois} | \Theta = \Theta^{(t)}}
$$

$$
\pi^{jk}_{j'k', is3}(t) = \frac{P_{ois} f_{ois} | \Theta = \Theta^{(t)}}{\sum_{jj'} \sum_{kk'} P_{ois} f_{ois} | \Theta = \Theta^{(t)}}
$$

Here, i indexes the family number. s is the offspring number in the i-th family. l_1 and l_1' represent the phased QTL genotype of the first parent. For example, $l_1 = 1$ and $l_1' = 0$ stand for Q_1/Q_0. h_1 and h_1' represent the phased marker genotype of the first parent. Likewise, l_2, l_2' and h_2 and h_2' represent the phased QTL and marker genotypes of the second parent. j, j' and k, k' stand for the phased QTL and marker genotypes of the children.

In the log-likelihood, only P_{i1} and P_{i2} involve the QTL allele frequencies. If we take derivatives of all the other terms in the L-function with respect to p_1 or D, they will equal zero. Thus, taking the derivative of the log-likelihood with respect to QTL allele frequency p_1, we have

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial p_1} &= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)[\frac{\partial \log P_{i1}}{\partial p_1} + \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_2 h_2}_{l_2' h_2', i2}(t)\{\frac{\partial \log P_{i2}}{\partial p_1} + 0\}] \\
&= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)[\frac{(-1)^{1+l_1} q_{h_1}}{P_{l_1 h_1}} + \frac{(-1)^{1+l_1'} q_{h_1'}}{P_{l_1' h_1'}} \\
&\quad + \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_2 h_2}_{l_2' h_2', i2}(t)\{\frac{(-1)^{1+l_2} q_{h_2}}{P_{l_2 h_2}} + \frac{(-1)^{1+l_2'} q_{h_2'}}{P_{l_2' h_2'}}\}] \\
&= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \frac{(-1)^{1+l_1} q_{h_1}}{P_{l_1 h_1}} \{\pi^{l_1 h_1}_{l_1' h_1', i1}(t) + \pi^{l_1' h_1'}_{l_1 h_1, i1}(t)\} \\
&\quad + \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \frac{(-1)^{1+l_2} q_{h_2}}{P_{l_2 h_2}} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)\{\pi^{l_2 h_2}_{l_2' h_2', i2}(t) + \pi^{l_2' h_2'}_{l_2 h_2, i2}(t)\} \\
&= \sum_{l_1 h_1} [\sum_i \sum_{l_1' h_1'} \{\pi^{l_1 h_1}_{l_1' h_1', i1}(t) + \pi^{l_1' h_1'}_{l_1 h_1, i1}(t)\}] \frac{(-1)^{1+l_1} q_{h_1}}{P_{l_1 h_1}} \\
&\quad + \sum_{l_2 h_2} [\sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2' h_2'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)\{\pi^{l_2 h_2}_{l_2' h_2', i2}(t) + \pi^{l_2' h_2'}_{l_2 h_2, i2}(t)\}] \frac{(-1)^{1+l_2} q_{h_2}}{P_{l_2 h_2}} \\
&= 0
\end{aligned}
$$

Here, when $P_{i1} = 0$, $\pi^{l_1 h_1}_{l_1' h_1', i1}(t) = 0$, so

$$
\pi^{l_1 h_1}_{l_1' h_1', i1}(t) \frac{\partial \log P_{i1}}{\partial p_1} = \pi^{l_1 h_1}_{l_1' h_1', i1}(t)(\frac{(-1)^{1+l_1} q_{h_1}}{P_{l_1 h_1}} + \frac{(-1)^{1+l_1'} q_{h_1'}}{P_{l_1' h_1'}})
$$

Similarly, when $P_{i2} = 0$, $\pi^{l_2 h_2}_{l_2' h_2', i2}(t) = 0$, so

$$
\pi^{l_2 h_2}_{l_2' h_2', i2}(t) \frac{\partial \log P_{i2}}{\partial p_1} = \pi^{l_2 h_2}_{l_2' h_2', i2}(t)(\frac{(-1)^{1+l_2} q_{h_2}}{P_{l_2 h_2}} + \frac{(-1)^{1+l_2'} q_{h_2'}}{P_{l_2' h_2'}})
$$

Likewise, only P_{i1} and P_{i2} involve the disequilibrium coefficient, D. Thus, taking

derivative with respect to D, we have

$$\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial D} &= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)[\frac{\partial \log P_{i1}}{\partial D} + \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_2 h_2}_{l_2' h_2', i2}(t)\{\frac{\partial \log P_{i2}}{\partial D} + 0\}] \\
&= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)[\frac{(-1)^{l_1+h_1}}{P_{l_1 h_1}} + \frac{(-1)^{l_1'+h_1'}}{P_{l_1' h_1'}} \\
&\quad + \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_2 h_2}_{l_2' h_2', i2}(t)\{\frac{(-1)^{l_2+h_2}}{P_{l_2 h_2}} + \frac{(-1)^{l_2'+h_2'}}{P_{l_2' h_2'}}\}] \\
&= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \frac{(-1)^{l_1+h_1}}{P_{l_1 h_1}} \{\pi^{l_1 h_1}_{l_1' h_1', i1}(t) + \pi^{l_1' h_1'}_{l_1 h_1, i1}(t)\} \\
&\quad + \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \frac{(-1)^{l_2+h_2}}{P_{l_2 h_2}} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)\{\pi^{l_2 h_2}_{l_2' h_2', i2}(t) + \pi^{l_2' h_2'}_{l_2 h_2, i2}(t)\} \\
&= \sum_{l_1 h_1} [\sum_i \sum_{l_1' h_1'} \{\pi^{l_1 h_1}_{l_1' h_1', i1}(t) + \pi^{l_1' h_1'}_{l_1 h_1, i1}(t)\}] \frac{(-1)^{l_1+h_1}}{P_{l_1 h_1}} \\
&\quad + \sum_{l_2 h_2} [\sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2' h_2'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)\{\pi^{l_2 h_2}_{l_2' h_2', i2}(t) + \pi^{l_2' h_2'}_{l_2 h_2, i2}(t)\}] \frac{(-1)^{l_2+h_2}}{P_{l_2 h_2}} \\
&= \sum_{l_1 h_1} a_{l_1 h_1}(t) \frac{(-1)^{l_1+h_1}}{P_{l_1 h_1}} + \sum_{l_2 h_2} b_{l_2 h_2}(t) \frac{(-1)^{l_2+h_2}}{P_{l_2 h_2}} \\
&= \sum_{l_1 h_1} (a_{l_1 h_1}(t) + b_{l_1 h_1}(t)) \frac{(-1)^{l_1+h_1}}{P_{l_1 h_1}} \\
&= 0
\end{aligned}$$

where

$$\begin{aligned}
a_{l_1 h_1}(t) &= \sum_i \sum_{l_1' h_1'} \{\pi^{l_1 h_1}_{l_1' h_1', i1}(t) + \pi^{l_1' h_1'}_{l_1 h_1, i1}(t)\} \\
b_{l_2 h_2}(t) &= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2' h_2'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)\{\pi^{l_2 h_2}_{l_2' h_2', i2}(t) + \pi^{l_2' h_2'}_{l_2 h_2, i2}(t)\}
\end{aligned}$$

To get closed form solutions for the QTL effected parameters, we need to take derivatives of the log-likelihood with respect to each QTL effect parameter. In the

log-likelihood, only f_{i1},f_{i2}, and f_{ois} are functions of QTL effect parameters. For simplicity, we assume that the i-family has s_i offspring.

Taking derivative of the log likelihood function with respect to the overall population mean of the trait value, we obtain

$$\begin{aligned}
\frac{\partial L(\Theta|\Theta^t)}{\partial \mu} &= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)[\frac{\partial \log f_{i1}}{\partial \mu} + \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_2 h_2}_{l_2' h_2', i2}(t)\{\frac{\partial \log f_{i2}}{\partial \mu} \\
&\quad + \sum_s \sum_{jj'} \sum_{kk'} \pi^{jk}_{j'k', is3}(t) \frac{\partial \log f_{ois}}{\partial \mu}\}] \\
&= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \frac{(y_{i1} - \mu - X_{l_1 l_1'} E)}{\sigma^2} \\
&\quad + \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \frac{(y_{i2} - \mu - X_{l_2 l_2'} E)}{\sigma^2} \\
&\quad + \sum_i \sum_s \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \sum_{jj'} \sum_{kk'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \pi^{jk}_{j'k', is3}(t) \\
&\quad \times \frac{(y_{ois} - \mu - X_{jj'} E)}{\sigma^2} \\
&= 0
\end{aligned}$$

To simplify the above expression, we define

$c_{l_1 l_1', i1}(t) = \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t)$

$c_{l_2 l_2', i2}(t) = \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{h_2 h_2'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t)$

$c_{jj', is3}(t) = \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \sum_{kk'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \pi^{jk}_{j'k', is3}(t)$.

Then

$\sum_{l_1 l_1'} c_{l_1 l_1', i1}(t) = \sum_{l_2 l_2'} c_{l_2 l_2', i2}(t) = 1$ and $\sum_s \sum_{jj'} c_{jj', is3}(t) = s_i$.

Then the original expression becomes

$$
\begin{aligned}
& \sum_i \sum_{l_1 l_1'} c_{l_1 l_1', i1}(t)(y_{i1} - \mu - X_{l_1 l_1'} E) + \sum_i \sum_{l_2 l_2'} c_{l_2 l_2', i2}(t)(y_{i2} - \mu - X_{l_2 l_2'} E) \\
& + \sum_i \sum_s \sum_{jj'} c_{jj', is3}(t)(y_{ois} - \mu - X_{jj'} E) \\
= & \sum_i \sum_{jj'} c_{jj', i1}(t)(y_{i1} - \mu - X_{jj'} E) + \sum_i \sum_{jj'} c_{jj', i2}(t)(y_{i2} - \mu - X_{jj'} E) \\
& + \sum_i \sum_s \sum_{jj'} c_{jj', is3}(t)(y_{ois} - \mu - X_{jj'} E) \\
= & \ 0
\end{aligned}
$$

Solving for μ, we obtain the following expression

$$
\begin{aligned}
\sum_i \sum_{jj'} (c_{jj', i1}(t) + c_{jj', i2}(t) + \sum_s c_{jj', is3}(t))\mu \ = \ & \sum_i (y_{i1} + y_{i2} + \sum_s y_{ois}) \\
& - \sum_i \sum_{jj'} (c_{jj', i1}(t) + c_{jj', i2}(t)) X_{jj'} E \\
& - \sum_i \sum_s \sum_{jj'} c_{jj', is3}(t) X_{jj'} E
\end{aligned}
$$

then

$$
\begin{aligned}
\mu^{(t+1)} \ = \ & \frac{1}{\sum_i (2 + s_i)} [\sum_i (y_{i1} + y_{i2} + \sum_s y_{ois}) \\
& - \sum_i \sum_{jj'} (c_{jj', i1}(t) + c_{jj', i2}(t)) X_{jj'} E^{(t)} - \sum_i \sum_s \sum_{jj'} c_{jj', is3}(t) X_{jj'} E^{(t)}]
\end{aligned}
$$

Similarly, taking derivative with respective to the additive effect of QTL, we can get

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^t)}{\partial a} &= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \frac{\partial (X_{l_1 l_1'} E)}{\partial a} \frac{(y_{i1} - \mu - X_{l_1 l_1'} E)}{\sigma^2} \\
&\quad + \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \frac{\partial (X_{l_2 l_2'} E)}{\partial a} \frac{(y_{i2} - \mu - X_{l_2 l_2'} E)}{\sigma^2} \\
&\quad + \sum_i \sum_s \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \sum_{jj'} \sum_{kk'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \pi^{jk}_{j'k', is3}(t) \\
&\quad \times \frac{\partial (X_{jj'} E)}{\partial a} \frac{(y_{ois} - \mu - X_{jj'} E)}{\sigma^2} \\
&= \sum_i \sum_{l_1 l_1'} c_{l_1 l_1', i1}(t) \frac{\partial (X_{l_1 l_1'} E)}{\partial a} \frac{(y_{i1} - \mu - X_{l_1 l_1'} E)}{\sigma^2} \\
&\quad + \sum_i \sum_{l_2 l_2'} c_{l_2 l_2', i2}(t) \frac{\partial (X_{l_2 l_2'} E)}{\partial a} \frac{(y_{i2} - \mu - X_{l_2 l_2'} E)}{\sigma^2} \\
&\quad + \sum_i \sum_s \sum_{jj'} c_{jj', is3}(t) \frac{\partial (X_{jj'} E)}{\partial a} \frac{(y_{ois} - \mu - X_{jj'} E)}{\sigma^2} \\
&= 0
\end{aligned}
$$

Then

$$
\begin{aligned}
&\sum_i [c_{11,i1}(t)(y_{i1} - \mu - a + \frac{d}{2}) - c_{00,i1}(t)(y_{i1} - \mu + a + \frac{d}{2}) \\
&+ c_{11,i2}(t)(y_{i2} - \mu - a + \frac{d}{2}) - c_{00,i2}(t)(y_{i2} - \mu + a + \frac{d}{2}) \\
&+ \sum_s \{c_{11,is3}(t)(y_{ois} - \mu - a + \frac{d}{2}) - c_{00,is3}(t)(y_{ois} - \mu + a + \frac{d}{2})\}] \\
= \; & 0
\end{aligned}
$$

Thus,

$$
a^{(t+1)} = [\sum_i \{(c_{11,i1}(t) - c_{00,i1}(t))(y_{i1} - \mu^{(t+1)} + \frac{d^{(t)}}{2}) + (c_{11,i2}(t) - c_{00,i2}(t))
$$

$$\times(y_{i2}-\mu^{(t+1)}+\frac{d^{(t)}}{2})+\sum_s(c_{11,i3}(t)-c_{00,i3}(t))(y_{ois}-\mu^{(t+1)}+\frac{d^{(t)}}{2})\}]$$
$$/[\sum_i\{c_{11,i1}(t)+c_{00,i1}(t)+c_{11,i2}(t)+c_{00,i2}(t)+\sum_s(c_{11,is3}(t)+c_{00,is3}(t))\}]$$

Likewise, taking derivative with respect to d, we get

$$\begin{aligned}\frac{\partial L(\Theta|\Theta^t)}{\partial d} &= \sum_i\sum_{l_1l_1'}\sum_{h_1h_1'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)\frac{\partial(X_{l_1l_1'}E)}{\partial d}\frac{(y_{i1}-\mu-X_{l_1l_1'}E)}{\sigma^2}\\ &+\sum_i\sum_{l_1l_1'}\sum_{h_1h_1'}\sum_{l_2l_2'}\sum_{h_2h_2'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)\pi^{l_2h_2}_{l_2'h_2',i2}(t)\frac{\partial(X_{l_2l_2'}E)}{\partial d}\frac{(y_{i2}-\mu-X_{l_2l_2'}E)}{\sigma^2}\\ &+\sum_i\sum_s\sum_{l_1l_1'}\sum_{h_1h_1'}\sum_{l_2l_2'}\sum_{h_2h_2'}\sum_{jj'}\sum_{kk'}\pi^{l_1h_1}_{l_1'h_1',i1}(t)\pi^{l_2h_2}_{l_2'h_2',i2}(t)\pi^{jk}_{j'k',is3}(t)\\ &\times\frac{\partial(X_{jj'}E)}{\partial d}\frac{(y_{ois}-\mu-X_{jj'}E)}{\sigma^2}\\ &= 0\end{aligned}$$

Solving for d, we get

$$\begin{aligned}d^{(t+1)} &= \frac{2}{\sum_i(2+s_i)}\sum_i[(c_{01,i1}(t)+c_{10,i1}(t)-c_{11,i1}(t)-c_{00,i1}(t))(y_{i1}-\mu^{(t+1)})\\ &+(c_{01,i2}(t)+c_{10,i2}(t)-c_{11,i2}(t)-c_{00,i2}(t))(y_{i2}-\mu^{(t+1)})\\ &+\sum_s\{(c_{01,is3}(t)+c_{10,is3}(t)-c_{11,is3}(t)-c_{00,is3}(t))(y_{ois}-\mu^{(t+1)})\}\\ &+\{c_{11,i1}(t)-c_{00,i1}(t)+c_{11,i2}(t)-c_{00,i2}(t)+\sum_s(c_{11,is3}(t)-c_{00,is3}(t))\}a^{(t+1)}]\end{aligned}$$

Likewise,

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^t)}{\partial \sigma^2} &= \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \frac{(y_{i1} - \mu - X_{l_1 l_1'} E)^2}{2\sigma^4} \\
&+ \sum_i \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \frac{(y_{i2} - \mu - X_{l_2 l_2'} E)^2}{2\sigma^4} \\
&+ \sum_i \sum_s \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \sum_{jj'} \sum_{kk'} \pi^{l_1 h_1}_{l_1' h_1', i1}(t) \pi^{l_2 h_2}_{l_2' h_2', i2}(t) \pi^{jk}_{j'k', is3}(t) \\
&\times \frac{(y_{ois} - \mu - X_{jj'} E)^2}{2\sigma^4} - \frac{\sum_i (2 + s_i)}{2\sigma^2} \\
&= 0
\end{aligned}
$$

Then,

$$
\begin{aligned}
(\sigma^2)^{(t+1)} &= \frac{1}{\sum_i (2 + s_i)} [\sum_i \sum_{jj'} \{ c_{jj', i1}(t) (y_{i1} - \mu^{(t+1)} - X_{jj'} E^{(t+1)})^2 \\
&+ c_{jj', i2}(t) (y_{i2} - \mu^{(t+1)} - X_{jj'} E^{(t+1)})^2 \\
&+ \sum_s c_{jj', is3}(t) (y_{ois} - \mu^{(t+1)} - X_{jj'} E^{(t+1)})^2 \}]
\end{aligned}
$$

Appendix B

Derivation of closed form solutions for QTL parameters in the two marker model

In this appendix, we will derive closed form solutions for QTL allele frequency, p_1 and the three disequilibrium coefficients D_{01}, D_{02} and D_{012}. In the log-likelihood, only P_{i1} and P_{i2} are functions of QTL allele frequency, p_1. Taking the derivative with respect to p_1, we get

$$\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial p_1} &= \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) [\frac{\partial \log P_{i1}}{\partial p_1} + \sum_{j_2 j_2'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) \\
&\quad \times \{\frac{\partial \log P_{i2}}{\partial p_1} + 0\}] \\
&= \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) [\frac{(-1)^{1+j_1} P_{l_1 h_1}}{P_{j_1 l_1 h_1}} + \frac{(-1)^{1+j_1'} P_{l_1' h_1'}}{P_{j_1' l_1' h_1'}} \\
&\quad + \sum_{j_2 j_2'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) \{\frac{(-1)^{1+j_2} P_{l_2 h_2}}{P_{j_2 l_2 h_2}} + \frac{(-1)^{1+j_2'} P_{l_2' h_2'}}{P_{j_2' l_2' h_2'}}\}] \\
&= \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \frac{(-1)^{1+j_1} P_{l_1 h_1}}{P_{j_1 l_1 h_1}} (\pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) + \pi^{j_1' l_1' h_1'}_{j_1 l_1 h_1, i1})(t)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{j_2 j_2'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \frac{(-1)^{1+j_2} P_{l_2 h_2}}{P_{j_2 l_2 h_2}} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) \\
&\times (\pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) + \pi^{j_2' l_2' h_2'}_{j_2 l_2 h_2, i2}(t)) \\
= \;& 0
\end{aligned}
$$

To simply the above expression, define

$$
\begin{aligned}
&a_{j_1 l_1 h_1}(t) = \textstyle\sum_i \sum_{j_1' l_1' h_1'} (\pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) + \pi^{j_1' l_1' h_1'}_{j_1 l_1 h_1, i1}(t)) \\
&b_{j_2 l_2 h_2}(t) = \textstyle\sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{j_2' l_2' h_2'} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) (\pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) + \pi^{j_2' l_2' h_2'}_{j_2 l_2 h_2, i2}(t)) \\
&c_{l_1 h_1}(t) = \textstyle\sum_{j_1} \frac{(-1)^{j_1}}{P_{j_1 l_1 h_1}} (a_{j_1 l_1 h_1}(t) + b_{j_1 l_1 h_1}(t))
\end{aligned}
$$

Then, the derivative becomes

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial p_1} &= \sum_{j_1 l_1 h_1} \frac{(-1)^{1+j_1} P_{l_1 h_1}}{P_{j_1 l_1 h_1}} a_{j_1 l_1 h_1}(t) + \sum_{j_2 l_2 h_2} \frac{(-1)^{1+j_2} P_{l_2 h_2}}{P_{j_2 l_2 h_2}} b_{j_2 l_2 h_2}(t) \\
&= \sum_{j_1 l_1 h_1} \frac{(-1)^{1+j_1} P_{l_1 h_1}}{P_{j_1 l_1 h_1}} (a_{j_1 l_1 h_1}(t) + b_{j_1 l_1 h_1}(t)) \\
&= \sum_{l_1 h_1} (-1)^1 c_{l_1 h_1}(t) P_{l_1 h_1} \\
&= 0
\end{aligned}
$$

Similarly, only P_{i1} and P_{i2} are functions of the three disequilibrium coefficients. taking derivative with respect of D_{012}, D_{01}, D_{02},we can get

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial D_{012}} &= \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \frac{(-1)^{j_1+l_1+h_1}}{P_{j_1 l_1 h_1}} (\pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) + \pi^{j_1' l_1' h_1'}_{j_1 l_1 h_1, i1}(t)) \\
&\quad + \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{j_2 j_2'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \frac{(-1)^{j_2+l_2+h_2}}{P_{j_2 l_2 h_2}} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) (\pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) \\
&\quad + \pi^{j_2' l_2' h_2'}_{j_2 l_2 h_2, i2}(t)) \\
&= \sum_{j_1 l_1 h_1} \frac{(-1)^{j_1+l_1+h_1}}{P_{j_1 l_1 h_1}} (a_{j_1 l_1 h_1}(t) + b_{j_1 l_1 h_1}(t))
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{l_1 h_1} (-1)^{l_1+h_1} c_{l_1 h_1}(t) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial D_{01}} &= \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \frac{(-1)^{j_1+l_1} w_{h_1}}{P_{j_1 l_1 h_1}} (\pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) + \pi^{j_1' l_1' h_1'}_{j_1 l_1 h_1, i1}(t)) \\
&\quad + \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{j_2 j_2'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \frac{(-1)^{j_2+l_2} w_{h_2}}{P_{j_2 l_2 h_2}} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) (\pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) \\
&\quad + \pi^{j_2' l_2' h_2'}_{j_2 l_2 h_2, i2}(t)) \\
&= \sum_{j_1 l_1 h_1} \frac{(-1)^{j_1+l_1} w_{h_1}}{P_{j_1 l_1 h_1}} (a_{j_1 l_1 h_1}(t) + b_{j_1 l_1 h_1}(t)) \\
&= \sum_{l_1 h_1} (-1)^{l_1} c_{l_1 h_1}(t) w_{h_1} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L(\Theta|\Theta^{(t)})}{\partial D_{02}} &= \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \frac{(-1)^{j_1+h_1} q_{l_1}}{P_{j_1 l_1 h_1}} (\pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) + \pi^{j_1' l_1' h_1'}_{j_1 l_1 h_1, i1}(t)) \\
&\quad + \sum_i \sum_{j_1 j_1'} \sum_{l_1 l_1'} \sum_{h_1 h_1'} \sum_{j_2 j_2'} \sum_{l_2 l_2'} \sum_{h_2 h_2'} \frac{(-1)^{j_2+h_2} q_{l_2}}{P_{j_2 l_2 h_2}} \pi^{j_1 l_1 h_1}_{j_1' l_1' h_1', i1}(t) (\pi^{j_2 l_2 h_2}_{j_2' l_2' h_2', i2}(t) \\
&\quad + \pi^{j_2' l_2' h_2'}_{j_2 l_2 h_2, i2}(t)) \\
&= \sum_{j_1 l_1 h_1} \frac{(-1)^{j_1+h_1} q_{l_1}}{P_{j_1 l_1 h_1}} (a_{j_1 l_1 h_1}(t) + b_{j_1 l_1 h_1}(t)) \\
&= \sum_{l_1 h_1} (-1)^{h_1} c_{l_1 h_1}(t) q_{l_1} \\
&= 0
\end{aligned}
$$

From above derivatives, we can get four expression with four unknown parameters,

$$\begin{cases} c_{00}(t)P_{00} + c_{01}(t)P_{01} + c_{10}(t)P_{10} + c_{11}(t)P_{11} = 0 \\ c_{00}(t) - c_{01}(t) - c_{10}(t) + c_{11}(t) = 0 \\ c_{00}(t)w_0 + c_{01}(t)w_1 - c_{10}(t)w_0 - c_{11}(t)w_1 = 0 \\ c_{00}(t)q_0 - c_{01}(t)q_0 + c_{10}(t)q_1 - c_{11}(t)q_1 = 0 \end{cases}$$

Solve for $c_{l_1h_1}(t)$, we can get $c_{00}(t) = c_{01}(t) = c_{10}(t) = c_{11}(t) = 0$. Then,

$$\begin{aligned} \sum_{j_1} \frac{(-1)^{j_1}}{P_{j_1 l_1 h_1}} (a_{j_1 l_1 h_1}(t) + b_{j_1 l_1 h_1}(t)) &= 0 \\ \Longrightarrow \frac{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t)}{P_{0 l_1 h_1}} &= \frac{a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)}{P_{1 l_1 h_1}} \\ \Longrightarrow P^{(t)}_{Q_1 M_{l_1} N_{h_1}} &= \frac{a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)}{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t)} P_{0 l_1 h_1} \\ \Longrightarrow P^{(t)}_{Q_1 M_{l_1} N_{h_1}} &= \frac{a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)}{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t)} (P_{l_1 h_1} - P_{1 l_1 h_1}) \\ \Longrightarrow P^{(t)}_{Q_1 M_{l_1} N_{h_1}} &= \frac{a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)}{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t) + a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)} P_{l_1 h_1} \end{aligned}$$

Similarly,

$$P^{(t)}_{Q_0 M_{l_1} N_{h_1}} = \frac{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t)}{a_{0 l_1 h_1}(t) + b_{0 l_1 h_1}(t) + a_{1 l_1 h_1}(t) + b_{1 l_1 h_1}(t)} P_{l_1 h_1}$$

Thus,

$$\begin{aligned} P^{(t)}_{M_{l_1} Q_0} &= P^{(t)}_{Q_0 M_{l_1} N_0} + P^{(t)}_{Q_0 M_{l_1} N_1} \\ P^{(t)}_{Q_0 N_{h_1}} &= P^{(t)}_{Q_0 M_0 N_{h_1}} + P^{(t)}_{Q_0 M_1 N_{h_1}} \\ p_j^{(t+1)} &= \sum_{l_1 h_1} P^{(t)}_{Q_j M_{l_1} N_{h_1}} \quad \text{for j=0, 1} \end{aligned}$$

$$
\begin{aligned}
D_{01}^{(t+1)} &= P_{Q_0 M_0}^{(t)} - p_0^{(t+1)} q_0 \\
D_{02}^{(t+1)} &= P_{Q_0 N_0}^{(t)} - p_0^{(t+1)} w_0 \\
D_{012}^{(t+1)} &= P_{Q_0 M_0 N_0}^{(t)} - w_0 D_{01}^{(t+1)} - q_0^{(t+1)} D_{02}^{(t+1)} - p_0^{(t+1)} P_{M_0 N_0}
\end{aligned}
$$